Sociophonetics of Popular Music:

Insights from Corpus Analysis and Speech Perception Experiments

Andy M. Gibson

This dissertation is submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy in Linguistics

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Department of Linguistics
University of Canterbury
New Zealand
September 2019

To my Dad, who taught me the joy of climbing hills on blustery days.

Abstract

This thesis examines the flexibility and context-sensitivity of speech perception by looking at a domain not often explored in the study of language cognition — popular music. Three empirical studies are presented. The first examines the current state of sociolinguistic variation in commercial popular music, while the second and third explore everyday listeners' perception of language in musical and non-musical contexts. The foundational assumption of the thesis is that the use of 'American English' in song is automatic for New Zealand singers, and constitutes a responsive style that is both accurate and consistent. The use of New Zealand English in song, by contrast, is stylised, involving an initiative act of identity and requiring effort and awareness. This will be discussed in Chapter 1, where I also introduce the term Standard Popular Music Singing Style (SPMSS) to refer to the US English-derived phonetic style dominant in popular song.

The first empirical study will be presented in Chapter 2. Using a systematically selected corpus of commercial pop and hip hop from NZ and the USA, analysis of non-prevocalic and linking /r/, and the vowels of the BATH, LOT and GOAT lexical sets confirm that SPMSS is highly normative in NZ music. Most pop singers closely follow US patterns, while several hip hop artists display elements of New Zealand English. This reflects the value placed on authenticity in hip hop, and also interacts with ethnicity, showing the use of different authentication practices by Pākehā (NZ European) and Māori/Pasifika artists. By looking at co-variation amongst the variables, I explore both the apparent identity goals of the artists, and the relative salience of the variables. Chapters 3 and 4 use the results of the corpus analysis to explore how the dominance of SPMSS affects speech processing.

The first of the two perception experiments is a phonetic categorisation task. Listeners decide whether they hear the word *bed* or *bad* in a condition where the stimuli are either set to music, or appear in one of two non-musical control conditions. The stimuli are on a resynthesised continuum between the DRESS and TRAP vowels, passing through an F1 space where the vowel is ambiguous and could either be perceived as a spoken NZE TRAP or a sung DRESS. When set to music, the NZ listeners perceive the vowel according to expectations of SPMSS (i.e. expecting US-derived vowel qualities). The second perception experiment is a lexical decision task that uses the natural speech of a NZ and a US speaker, once again in musical and non-musical conditions. Participants' processing of the US voice is facilitated in the music condition, becoming faster than reaction times to their native dialect.

Bringing the results of the corpus and perception studies together, this thesis shows that SPMSS is highly normative in NZ popular music not just for performers, but also in the minds of the general music-listening public. I argue that many New Zealanders are bidialectal, with native-like knowledge of SPMSS. Speech and song are two highly distinct and perceptually contrastive contexts of language use. By differing from conversational language across a range of perceptual and cognitive dimensions, language heard or produced in song is likely to encode and activate a distinct subset of auditory memories. The contextual specificity of such networks may then allow for the abstraction of an independent sub-system of sociophonetic knowledge specific to the musical context.

Acknowledgements

Jen, thank you for seeing my potential, even when all outward manifestations of productivity lay dormant. You kept on believing that I had a PhD in me, and your patience and persistence with me both prior to, and during my enrolment have had a profoundly positive impact on me. One of the highlights of my time at NZILBB has been the intellectual treat of hearing your comments and responses to a wide range of issues in Socio meetings, reading groups, and stats chats over these four years. I am truly grateful to have learnt so much from you. Lynn, you continually reminded me what optimism and positivity look like, along with providing empathy and support during the difficult times. Every chat and every meeting left me feeling like this mission was a bit more possible than it had seemed before. Thank you for insisting that it was within my abilities to 'get it done'! Catherine, thank you for steering me through my cognitive neuroscience learning. I'm so grateful for the hands on training with EEG, and for all your help when I was battling with E-Prime. Thanks for being supportive and enthusiastic even when it took me forever to get anything done.

Coming back to Canterbury after ten years, it was so wonderful to reconnect with old friends like Emma, Heidi, Margaret and Jeanette on-site, while visits from Katie, Abby, and Anita were all real highlights. I then had the pleasure of getting to know so many other wonderful PhD students, post-docs and faculty through the amazing research community that is NZILBB. So many people here have helped me in big and small ways: Kevin, Vica, Petya, Maria, Matthias, Keyi, Ksenia, Xuan, Mohammed, Mineko, Clay, Yoonmi, Scott, Doreen, Ruth, Arshad, Muneir, Jacq (highlighters and curry), Stephanie, Merten, Vicky, Sidney, Moonsun, Wakayo, Donald, Susan, Marie, James, Jeremy, Dan, Simon, Alison, Ryan, Darcy (and Biscuit). Many thanks to both the friends and strangers that participated in the experiments, to the two colleagues who recorded the stimuli, and a second shout-out to Ryan for letting me use Science Music. Jonathan Wiltshire, thank you for helping me with E-Prime. Robert Fromont, I honestly don't know what I would have done without you (aside from amazing LaBB-CAT support, the lexical frequencies in song were entirely collected and calculated by Robert). A big thank you to all of the singers and rappers who replied to my requests for info about their ethnicities. Jody Lloyd, thank you for being the great champion of own-accent singing and rap in NZ. I look forward to making more music together.

I owe a great debt to Allan Bell. Your influence continues to flow through these pages. Thank you for all the opportunities you have given me, and for your friendship. A heartfelt thanks to Janet Holmes and Paul Warren for sparking my interest in linguistics many years ago, and also for enriching collaborations in more recent years. Big love to Bronwyn and Laura, my original ling crew. Thanks to Malcah Yaeger-Dror for keeping me in the loop, and to Dave Britain for collegiality and cocktails. To Sarah Hawkins, thank you for all the chats — your wisdom helped me solve some big problems! Caitlin Smith, thank you for reminding me where all these research questions came from.

I am so grateful for my family. Huge love to Alana, Riley, Dylan, Lucas, Tolstoy, Caspian, Ambrosia, Estonia and Olivia. Mum, you're the best... you understand me so well, and your support through this PhD journey has been amazing. Jo, Karen, and Debs, I love each of you so much. Finally, to Gabriel — thank you for being by my side through this whole process and for bringing joy into my daily life. Here's to the continuation of this beautiful thing we have.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Contextualising the Sociophonetics of Popular Music

1.1 American Accents in NZ Popular Music

I grew up writing songs. It was only when I took my first class in phonetics that I realised that I sang my songs in an 'American accent'[1]. That realisation did not happen across all aspects of my singing accent in one moment, but spread gradually from word to word, from vowel to vowel. For example, I became aware of pronouncing *can't* with [ae] in song, despite pronouncing it with [aː] in speech. Through a succession of noticings, I started a process of transitioning my singing accent to something closer to my native New Zealand English (NZE) dialect. This process took time, effort, and awareness. My experience differed from the one described by Trudgill (1983) for British artists using American English (AmE) variants in song. Trudgill described these artists as *trying to sound American.* My experience, by contrast, had been one of *trying to sound like a New Zealander* in song. Many of the insights presented in Trudgill (1983) fit with my experiences, including the key idea of conflicting identity motivations — wanting to sound like an authentic member of a certain tradition of song, and simultaneously wanting to project an 'authentic' identity.

Trudgill (1983) suggested that adopting an American-like accent in song might be driven by a desire for context-appropriateness, but ultimately dropped this line of argumentation as lacking in explanatory power. I argue in this thesis that exemplar theories of language representation provide a framework in which appropriateness to context becomes a viable and testable explanation for the processes involved in singing accents. Social constructionist approaches to language style in sociolinguistics (Eckert, 2012) which have emerged since that early work also provide valuable tools for examining this issue. They highlight the dynamic flexibility exercised by speakers in their construction and reconstruction of identities, personas and stances in situated discourse.

Clearly, my personal experience with the phonetics of singing is anecdotal. In my Masters work (Gibson, 2010b; Gibson and Bell, 2012), I explored the possi-

[1]'Accent' will be used in the sense of a social or regional speech style which differs from another primarily on phonetic/phonological grounds. Use of the word in the context of prosody will be avoided.

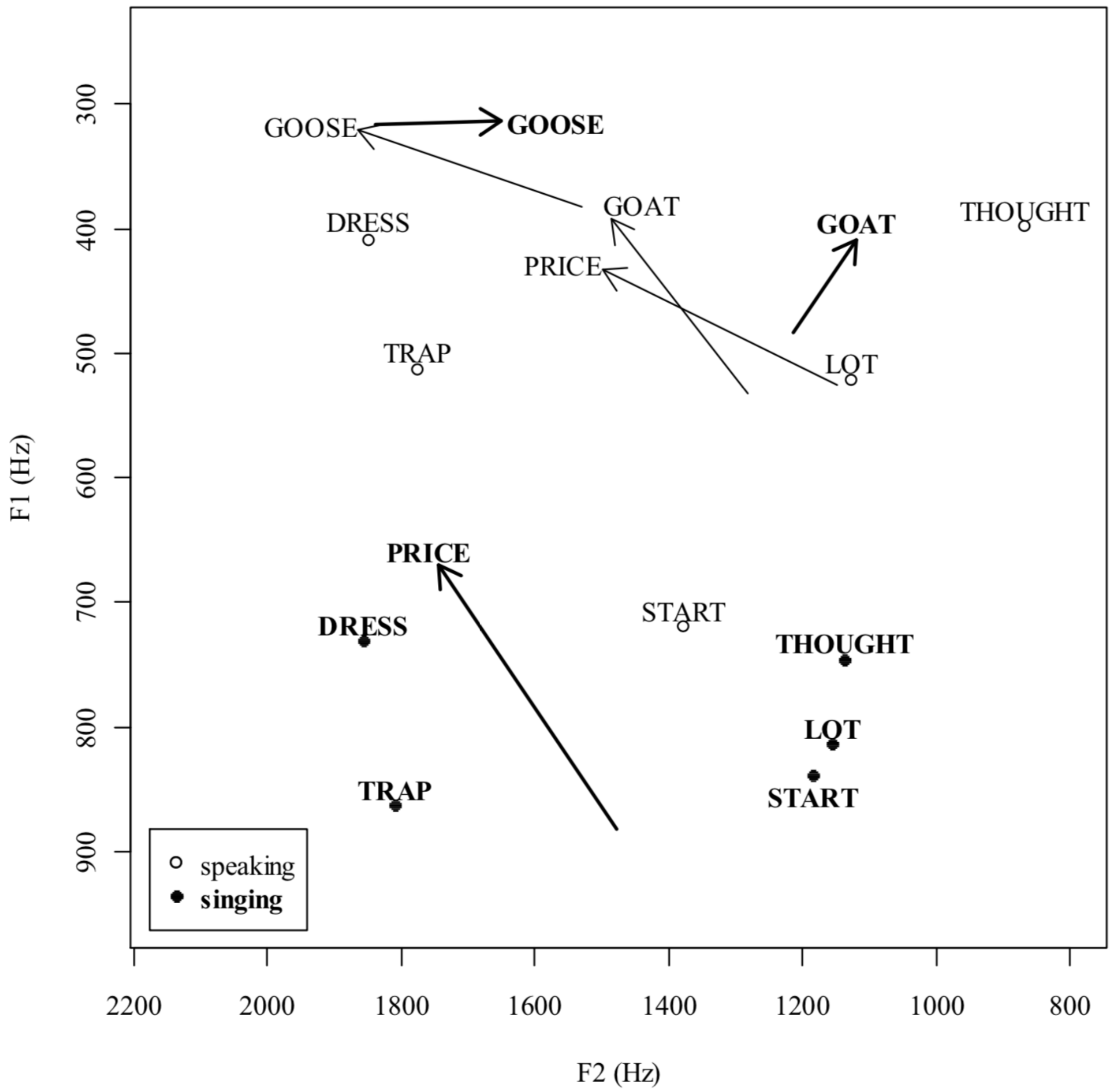


Figure 1.1: Mean F1 and F2 of sung and spoken vowels for Dylan Storey, reproduced from Gibson (2010b).

bility that my perspective was not unique. By conducting an acoustic analysis of eight vowel variables in the singing and speech of three New Zealand (NZ) singer-songwriters, and by interviewing them about their experiences, I found strong evidence that the 'default' accent for popular song for these New Zealanders was also an American-derived one. While Trudgill (1983), and many subsequent studies of singing accents, focused on variables which have, at least, marker status in distinguishing American and British Englishes, my Masters project focused largely on less salient variables. The vowels studied (DRESS, TRAP, THOUGHT, LOT, START, GOOSE, GOAT, and PRICE) were analysed in three contexts: the isolated sung performances from recordings by each of the artists, the same lyrics recited by the artists, and interview speech. The reciting of lyrics was indistinguishable from the interview speech. But these spoken styles were dramatically different from the sung performances. Figure 1.1, reproduced from Gibson (2010b), shows these dramatic differences for one of the singers, Dylan Storey.

There are some differences between singing and speech which may relate to singing technique, notably a tendency for greater sonority, that is, more open vowels, in song. The role of singing-technique related factors is an important point to keep in mind in any study of the phonetics of song, as argued for by Andres Morrissey (2008). However, several of the differences are clearly dialect based, reflecting differences between NZE and American Englishes. DRESS and TRAP are raised in speech (and NZE), but open in song (and AmE). THOUGHT and LOT are both open and unrounded (and perhaps merged) in song (as they are in many dialects of AmE), while they are rounded and mid to high back vowels in spoken NZE. While GOOSE fronting occurs in both AmE and NZE, there are differences in its dynamism, with a fronting trajectory in NZE and a retracting one in AmE, a subtlety reflected in Dylan's speech and singing, respectively. An opposing direction of F2 movement is also seen in GOAT, a front-rising diphthong in NZE and a back-rising diphthong (or a high back monophthong) in AmE.

The comparisons between singing and speech were largely consistent across the three singers, despite their differences in genre (jazz-influenced pop, blues-rock, and indie folk) and despite differences in their intentions around identity projection. One of the singers said that he had never thought about his singing accent and had no desire to sound like a New Zealander in song, while the other two singers both stated plainly that they would like to use NZE in their songs and found it difficult to do so. Importantly, the dialect produced in song was strikingly similar despite these differing identity orientations. Of the two singers who stated having some desire to use NZE in their singing, one reported re-recording a particular song, line by line, multiple times, in order to produce a sung NZE style. The other (Dylan Storey) had specific awareness of one variable, the GOAT vowel, and it was in this vowel where he produced a small number of tokens where his singing and his speech matched phonetically. As can be seen in Figure 1.1, those tokens were eclipsed by the general tendency to use the AmE form once an average of all tokens was taken.

My previous work thus looked in detail at the phonetic differences between singing and speech for three NZ singers, and found evidence that an American-derived accent is default for them, while the use of NZE in song requires effort and awareness, and is reported as being 'difficult'. Those results suggested that my personal experiences were not unique, and that a degree of automaticity was involved in the production of AmE features in song for other NZ singers as well. The present project investigates the cognitive, social and linguistic processes behind this phenomenon, and examines the extent to which this pattern exists in the wider context of NZ popular music by building a corpus of songs. The corpus compares the singing styles of NZ and USA artists in commercial pop and hip hop genres (Chapter 2). This is followed by a novel method for investigating the dominance of AmE in NZ song: Chapters 3 and 4 explore the expectations of the general music-listening public in NZ about accents in popular music contexts.

The conclusions of my Masters work (Gibson, 2010b; Gibson and Bell, 2012) form the foundational assumptions of the present project:

- North American-influenced singing styles are so dominant in popular music that it requires conscious effort for New Zealanders to sing with the accent they use in speech.

- A levelled variety of American-derived English is the default for NZ singers across the full vowel space, not a stylisation restricted to salient variables.

- Use of NZE phonetic variants in singing, for example to project an 'authentic' identity, represents an initiative style-shift involving intention.

This thesis will further explore and verify these foundational assumptions, but the question which frames the present project is: *Why?* Why do New Zealand singers find it difficult to sing in a New Zealand accent? To begin to unpack this question, this introductory chapter reviews a range of literature relating to language cognition, and then considers the question from a sociohistorical and sociolinguistic perspective.

Since I take as my starting position the normative status of AmE in song, I prefer not to continue to refer to it with reference to 'America' since this is problematic from multiple perspectives. I put forward here the term Standard Popular Music Singing Style (SPMSS)[2], which gives this sung variety of English some room to move, without geographical constraints or potentially tenuous connections to specific dialects of US English. SPMSS exists alongside Hip Hop Nation Language (HHNL), which is derived from African American English (AAE) and has become an important part of hip hop culture world-wide (Alim et al., 2009). It should be emphasised at this point that the present project deals specifically with music that is commercialised and marketed. There are many genres and domains of singing that are not addressed by the research presented in this thesis.

The questions at the heart of this work are about how our experience with sound in the past shapes our processing of sound in the present. In both linguistics and the cognitive neuroscience of language, there is an increasing interest in the role of statistical learning in the language system. Across a wide range of disciplines in language science, there is a move away from 'the notion of an encapsulated language system' (Hasson et al., 2018, p. 151). In sociolinguistics, the move has come about through decades of research into the structured heterogeneity of language that gradually revealed the central role of indexicality in language processing. In language acquisition, studies revealed statistical learning by infants of transitional probabilities in the language signal (Saffran et al., 1996). In computational linguistics, rule-based approaches gave way to machine learning, forming statistical models based on large corpora. In cognitive neuroscience, 'neurobiological findings are in fact tearing away at the set of functions purportedly assigned to the classic language areas (or language module/system)' (Hasson et al., 2018, p. 151), increasingly showing that language functions involve widely distributed brain networks. All of these approaches culminate in the foregrounding of two important claims: language cannot be understood without also understanding general properties of learning such as memory and attention; and, language cannot be understood outside of its contexts of use.

The production and perception parts of this thesis are somewhat cross-disciplinary, involving first a sociolinguistic analysis of popular music (Chapter 2), and then psycholinguistic studies of speech perception (Chapters 3 and 4). The underlying research questions that drive those studies, however, are much more cross-disciplinary than that. The two parts of this thesis fit together because they both ask how contextualised memories shape linguistic representation and processing. To understand how sung memories might be stored, we[3] need to consider language and music

[2] This acronym builds on that used by Wilson (2017): CCSS — Classical Choral Singing Style.

[3] Any use of *we* in this thesis is meant as inclusive, either you, the reader and me, the author, or *we* as humans — or occasionally, as in this case, *we* as language researchers.

cognition.

1.2 Language and Music Cognition

> Language and music, two of the most unique human cognitive abilities, are combined in song, rendering it an ecological model for comparing speech and music cognition (Gordon, 2010, p. 1).

Singing is at once linguistic, physical, musical and cultural. It is involved in the emotional bonding of parent and child (Fancourt and Perkins, 2018), and may have been a key precursor to the evolution of language (Mithen, 2011). There is also robust evidence that musical training improves a range of cognitive abilities including language processing (Patel, 2012). Given its universality and centrality to human experience, and given its obvious connections to *the social*, it is surprising how little attention it has received in sociolinguistics. There is a strong tradition, however, for the study of language and music processing in cognitive neuroscience. While it is beyond the scope of this thesis to engage with this vast literature in any detail, I briefly touch on it here to provide a basic backdrop to the sociolinguistic and psycholinguistic questions addressed in this thesis.

Language and music share cognitive resources (Gordon, 2010). There is also evidence, however, for modularity of language and music, particularly through the study of brain lesions (Peretz et al., 2004), including cases that provide evidence of double dissociation between amusia and aphasia. Proponents of modularity of language and music suggest that any shared cognitive resources are domain-general mechanisms such as attention (Schellenberg and Peretz, 2008), but as described above, there is increasing evidence that even the supposedly 'core' language functions also draw on widely distributed functional networks. Whether music and language utilise fully distinct brain regions or not, melodies sung with lyrics are a special instance combining music and language cognition. One study of this combination of music and language concluded that 'the text and the melody of a song have separate representations in memory, making singing a dual task to perform, at least in the first steps of learning' (Racette and Peretz, 2007, p. 242). There is also evidence that processing musical information may incur a cost on the processing of denotational aspects of language. In an experiment that required participants to remember either the melody, the words, or both, when exposed to short spoken and sung phrases, van Besouw et al. (2005) found 'a decrease in the amount of linguistic information retained by subjects for sung phrases' (p. 129).

Singing and speech are likely to be associated with different cognitive processes for a number of reasons. They involve systematic differences at the acoustic level, for example, through the co-occurrence of musical instrumentation with singing. There are also social and functional differences between the two registers, one of which is the degree to which the speaker/singer wishes to communicate referential content. Many experiences with singing are mediated, with no potential for interaction or reply, despite the potential for strong emotional engagement of a listener with the singer's voice. A greater focus on form than meaning in singing than speech may divide these sub-systems at a neural level, with the former involving a reduced urge to communicate. Most work on the intelligibility of lyrics has involved classical singing styles, and shown that intelligibility is much lower for singing than speech

(Collister and Huron, 2008), though these rates are conditioned by similar factors as is intelligibility of speech (Heinrich et al., 2015).[4] A focus more on form than meaning in song might lead to a greater focus on surface features of the voice, which increases the effects of episodic memory on subsequent recall (Goldinger, 1996b).

Peretz and Coltheart (2003) present a modular model of speech and music processing. Through a range of processes such as acoustic to phonological conversion, rhythm analysis and pitch contour analysis, acoustic input makes contact with a 'musical lexicon' and a 'phonological lexicon'. It is at the stage of these lexicons that associative memories are allowed to play a role. This model does not fit well with findings in the literature in sociophonetics and laboratory phonology (to be reviewed below) which show that associative memories allow us to bypass 'conversion' processes altogether. Categories emerge through co-occurrence patterns between associative memory, acoustic input, and the lexicon (e.g. Pierrehumbert, 2016; Drager, 2010). Rather than separate phonological and musical lexicons, the storage of words encountered with and without musical elements could form neurally distinct clusters of memory traces.

1.3 High-Fidelity Imitation and the Ritual of Song

The tension between structure and agency is a core issue straddling sociolinguistics (Bell, 2001) and language processing — speakers are subject to structural forces through their past experience with language, but also have agency, using language to achieve identity goals. The pressure to conform is evidenced across multiple human activities, including language: 'people adapt their speech patterns to their speech community even without overt pressures and rewards' (Pierrehumbert, 2001, p. 13). Despite some claims to the contrary (Trudgill, 2014), there is evidence that phonetic convergence with interlocutors involves social attitudes (Babel, 2012). A controversial question is whether such adaptation occurs in the absence of face-to-face interaction, though perceptual studies suggest it might (see Pardo, 2013, for a review).

Imitation, of course, extends far beyond language. There is a clear role for face-to-face transmission of detailed behaviours in human evolution, resulting in the 'cumulative culture' responsible for the development of tools across millenia. Legare and Nielsen (2015, p. 689) state that 'learning social conventions requires close conformity ... through high fidelity imitation'. This relates to the suggestion that 'mirror neurons evolved to support an abstract manual gestural system that was then adapted to vocal tract behaviors' (Hickok, 2010, p. 3). High-fidelity imitation can be causally opaque, not having a clear effect on the world, but performed all the same due to 'a willingness to rely on faith in cultural traditions' (Legare and Nielsen, 2015, p. 690).

Imitation is thus the stuff of ritual, and according to Watts and Andres Morrissey (2019, p. 35), so is song: 'all forms of music, hence all song genres, ultimately derive from the symbolic container of ritual'. 'Rituals are consensual group behaviors that frequently involve group coordination and synchrony' (Legare and Nielsen, 2015,

[4]Note, however, that Potter (1998) claims that the urge to communicate in song actually increased in the second half of the twentieth century as singers moved away from classical singing styles, allowing 'a return to singing as carrier of text, a vehicle for the articulation of meanings' (p. 189).

p. 692), and the rhythmicity involved in music-making makes it a particularly good tool for coordinated joint action (Hawkins, 2014). There is also recent evidence for complex patterns of speech convergence and divergence to task-irrelevant instrumental music, at least with respect to pitch, intensity and speech rate (Podlubny, 2019). Whether the instinct towards high-fidelity imitation applies to the case of listening to a recorded song over and over is something many would question. Trudgill (2014, p. 220, emphasis in original) states:

> ... it is the *interaction* itself which is crucial, and not the repeated exposure. Repeated exposure does indeed come from the media. But one does not interact verbally with the TV. Lexical change does not require interaction; but we have been presented with no convincing evidence so far that core phonological and grammatical change can be diffused without it. To repeat Labov's point, diffusion is purely a matter of who interacts most often with who.

The subject matter of this thesis is not diffusion of phonetic variants from one speech community to another, so much as it is a case study in mediated second-dialect acquisition. This acquisition of the song register[5], in all its phonetic richness, relies most definitely on repeated exposure, but it may also involve a kind of unusual spatially and temporally displaced verbal interaction, in the form of 'singing along'. For this, our evolved endowment for high-fidelity imitation may be quite useful. This brings us to the relationship between production and perception, which has recently been modelled quite explicitly in the context of exemplar theory.

1.4 Exemplar Theory and the Importance of Context

> [A] word's production and perception will be influenced by the full range of contexts (linguistic, social, environmental) in which we have previously encountered it. (Hay, 2018, p. 6)

There is now a great deal of evidence for the role of episodic memory in the tracking of co-occurrences between linguistic and social patterns, and the fine phonetic variability of words (much of which will be introduced in the literature reviews contained in the following chapters, but which includes for example Strand, 1999; Seyfarth, 2014; Hay and Foulkes, 2016; Hay et al., 2018). This section walks through some of the basic concepts of exemplar theory, leading to the way it handles the link between production and perception, which motivates my attention to both the production and perception of song in this thesis.

Episodic memories allow us to transport ourselves to a specific moment experienced in the past, and essentially 'relive it'. They are distinct from our *explicit/semantic memory* for facts about the world, or the *implicit/procedural memory* that allows us to ride a bike (see, e.g. Tulving, 2002). Since language forms a prominent part of our experiences from the very beginning of our lives, we have many, many episodic memories of speech. Hybrid models of language processing

[5] I use the term register to refer to the distinction between song and speech. For more on the concept of register, see Halliday (1978); Finegan and Biber (1994); Agha (2003).

(Pierrehumbert, 2006, 2016) assert the importance of both episodic memories, rich in detail, and semantic knowledge abstracted from regularities across such memories. A word is an example of such an abstraction: 'The lexicon and the grammar ... represent two degrees of generalization over the same memories and are thus strongly related to each other' (Pierrehumbert, 2001, p. 2). A word connects semantic knowledge to a cluster of acoustic patterns that bear some similarity to one another, as well as the concatenation of a series of abstract sound categories, themselves based on remembered clusters, organised by similarity. Determining acoustic similarity is thus a central concept:

> In an exemplar model, each category is represented in memory by a large cloud of remembered tokens of that category. These memories are organized in a cognitive map, so that memories of highly similar instances are close to each other and memories of dissimilar instances are far apart ... The entire system is then a mapping between points in a phonetic parameter space and the labels of the categorization system. (Pierrehumbert, 2001, pp. 3–4)

A word is connected to non-linguistic information, within episodic memories of specific experiences. Just as the /æ/ phoneme is an abstraction from a cluster of remembered sounds, the labels for 'male' and 'female', or 'inside' and 'outside' are abstractions emerging from a multitude of episodic memories. This non-linguistic parameter space would include all the kinds of abstractions employed by sociolinguists to described social meanings: stances (Jaffe, 2015), characterological figures (Agha, 2005) and macro-social categories (Labov, 1966). These socio-contextual categories co-occur with phonetic variants at different rates, and it is through such co-occurrences that the mind forms expectations about the types of phenomena most likely to occur in any given situation.

Such representations can be conceived of as being encoded and recalled in the form of a neural 'fingerprint' — a distributed network of the sensory input and internal states that co-occurred with the word at the time of its perception. This is likely to hold some degree of neurophysiological validity, and has been explored in fMRI studies showing the reactivation of brain states associated with context-specific experiences with a stimulus (Danker and Anderson, 2010). Another striking fMRI finding showed the existence of reliable neural signatures for semantic fields (Huth et al., 2016), suggesting the kind of similarity-determined cognitive map that is a core tenet of exemplar theories.

Foundational perception experiments arguing for the retention of acoustic detail in memories of language stimuli showed that speech recognition in adverse conditions is aided when a repeated stimulus occurs in the same voice as it was originally heard (e.g. Palmeri et al., 1993). Pufahl and Samuel (2014) extended this finding to non-linguistic incidental co-occurring auditory input, such as a phone ringing or a dog barking. A word is processed more accurately on repetition if the exact same dog bark or phone ring is repeated, and the inverse relationship is also true, showing that 'representations stored in the mental lexicon are not limited to linguistic information, nor are they limited to the addition of information from highly related sources like voices. Instead, these representations appear to reflect more episodic traces of words and co-occurring auditory events, even from unrelated sources like background sounds' (p. 28). This finding is clearly of central importance to the present project, given the fact that popular music is consistently accompanied by co-occurring auditory events that are highly systematic, having predictable structures connecting the voice to the instrumentation, and the harmonic and rhythmic structures of one song to those of another.

1.4.1 Speech production and speech perception

> [A]uditory–motor interactions in the acquisition of new vocabulary involve generating a sensory representation of the new word that codes the sequence of segments or syllables. This sensory representation can then be used to guide motor articulatory sequences (Hickok and Poeppel, 2007, p. 399)

Memories for words are distributed across many brain regions including sensory, executive and limbic areas. Amongst these, the motor cortex plays an important role. Given the striking finding that monkeys have difficulty forming long term memory for auditory stimuli (but not visual or tactile stimuli), Schulze et al. (2012) argue that our ability to store detailed memories of sound is highly dependent on the oral-motor skills associated with language.[6] Since detailed, enduring memories of sound are at the heart of exemplar theories of language, the role of the motor system in the formation and maintenance of memories of speech cannot be ignored. In an important early paper applying exemplar theory principles to language processing, Johnson (1997, p. 154) stated that ‘the production–perception link is based on one’s own speech’, but also argued that when processing the speech of others we form ‘gestural mirages’, that is, we imagine the gestures which might have led to the acoustics produced. This approach is influenced by the motor theory of speech perception (Liberman and Mattingly, 1985), but departs from it by emphasising the acoustics of speech, and considering speech production to involve acoustic targets. The importance of the motor system to auditory learning is also now beginning to receive attention in cognitive neuroscience (Kraus and White-Schwoch, 2015).

The production–perception loop is shaped by one’s own production, and by phonetically detailed memories of the speech of others. The combined cluster of memories can be thought of as a cloud of exemplars, which inform the selection of targets for further speech production. Pierrehumbert (2001) presented the first application of an exemplar theory approach to speech production, in which ‘production of the phonological category represented by any specific label involves making a random selection from the exemplar cloud for that label’ (Pierrehumbert, 2002, pp. 114–115).

While always assumed to be present, exemplar models still have not satisfactorily implemented the ‘hidden systematicity which a more complete model should capture’ (Pierrehumbert, 2002, pp. 115). In an ideal model, a cloud of acoustic phonetic memories should be viewed as having connections to myriad labels, each of which emerged out of exemplars of its own. These labels are abstractions based on patterns of co-occurrence amongst and between both linguistic and non-linguistic experiences. In order to make modelling feasible, this multi-dimensional property is stripped from the model: ‘The aggregate effect ... is random variation over the exemplar cloud’

[6] Though note the wealth of evidence (reviewed by Hickok, 2010) demonstrating that the ability to produce speech is not a necessary condition for speech perception.

(Pierrehumbert, 2002, p. 115). If included, the structured heterogeneity would allow the speech production system to pick a target not across the full set of exemplars for a phoneme, but from a sub-cloud of exemplars that sit at a point in multi-dimensional space most closely aligned with the currently activated cognitive scene.

1.4.2 An implementation of the production–perception loop

> [S]ound change is as much in the ear of the listener as in the mouth of the speaker (Harrington et al., 2019, p. 3)

Though early computational implementations of an exemplar approach tended to focus on either perception *or* production, the two have recently been combined in a single implementation (Todd et al., 2019). This implementation of exemplar theory explicitly models the production–perception loop. It includes mechanisms by which high frequency words can lead in one type of sound change and lag in another, and change at the same rate as lower frequency words in yet another type of sound change. These simulations successfully emulate the empirical results of three different types of documented sound change. There are adjustable parameters on both the production and perception sides of transmission, each in turn affecting the storage and updating of exemplars. The model does not implement any social or contextual weighting, but this is a possible next step. I will thus summarise the model here, but with the addition of a basic conception of how context, defined here as singing vs. speech, might be included in a future implementation. Figure 3 from Todd et al. (2019) is reproduced in Figure 1.2.[7]

The Todd et al. (2019) model begins with a set of exemplars in the mind of a speaker-listener. An instance of language use begins by selecting a type (a word, perhaps the word *bed*, made up of both a DRESS vowel phoneme[8] and its surrounding /bVd/ phonetic frame). This is followed in the original model by selection of a target (one of the remembered exemplars of that word). Between these steps, we could add an additional step to determine whether the situation for speech production is Context A or Context B. This would add another dimension of detail to be encoded with each exemplar, giving each one: an acoustic phonetic property (represented as colour gradient), a phonemic category (represented as the colour of the shape's outline), a phonetic frame (represented as shape), and a contextual category (represented, perhaps, by size). For the purposes of this discussion, Context A can be 'conversation' and Context B can be 'song'. In this case, the speaker-listener wants to sing the word *bed*, so they select specifically from the sub-cloud of their memories of *bed* that were also sung. Importantly, the selection ignores encounters with *bed* in conversation.

Once a production target has been selected, the target vowel quality is subject to bias and imprecision, and then the resulting acoustic form is produced and trans-

[7] Todd et al. (2019) is under a Creative Commons Attribution-NonCommercial-No Derivatives license (CC BY NC ND), the details of which can be viewed at `https://creativecommons.org/licenses/by-nc-nd/4.0/`. Any further reproduction of this image must comply with the terms of that license. In addition to the figure itself, I quote here the part of the original caption relevant to my discussion, for the full caption detailing other aspects of the model than the exemplar store, see the source.

[8] Throughout this thesis I refer to vowels by the name of the lexical set to which they belong. DRESS, here, is a way of referring to the short front vowel /e/. For an introduction to lexical sets, see Wells (1982).

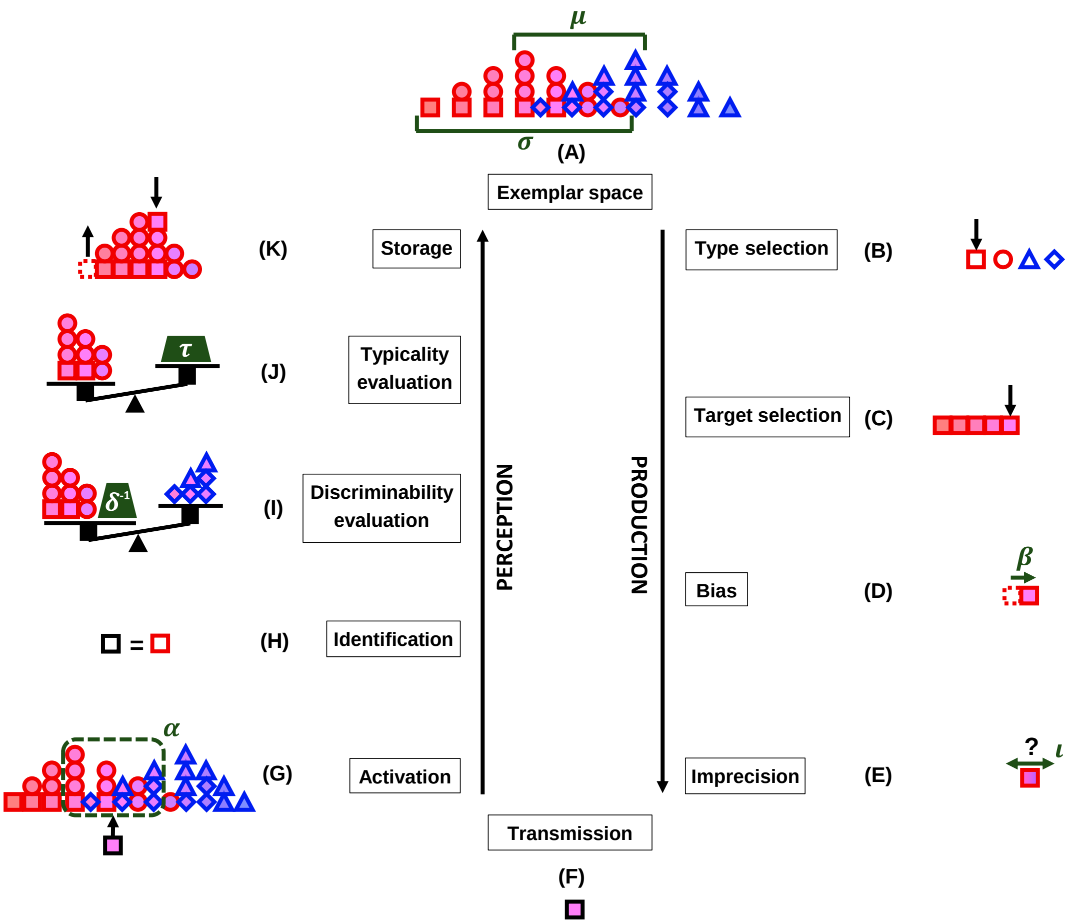


Figure 1.2: Reproduction of Figure 3 from Todd et al. (2019, p. 7). 'Schematic illustration of processes in the model, forming a closed loop between production and perception. Outline colors represent phoneme category membership (e.g. /æ/), shapes represent phonological frame (e.g. /m_p/) – so that colored shape-outlines represent types (e.g. "map") – and fill colors and horizontal positions represent perceptual-acoustic value (e.g. vowel F1). Dark green components with Greek letters indicate parameters of the model. (A) Two partially-overlapping categories exist in an exemplar space...' (partial caption from Figure 3 of Todd et al., 2019, p. 7). This figure is reproduced under a Creative Commons Attribution-NonCommercial-No Derivatives license (which can be viewed at `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

mitted. This acoustic form is represented by just one dimension in the model, in this case let's call it F1. On the perception side, a window of exemplars is activated by the acoustics of the incoming vowel. Exemplars with an F1 similar to the input become activated. Some of these exemplars would be sung DRESS vowels, many would be conversational TRAP vowels[9], and there would also be activation for memories, for example, of sung *red* and spoken *cat*. The word *bed* is identified using the phonetic frame[10], and the context is also identified, perhaps using the presence or absence of background music in the acoustics. After this identification process, the token is evaluated with respect to the other activated exemplars in the acous-

[9]See Section 3.2 for the empirical motivations behind this example.
[10]This is possible in the current implementation because it does not allow for minimal pairs.

tic space, and if the token is acceptably discriminable and typical (both based on adjustable parameters), it is stored in memory, and will affect the next iteration of production and perception. With respect to the computational implementation of exemplar theory, this model is a promising step forward.

This description of exemplar theory may seem perhaps too a-linguistic, robbing language of its special status in the mind. After all, auditory language processing involves incredibly fine-grained judgements of acoustics (and gestures), and the number of abstractions required to build a phonology, let alone a lexicon, is massive. As discussed above, however, there is increasing evidence that language can indeed be managed by general learning mechanisms: Naive Discriminative Learning models (Baayen et al., 2011), for example, 'provide a clear demonstration of how a powerful and rudimentary learning mechanism can explain phenomena that have been typically addressed by linguistic (language-essential) formalisms' (Hasson et al., 2018, p. 147). Perhaps the granularity of abstraction is coarser in the non-linguistic dimensions of these indexicality matrices, despite being supported by the same underlying adaptive learning processes. This is an issue that will, hopefully, become less ethereal and more testable, through the joint efforts of computational linguistics and cognitive neuroscience. Behavioural psychology (including the methods employed in sociophonetic research) and sociolinguistics also have roles to play, and this project will advocate for a high degree of specificity and gradience in our conception of non-linguistic abstractions, enabling precise relationships to build up between phonetic forms and the cognitive scenes in which they occur.

To close this section, before turning from the cognitive to the social, I return once more to Hasson's 2018 review article cited above. The central focus of the paper is 'the necessity of taking context as a serious topic of study, modeling it formally and acknowledging the limitations on external validity when studying language comprehension outside context' (p. 135).

> In other words, the co-occurrence of elements, irrespective of the 'cognitive category' one assigns them (lexeme, phoneme, visual element), could serve as foundation for a comprehensive statistical model of language representations. Although there are clearly levels of representation where these are distinct units, for purposes of neurobiological computations underlying language comprehension, such multi-modal elements may be bound together. (Hasson et al., 2018, p. 146)

This is a good summary of exemplar approaches to language perception and production, even though it comes from a neurobiology perspective. It suggests that increased collaboration between computational linguistics, laboratory phonology and cognitive neuroscience would be fruitful. No matter how well we model memory — the structures of the past — we will ultimately have to also incorporate the role of speaker agency in any complete model of language processing. This is where sociolinguistics is strong.

1.5 A Sociolinguistics of Popular Music

Through a wide range of work from the 'third wave' (Eckert, 2012), we have learnt that language forms are related to many social meanings, and these meanings shift dynamically in response to context. The speaker has a range of identity and interactional goals which they achieve through the construction of styles and stances, through a bricolage (Hebdige, 1979; Eckert, 2000) of elements including phonetic styles, but also extending far beyond into a range of other semiotic modes. When agentively constructing a style, the speaker takes *initiative* to shape the language situation, rather than simply taking a *responsive* approach (Bell, 2001). In an exemplar model such as the one by Todd et al. (2019) described above, it might be possible to include a step in which the motivations of the speaker affect target selection. These could be modelled by allowing the speaker to pay attention to any dimension of context represented in the current scene, or indeed to call forth categories from memories not represented in the current scene. Sampling for the speech production target could come from memories of characterological figures (Agha, 2005), for example, and would result in referee design (Bell, 1984).

There is a constant tension between contextual norms and speaker motivations, represented through dualisms such as structure vs. agency (Giddens, 1979); responsive vs. initiative (Bell, 2001, which presents a fuller summary of these dichotomous terms in sociolinguistics); imitation vs. innovation (Legare and Nielsen, 2015); and centripetal vs. centrifugal (Bakhtin, 1981). For both sociolinguistic and psycholinguistic reasons, we automate some language behaviours, leading to *linguistic routines* (Hymes, 1968). These vocal habits reflect the manifestation of societal structures in the individual, as a form of *habitus* (Bourdieu, 1991). In the chapters that follow, we see that convention looms large in the context of commercial popular music.

1.5.1 (The lack of) a dialectology of popular music

No systematic dialectology of English popular song has yet been attempted. The dialectology of spoken English, of course, goes back centuries, and provides the basis upon which sociolinguistic analyses are able to occur, through the demarcation of geographically-defined speech communities. The sociolinguistics of popular song still resembles a patchwork. One of the most advanced areas of research is on the global spread of hip hop and its language forms (Pennycook, 2007). This work tends to border sociolinguistics and cultural studies, as does a recent collection on language in popular culture (Werner, 2018). From a more variationist perspective, the foundational study of the pronunciation of English in popular music singing (Trudgill, 1983) identified the use of ‘American’ variants in the songs of British musicians in the 1960s and 1970s, and treated these shifts as a ‘symbolic tribute to the origins of popular music’ (Konert-Panek, 2018, p. 155). Trudgill considered the adoption of American variants to be primarily a matter of intentional (and sometimes inaccurate) imitation, while more recent evidence (Beal, 2009; Gibson and Bell, 2012) suggests that these shifts happen largely unconsciously, and that the use of one’s ‘own’ phonetic style in song, even when desirable from an authenticity perspective, takes effort and conscious control.

Coupland (2011) cut across much of this research by suggesting that these studies have been limited by an overly literal interpretation of ‘place’ as region or nation. He theorised popular song as a ‘field of performance organised according to genre’ (Coupland, 2011, p. 573)[11], where place is understood as a specific socio-cultural

[11] Terminological note: I use genre throughout this thesis to refer to musical style. Other authors (e.g. Squires, 2018) treat singing as a genre in its traditional linguistic sense, as distinct for example

context. This reframing opens up the possibility of a dialectology of popular song organised primarily by genre rather than geography. While I do foreground genre in my analysis of the current prevalence of SPMSS in NZ popular music, the variationist approach taken is perhaps still not quite in the spirit of Coupland's suggestions. Before moving to the linguistic detail, I consider first the social and historical conditions surrounding the emergence of recorded popular music, in particular, the early dominance of music from the USA. I put forward here an analogy between the establishment of popular music singing accents and the development of post-colonial Englishes (Schneider, 2003, 2007). The basic idea is that early US cultural dominance established a kind of founder effect (Mufwene, 1996). The first dialect to get to the metaphorical 'island of pop music' was the one to establish long-lasting conventions. The unique potential in commercial popular music is that its entire recorded history is available, waiting to be analysed.

1.5.2 Early sound recordings

Hickey (2017) argues for the importance and potential to sociolinguistics of studying early recordings. The benefits of such an approach have perhaps been best demonstrated through the Origins of New Zealand English (ONZE) project (Gordon et al., 2004; Hay et al., 2015) which includes recordings that span much of the development of NZE. This provides a unique opportunity to address questions of dialect formation with respect to both social (Gordon et al., 2004; Schneider, 2007), and non-social (Trudgill, 2004) processes. The earliest recordings of the human voice, however, date back much further than the earliest ONZE recordings. While the first recording of all was made in 1860, the mass distribution of recorded sound as a form of novelty entertainment began with Thomas Edison's phonograph, which was first heard in NZ in 1879 (Hoar, 2012).

From very early on in the development of sound recording technology, reproduction of song was placed at the centre of the enterprise. This makes a diachronic dialectology of singing styles invitingly possible. Recorded popular music, by its very nature, is available for analysis from its conception through to the present time, offering an extremely valuable corpus of data which has barely been mined in the sociolinguistic literature, in part because performed language was traditionally shunned in favour of the vernacular. Performed language is increasingly considered to be an important tool for understanding sociolinguistic processes (see Bell and Gibson, 2011b, for a review). While it is not my intention to analyse the path by which American dialects took centre stage in recorded music, it is worth briefly considering an early example.

Henry Burr (the stage name for Harry McClaskey, who was born in 1882 in a border town in New Brunswick, Canada) is described as 'The Original King of Pop' in the title of a collection album released by Archeophone Records (ARCH5502), and was a prolific early star of the American recording industry. These songs recorded in the first and second decades of the twentieth century are striking — they were made long before the style so dominant in popular music today took root. BATH is realised with [ɑː] (e.g. *last* in 'The Holy City'), and /r/ is categorically tapped or trilled in all environments: syllable onset, intervocalically and in non-prevocalic position. Though outside of my present scope, an analysis of how and when North

from a sermon or sports commentary.

American dialects began to supplant the early adherence to British standards would help to shed light on the questions asked in this thesis.

1.5.3 Cultural Imperialism and the Dynamic Model

The current dominance of SPMSS in popular song has parallels to British colonialism and the formation of post-colonial Englishes (Schneider, 2007), but in the cultural sphere. In the context of NZ music specifically, Shuker and Pickering (1994) define *cultural imperialism* as 'a concept analogous to the historical political and economic subjugation of the Third World by the colonising powers in the nineteenth century.'

> This economic and political imperialism also has a cultural aspect: 'namely the ways in which the transmission of certain products, fashions and styles from the dominant nations to the dependent markets leads to the creation of particular patterns of demand and consumption which are underpinned by and endorse the cultural values, ideals and practices of their dominant origin' (Shuker and Pickering, 1994, p. 277, quoting O'Sullivan, 1983, p. 62)

Through dominance beginning in the early stages of recorded popular music, the USA became and remained the centre of commercialised culture. The American-derived varieties of English used in mass-distributed recordings took root as part of the aesthetic of rhythm&blues, country, jazz, and rock&roll. In the following paragraphs, I adapt the logic of Schneider's 2007 Dynamic Model of Post-colonial Englishes to the history of recorded popular music, adapting quotes from his model to this analogy using square brackets. We can think of the infiltration of recorded songs as the 'settler strand', and music listeners around the world from all kinds of language backgrounds as the 'indigenous strand'.

In Phase One, Foundation, people outside of the USA begin to use AmE in the context of songs. They sing along to hit singles, and covers bands spring up, who faithfully reproduce those hits. Phase Two, Exonormative Stabilisation, refers mainly to the settler strand. In this contortion of the model to fit cultural imperialism rather than physical colonialism, aspects relating to the settler strand's motives and adjustments are ignored, since the flow of linguistic influence is effectively unidirectional.

In Phase Three, Nativization, 'a new identity emerges': non-US music listeners 'realize that something fundamental has been changing for good: traditional [ways of singing songs] are discerned as no longer conforming to a changed reality, and the potentially [error-prone] process of gradually replacing them with something different ... is in full swing' (p. 247). Singers in both English speaking and non-English speaking countries are at the forefront of this process. They 'undergo a process of linguistic and cultural assimilation and large-scale second language [or dialect] acquisition', all of this strictly in the context of listening to, and performing popular song, while conversational speech remains unaffected. At this stage, rather than just adopting US hit songs, singers extend AmE to their own compositions. Musically and phonetically, they create a coherent style across original and adopted songs.

Towards the end of the Nativization phase, there are the first signs of independence:

> many countries gain [commercial] independence [by producing highly successful local artists] but retain a close bond of cultural and psychological association with the mother country, a process that results in a kind of 'semi-autonomy' in their identity construction ... When the 'mother country' is felt to be less and less of a 'mother', the offspring will start going their own ways, [musically] and linguistically — slowly and hesitantly at first, gaining momentum and confidence as time passes. (p. 247)

As described by Trudgill (1983), there were several signs of this burgeoning independence in the British music scene in the 1970s, particularly with the huge commercial success of The Beatles, but also through the emergence of the punk movement. In New Zealand, such signs took longer to emerge, with the Flying Nun bands of the 1980s being an obvious sign that independence was emerging. However, as we approach the third decade of the twenty-first century, SPMSS remains extremely dominant. There continues to be an unbalanced (though no longer uni-directional) flow of cultural content from the 'mother country' outwards to smaller music consumer markets. And despite the emergence of mass decentralisation in content creation, a quick search of young singers from outside of the USA on YouTube will support the idea that the cultural colonies remain in Phase 3.

> The fourth phase, Endonormative Stabilization, is marked by the gradual adoption and acceptance of an indigenous linguistic norm, supported by a new, locally rooted linguistic self-confidence. (p. 249)

In various hip hop communities around the world, we see evidence of Phase Four (Pennycook and Mitchell, 2009; Mitchell, 2008; Williams, 2017). There is language and dialect mixing and the dichotomy of global vs. local is broken as communities accept the transcultural nature of their situation, and embrace the 'glocal' (Mitchell, 2008). O'Hanlon (2006) provides good evidence that Australian hip hop has reached Phase Four (though see comments on ethnicity in Section 2.5.4), but such independence is by no means universal (Stæhr and Madsen, 2017). Hip hop could also be seen as having its own separate history, since it has a different cultural centre, and a different timeline to the massive spread of, for example, jazz, or rock&roll in the 1950s.

Phase Five, Differentiation, may still be far off. As Schneider (2003, p. 253) describes, a Phase Five community will have freed itself from the 'external dominant source of power and orientation', having developed 'an attitude of relying on one's own strengths, with no need to be compared to others. As a reflection of this new identity, a new [sung] language variety has emerged'.

Obviously, there are many ways in which the parallels to Schneider's model are a stretch. But it provides a framework for thinking about singing accents, as they are now, in their historical context, both looking backwards and forwards. As I see it, there are two potential outcomes. Firstly, if the model does in fact stretch as far as the mediated spread of a context-specific register, then more and more communities of practice (Wenger, 1998) within the global music industry may enter Phase Four. This would mean the stabilisation of singing styles where local speech varieties either replace or blend with SPMSS, as we have seen in some hip hop communities. These need not map onto any region-based spoken dialects, but would rather represent the formation of communities with structure at both global and local levels, along the lines of the Hip Hop Nation (Alim, 2009). Once stable, these varieties may then extend to Phase Five and exhibit greater diversity along sub-cultural stylistic lines.

Alternatively, it may be that the paths of influence are now so pervasive that SPMSS has reached a stable equilibrium as a cultural lingua franca for popular song. It may be that 'own-accent' singing (a concept I introduce below) will continue to be an oppositional practice involving effort and awareness. Some of these vocalists will make a dent in the SPMSS fortress, but through their very act of differentiation will become associated (both indexically in the minds of listeners and overtly through marketing) with a particular musical sub-genre, and thus marginalised, allowing SPMSS to stabilise in the mainstream and continue on, with global reach.

One important point to emphasise is this: singing in popular song and conversational speech are so contextually divorced from one another that a complete overlap between the two, even in small tight-knit communities of practice, is unlikely. Only through a concerted effort by a singer, likely compelled by some ideology of authenticity (discussed in the next section), will the two be forced into alignment now that the inertia of US-accented recorded music has become so strong. Unlike in the face-to-face community, the voices of the past do not disappear after a season. They can continue to have an influence as long as there are ears to listen (and technology to reproduce recordings).

1.5.4 Genre norms, authentication practices and awareness

> [I]t is an ideological premise of the folk/country genre ... that 'character' and 'person' should not be distinct identities – that is, that the singer should sing sincerely as him/herself. (Coupland, 2011, p. 591)

Through a process of indexical bleaching (Squires, 2014) which will be described in more detail in Chapter 5, singing accents no longer 'sound American' — place meanings have been backgrounded in the context of song. They only resurface now when a distinction between a singer and their performance persona is highlighted. Such a distinction is well captured by the concept of role distance (Goffman, 1981, p. 144), particularly with respect to the roles *author* and *animator*. In an opera or a Broadway musical, there is clearly a long distance between the author (the composer/lyricist) and the animator (the singer). During the 1960s musicians began to emphasise their role as creators of song, rather than interpreters of song (Potter, 1998). Put another way, they began to see it as desirable to reduce the large author–animator role distance that had been typical in popular song. The Beatles, Potter (1998) argues, were the first major pop group to emphasise themselves as writers in this way. While it is generally accepted in the literature that authenticity should be seen as something people do, not something people have (Bucholtz and Hall, 2005), authenticity, in both cultural studies and amongst musicians themselves, is still often framed in essentialist terms (as attested by Harrison, 2008). This 'ontological' way of viewing authenticity (Coupland, 2003) revolves around 'being yourself'. The minimisation of role distance in song thus becomes an authentication practice ('a social process played out in discourse', Bucholtz and Hall, 2005, p. 601), a way to express a close relationship between a singer and their song.

Other genres, notably punk, emphasised anti-mainstream stances, and place and class meanings were foregrounded to demonstrate opposition to the homogeneity of popular music. Hip hop emerged as a genre which emphasised both of these ideologies: the authentic representation of self, and resistance against the mainstream. Folk music (as a marketed genre, not in the sense defined by Watts and Andres Morrissey, 2019, discussed below) also adopts this ideology. Coupland (2011) provides a nuanced discussion of this process of 'fusing person and character' with respect to James Taylor:

> Taylor performs sincerity in allowing us to learn or infer details of his autobiographical 'person' through his lyrics. (p. 591)

A likely consequence, for a singer, of trying to fuse person and character is that they will become aware of features of their phonetic style that differ between singing and speech. Throughout this thesis, I use the terms 'own-accent' singing or 'own-accent' rap to refer to the process by which a singer transfers a phonetic feature from their speech style to their musical performance. Inherent in this term is the ideological construct of the 'authentic self'. There has been a gradual trend towards own-accent vocal styles in the genres discussed above, and occasionally in mainstream pop. However, as this thesis will demonstrate in the context of NZ music, the SPMSS norm is still dominant.

Awareness is a crucial part of authenticity, since it involves the application of ideological motivations. The concept of awareness is closely related to the concept of salience. Rácz (2013) provides a very careful description of salience as it is used in different disciplines. In the visual cognition literature, salience refers to bottom up effects on attentional direction, and is generally restricted to intensity and local contrast effects, with little reference to prior experience (Rácz, 2013). A salient linguistic variable, by contrast, is defined by its use for social indexation, either with or without consciousness. There is a common thread to the various definitions of salience, however:

> In both cases, an entity juts out due to its dissimilarity to the rest of the structure. Dissimilarity is not an inherent property. It is assessed in comparison with the rest of the structure. (Rácz, 2013, p. 50)

1.5.5 Defining the scope of 'popular song'

Recall that my focus in this thesis is on the singing that is commercialised and marketed. It is a huge industry, and as I will demonstrate, it is worthy of sociophonetic analysis. However, commercial music only comprises a subset of 'song', which ranges from the intimacy of a lullaby to the symbolism of a national anthem. There are many types of singing which do not use SPMSS, not through opposition, but simply through its irrelevance to their artistic tradition. Reggae and choral music are two such exceptions. Reggae has a specific regional context, closely tied to Jamaican culture. This type of well-defined boundary around both the musical style and other characteristics can lead a genre to differ in language style from the American-influenced norm, and indeed become a cultural centre in itself. Reggae artists from outside of Jamaica, for example, routinely adopt a Jamaican phonetic style in song. Gerfer (2018) provides evidence of this, showing the use of phonetic, morphosyntactic and lexical features of Jamaican Creole and Jamaican English by non-Jamaican reggae artists, though Westphal (2018) suggests that such uptake of Jamaican features is based on a restricted range of stereotyped features, rather than the kind of native-like command described by Gibson and Bell (2012) for New Zealanders' use of SPMSS. There is also evidence of an inverse direction of influence in work by Wilson (2017) that shows choir conductors and singers adopting Southern British English features (alongside some use of Trinidadian Standard English) in the context of choral music (which she dubs CCSS, classical choral singing style).

One of the most important exceptions to the type of data I look at here is 'folk song'. Watts and Andres Morrissey (2019, p. 109) stress that 'folk song is not commercially motivated'. Their definition of folk song is strong on process not product — folk song is *musicking* not music — and strong on function. Folk song is:

> 'Any singing activity used originally to create, constitute and construct a group of people, a "folk", to bond them together for a space and time and in a specific location. (p. 11)

This embodied and social approach to singing is indeed very distant from the mass commercialisation of what I call 'popular song', which *is* focused on product, in the form of the textual artefacts of recorded songs.

1.6 Assumptions, Questions and Hypotheses

Returning to the core issues introduced at the start of this chapter, I now present in more detail how the conclusions of my previous work (Gibson, 2010b; Gibson and Bell, 2012) inform the present project. Three key assumptions can be summarised as follows:

- An 'American-influenced' accent is normative in NZ popular music.
- It occurs in both salient and non-salient variables for most NZ singers.
- It takes effort and awareness for a New Zealander to use NZE features in song.

This quote from Dylan Storey, one of the NZ singer/songwriters analysed in Gibson (2010b), encapsulates the issue:

> Once you start thinking about it ... it starts to become painful to blatantly sing American vowels, but going the other way is quite difficult too, you have to be really conscious.

Despite a desire to project an authentic persona in song, the acoustic analysis of Dylan Storey's singing showed that the use of NZE is indeed 'difficult' for him, at least in the sense that he produced NZE variants rarely, and only in variables where he reported conscious awareness of the distinction between NZE and the American style. This thesis builds on these prior findings by considering the cognitive and sociolinguistic reasons for this phenomenon, and is framed by the following question:

Why might it be difficult for New Zealanders to sing the way they speak?

The literature reviewed above gives an outline of some of the many factors which I see as providing insights into the above question. In this thesis, I search for answers by exploring the role of episodic memory in language cognition. Drawing on both exemplar theory and sociolinguistic studies of style, I present specific hypotheses below. The hypotheses emerge from a higher-level formulation of my response to the question just stated. My speculative answer to the framing question is as follows:

> Speech and non-speech memories are woven together in clusters determined by similarity in a multi-dimensional space. Central to language production is a tension between automatic and agentive behaviour. The phonetic variant most likely to be produced in the absence of initiative intent is the one most often encountered in similar previous contexts. Meanwhile, initiative intent allows a speaker to actively foreground or background any number of dimensions of indexical meaning. By searching and contorting their memory space, the speaker can produce a phonetic variant that is *not* the most likely fit with the context. Such initiative acts of identity can then forge new indexical associations for both speaker and listener as they are woven into memory.

This perspective will be examined through six hypotheses. A succinct version of each is presented below. The first four hypotheses are exploratory — they help me to structure my analysis of the corpus data presented in Chapter 2. The two hypotheses for speech perception, by contrast, were formulated and preregistered prior to data collection, and thus involve confirmatory hypothesis testing, as presented in Chapters 3 and 4. Each of the hypotheses will be defined in more detail in the chapters that follow.

Hypotheses for corpus analysis:

1. Dominance Hypothesis: Features of SPMSS will be prevalent in NZ singing and rap.

2. Accuracy Hypothesis: In the absence of an intention to produce NZE, NZ performers will be accurate in their adoption of the SPMSS model.

3. Genre Hypothesis: Genre will structure variation more strongly than speaker characteristics.

4. Salience Hypothesis: The desire to present an ‘authentic identity’ in song will be enacted in more sociolinguistically salient variables at more contextually salient sites.

Hypotheses for perception experiments:

5. Reference Frame Hypothesis: New Zealand listeners will shift their phonemic reference frame in expectation of SPMSS in a pop music context.

6. Lexical Access Hypothesis: New Zealand listeners’ lexical access will be facilitated for a US voice when it occurs in a pop music context.

These hypotheses speak to only a subset of issues involved in the perception and production of language forms in popular music contexts. A wide range of approaches are possible, and at present there is very little research on the subject. It is hoped that the research presented in this thesis will demonstrate that the study of sociophonetic processes in song can provide us with fresh perspectives on language cognition and how it interacts with the social motivations of language users.

1.7 Thesis Outline

This first chapter has outlined a series of interconnected themes that have led to the formulation of the above hypotheses. Beginning with the role of memory in auditory learning, especially in the statistical learning of co-occurrence patterns, we moved on to the socio-historical conditions through which SPMSS came to dominate popular music singing. An emphasis was placed on the tension between convention and innovation, and the role that ideologies of authenticity play within the midst of that tension. In Chapter 5, I will return again to some of these themes, equipped with the results of the empirical studies presented in the next three chapters, which I briefly outline below.

Chapter 2 analyses sociolinguistic patterns in the Phonetics of Popular Song (PoPS) corpus, beginning with the methods by which this corpus was built. It also introduces some other points of methodology to be used throughout the thesis, including the development of measures of relative lexical frequency in song as compared to speech. A mixture of auditory and acoustic analyses of BATH, post-vocalic /r/, LOT and GOAT will be presented, and the way these features co-vary across individual performers will be considered. With the results of each variable, I will examine the normativity of SPMSS in pop music, and the identity goals and salience required for innovation away from that norm.

Chapter 3 presents the first of two perception experiments, a phonetic categorisation task (PCT) exploring the expectations listeners have about where the boundary between vocalic phonemes will be in singing versus speech. Production data for the DRESS and TRAP vowels from various sources will be presented, followed by the PCT, which plays participants words along a re-synthesised continuum from DRESS–TRAP in the context of Music, Noise or Silence. This is followed by a lexical decision task (LDT), in Chapter 4, which explores ease of processing lexical items in congruent and incongruent voice-context pairings. Recordings of US and NZ speakers are presented once again in Music, Noise or Silence. The US voice is congruent with the musical context, while the NZ voice is not. In Chapter 5, I return to exemplar theory, using the results to consider in more detail how the role of context shapes the storage and categorisation of acoustic memories.

Chapter 2

The Phonetics of Popular Song: Creation and Analysis of a Corpus

2.1 Variationist Approaches to Singing Accents

In this chapter I will present the Phonetics of Popular Song (PoPS) corpus and analyse a series of sociolinguistic variables in order to explore the background assumptions introduced in Chapter 1, particularly to determine the prevalence of SPMSS in NZ popular music. The results of the analysis also provide an empirical basis for the perception experiments which follow in Chapters 3 and 4. As discussed in the previous chapter, singing voices from the USA have been a dominant force in the global industry of recorded music since the first half of the twentieth century. While there has been interest for some time in the ways non-Americans adopt US pronunciation styles in song, little research has focused on the dialectology of US song itself, with the exception of a few early comments on regional phonetic variation (Sackett, 1979), and some good work on the sociolinguistics of US hip hop (e.g. Alim, 2002; Hess, 2009), though these studies only rarely look specifically at phonetics (e.g. Taylor, 2011). Much of the sociolinguistics of popular song has, rather, consisted of case studies about artists who are not from the USA, either focusing on individuals (Beal, 2009; Flanagan, 2019; Heuer, 2017; Jansen and Westphal, 2017; Konert-Panek, 2017a,b, 2018; Duncan, 2017; Eberhardt and Freeman, 2015) or a small selection of artists (Trudgill, 1983; Simpson, 1999; Gerfer, 2018; Westphal, 2018). Studies by O'Hanlon (2006) and Coddington (2004) on the phonetics of Australian and New Zealand pop and hip hop were notable for including a larger number and range of artists.

Several studies of singing accents by non-US artists assume that the style shift from speech to singing is an initiative one, rather than a response to norms. With respect to Australian singer Lenka, for example, Yang (2018, p. 202) describes Lenka's use of American variants as intentional, and as being a way to 'connect with her fans worldwide'. Konert-Panek (2017b) presents compelling evidence, however, of an intentional shift to AmE in song by UK artist Adele. Her use of SPMSS increased as her commercial acclaim grew. Trudgill's 1983 examples of hyper-correction in song also provide evidence of an intent to use AmE. Such examples, however, appear to be rare exceptions. An exhaustive study of hyper-correction would perhaps show its prevalence decreasing as the 'cultural colonies' move through the Nativization phase (Schneider, 2007) and develop a native-like command of SPMSS in song.

While my focus is on the adoption of phonetic features of SPMSS by native speakers of English, this process of cultural flow also occurs at the level of language choice. See, for example, Zhou and Moody (2017) on the use of English in Chinese singing competitions. Studying the singing and speech of non-native speakers of English is instructive. Comparing singing and speech outside of recorded popular music, Mageau et al. (2019) had non-native speakers of English (who were not professional singers) sing and recite the words to a nursery rhyme. In a subsequent task, the voices were rated by listeners as less accented in song than speech (supporting the findings of Hagen et al., 2011). Mageau et al. (2019) conclude that this was because in singing, prosodic cues to non-nativeness are removed from the signal. The idea that there could be better imitation of English phonetics in song than speech at the segmental level, however, remains an intriguing possibility for future study. Mageau et al. (2019) also present a comparison of speech and singing production amongst a set of participants who were native speakers of Canadian English. There were no large differences in the vowel formants of the KIT and STRUT vowels between singing and speech. This was taken as evidence that only experienced singers shift their phonetic style in song. However, since these singers are already native speakers of a North American dialect of English, it may simply be the case that they do not *need* to shift their vowel qualities when singing, since their own dialect already approximates the Standard Popular Music Singing Style (SPMSS), at least for these vowels. It would be interesting to analyse the full range of vowel variables in this dataset to see whether any signs of specifically local dialect features are present in the participants' singing style.

Stone et al. (1999) also found little difference between singing and speech for Southern American participants singing a country song and the national anthem, and a similar interpretation can be made with respect to these results, Americans don't need to shift their accent when they sing. But of course, there is a wealth of diversity within the United States, as amply demonstrated for example by Labov et al. (2006). The characterisations of AmE presented in many studies of singing accents are over-simplified. Regional and social variation between artists in the USA needs to be considered alongside the uptake of the normative singing style by non-Americans.

Coupland's 2011 discussion of the roles of vernacularity and place in popular music by artists from the USA stands apart from the studies cited above, bringing theoretical depth, and criticising the tendency to reduce these rich sociolinguistic texts to the analysis of whether artists 'do or do not maintain features of their national or regional accents in singing' (p. 573). While this very question remains at the empirical centre of the present thesis, I intend to use it as a tool for better understanding language cognition, particularly the role of context in storing and accessing the highly variable acoustic signals associated with words. This work is thus situated more in laboratory phonology than sociolinguistics, and has different goals to many of the above studies, though an understanding of identity and authenticity practices is still crucial.

2.1.1 The importance of salience in identity construction

Salience is a difficult concept in sociolinguistics. Everyone seems to realise that it is crucial, and yet it is still far from having a consistent theoretical framework. It is the ideological layer of semiotic associations in linguistic anthropology (Woolard, 2008), the n+1th order of indexicality (Silverstein, 2003), it distinguishes the marker and the stereotype from the indicator and forms the critical distinction between sound change from above and sound change from below (Labov, 1966).

In a study comparing singing and interviews with Hebrew speakers in Israel, Yaeger-Dror (1991, p. 312) showed how ‘different song genres have different dialect targets’ and that these styles are different from the speech styles of the singers. An important aspect of this study was a focus on cognitive salience. Following the insights of that study, I claim throughout this chapter that words in prominent positions allow singers to enact their conscious identity goals. Cognitive salience was hypothesised by Yaeger-Dror (1991) to be greater on open class words and infrequent words, while the occurrence of one sociolinguistically salient variable was hypothesised to increase the salience of surrounding variables.

Babel (2016) marks an important collection attempting to tackle the issues of awareness and control in sociolinguistic production and perception from multiple perspectives. These two connected phenomena can be seen as the result of salience, which is given its most in-depth treatment from a sociolinguistic perspective by Rácz (2013), with the fundamental logic of that treatment finding further empirical support in (Racz et al., 2017). Rácz (2013) contrasts ‘sociolinguistic salience’ to the largely bottom-up notion of salience used in the visual cognition literature. In the latter, a feature of a single visual field is salient through its intensity or its contrast to its surrounds. In sociolinguistics, however, salience denotes something which ‘juts out’ from a more complex frame of reference: ‘A segment is salient if it has a large surprisal value when compared to an array of language input’ (p. 51). Bottom-up auditory prominence is often conflated with this more high level definition in the sociolinguistic literature.

In Hay et al. (2018), we distinguished expectation-driven (top down) salience from stimulus-driven (bottom-up) salience, focusing on the role of novelty in triggering a shift in attention to a stimulus which is not expected given prior experience. I use the terms ‘sociolinguistic salience’ and ‘contextual salience’ in this chapter. The latter term is used as an intermediate position, in which the jutting out of a variant relates to neither a ‘wide array’ of language experience nor a strictly perceptual local prominence. It is attention-grabbing at the level of the local context through either perceptual prominence or through informativity (as characterised in the most basic sense by low frequency lexical items, though this depends on the context leading up to the word). I believe that the systematic analysis of popular music singing and rap styles holds much promise for gaining a better understanding of these difficult concepts.

With these larger themes in mind, I turn now to the rationale behind the design of the PoPS corpus, and to the important role that sociolinguistic salience plays in the variables chosen for analysis. Contrary to the majority of the studies cited above, I focus on both the US artists who are part of the tradition from which the SPMSS evolved, and on NZ singers who grew up in one of the popular music ‘colonies’. I focus on commercial pop and hip hop songs. Pop music is included because it is where the dominance of SPMSS is expected to be most robust, and hip hop is included for the opposite reason — it is a site where authority is explicitly contested, and where display of one's own specific regional origin is highly valued (Hess, 2009; Pennycook and Mitchell, 2009). I acknowledge the importance of the commercial vs. underground dichotomy to the issues at hand, but I sacrifice the analysis of this distinction in order to allow for greater systematicity in comparisons across the groups of songs that are included, all of which come from commercial music charts.[1] Whether hip hop culture maintains its traditional values of authenticity in the commercial context is an interesting topic in its own right. This thesis will thus contribute to the debate around the effects of commercialisation on hip hop (see, Oware, 2014).

Trudgill's 1983 exploration of the pronunciation patterns of The Beatles, The Rolling Stones and a selection of other artists, particularly from the punk movement, set the agenda for the sociolinguistics of popular music. Many of the insights of that study continue to hold explanatory power. The titular concept of 'conflicting identities' (which draws on the acts of identity framework, Le Page and Tabouret-Keller, 1985) is crucial to any understanding of why artists choose to sing the way they do. Trudgill's original study also involved an orientation towards salience, by choosing to analyse variables which are known by language users to distinguish Southern British English (SBE) from AmE. He also made claims about the intentionality of the artists, stating that British artists 'put on' an American accent, and that over time as they introduced British English features into their singing accent, they were 'trying *less hard* to sound like Americans' (p. 154, emphasis in original). Subsequent work has challenged the directionality of this intention. In her analysis of Arctic Monkeys vocalist Alex Turner, who uses features of his Northern English dialect in song, Beal (2009) argued that the normativity of AmE accent in song is so strong that the use of regionally marked features requires effort. In my own work, I provided evidence that this normativity is indeed systematic, and affects the whole vowel space (Gibson, 2010b; Gibson and Bell, 2012). Looking at a range of non-salient vocalic variables, three NZ singers were shown to shift their entire vowel space away from NZE when singing — effectively adopting some version of AmE phonology, complete with the low-back COT–CAUGHT merger. Through interviews about the identities they wished to project through their persona as a singer, and analysis of cases where conscious attention was brought to bear on the phonetics of their performance, this study showed that even the intention to adopt NZE in singing was not sufficient. The intention to sing in one's own (non-US) accent needs to be coupled with a high level of awareness of the distinction between the NZE and AmE variants.

While Gibson and Bell (2012) demonstrated the normativity of SPMSS in song by providing evidence of its adoption in non-salient variables, the present analysis returns to the salient variables addressed in so many of the studies cited at the start of this section. These variables may reveal the initiative use of NZE in sung performance. Following the logic of Gibson and Bell (2012), this study is interested in these salient sites not because they offer examples of AmE accented singing (which is found across a much wider range of variables), but because it is in these salient sites that own-accent singing is most likely to come about. Such usages of NZE

[1]Note, however, that the US music industry is gargantuan in comparison to the NZ music industry, so the large majority of NZ music in the sample is inherently less commercial than its US counterpart.

constitute initiative acts of identity that require a relatively high degree of cognitive access to variables and their social meanings. The songs analysed here are clearly examples of what Bell and Gibson (2011b) termed 'staged performance', which is strong on the 'meta', involving high degrees of planning, rehearsal and reflexivity.

After a description of the methods used to build the PoPS corpus in Section 2.3, the analysis begins with two of the most salient features that distinguish SPMSS from NZE, the BATH vowel (Section 2.4) and rhoticity (Section 2.5). These variables are examples of distinctions between the dialects which are relatively accessible to performers' awareness since they are, at least in some sense, categorical. There is a cross-dialectal difference at the phonemic level in the case of BATH, with the lexical set being aligned with PALM in NZ and TRAP in the USA. This affects the rhymes that an artist can or can't use, and is therefore particularly likely to be brought to attention during the songwriting process. Rhoticity is also relatively cognitively accessible, since it involves a presence vs. absence distinction in the realisation of non-prevocalic /r/. The analysis of non-prevocalic /r/ will be followed by an analysis of its intervocalic counterpart, linking /r/, in Section 2.6.

An acoustic analysis of the LOT vowel follows in Section 2.7. BATH, rhoticity and LOT are all members of the USA-5 (Simpson, 1999) set of variables originally studied by Trudgill (1983).[2] LOT is of particular interest since it occupies a moderate position in terms of salience, and involves more gradient phonetic inter-dialectal differences. As the discussion develops, I will reflect increasingly on individual performers. Section 2.8 draws together the results of all the other variables to look at the patterns with which NZ artists adopt combinations of NZ features. The conclusions made about individuals on the basis of BATH, LOT and non-prevocalic /r/ will be tested against a basic auditory analysis of the GOAT vowel, which appears to be attracting some salience, especially amongst own-accent NZ hip hop artists.

2.2 Research Questions and Hypotheses

The question 'why might it be difficult for New Zealanders to sing in NZE?', which frames this research project, is based on the assumption that NZ artists' performance of SPMSS occurs without effort. There is no obvious way to directly assess the 'difficulty' of using NZE in song in an analysis of recorded music. I therefore propose here four more focused research questions that can be explored through analysis of the corpus. These questions will guide the analysis and discussion presented throughout this chapter:

1. Do prominent New Zealand popular music performers (still) use (US-derived) SPMSS variants in their vocal performances?

2. Are New Zealand singers and rappers able to accurately reproduce the norm provided by artists from the US?

3. Does genre determine phonetic style in song to a greater extent than the cluster of influences associated with a performer's own background, including their place of origin, ethnicity and gender?

[2] The others were intervocalic /t/ and PRICE.

4. How can we assess the role of sociolinguistic salience and, relatedly, intentionality in the phonetics of popular music performance?

Though the analysis of the corpus data is more by way of exploration than confirmatory hypothesis testing, I offer here a high-level hypothesis for each of the above questions, along with some considerations of the type of evidence that could conceivably falsify the predictions made. These hypotheses will be applied when introducing each variable to give more specific predictions, and will be considered again in the discussion of each variable to determine how the results support or negate the predictions made.

- Dominance Hypothesis — SPMSS variants will be prevalent in the NZ performances.

  Counter-evidence to this hypothesis would be an absence of SPMSS variants in the vocal performances of the NZ artists.

- Accuracy Hypothesis — If SPMSS is the default singing style, involving some degree of automaticity or habitus (Bourdieu, 1991) in its production by NZ artists, then it could be considered a part of NZ singers' 'native' repertoire. Performance will thus be accurate and consistent, and will not bear hallmarks of stylisation, such as mis-realisation and overshoot (Bell and Gibson, 2011a). In contrast, use of NZE in song is predicted to involve greater awareness (the initiative projection of an 'authentic self'), and will thus show signs of stylisation.

  The Accuracy Hypothesis forms part of the foundational assumption of the thesis, that was outlined in Chapter 1 (and justified by the argumentation and evidence put forward by Gibson and Bell, 2012), and also forms a hypothesis to be further tested. Evidence against the Accuracy Hypothesis could come from examples of inaccuracy or inconsistency in the adoption of SPMSS variants, for example through overshoot (in either frequency or degree of articulation), or hyper-correction, which would provide strong evidence against the hypothesis. As outlined in Bell and Gibson (2011a), while such signs of innacuracy can be either accidental (cf. Le Page's riders to linguistic modification) or strategic (i.e. deauthentication[3]), they do suggest an intentionality to the use of AmE variants. The classic example of this was the hyper-correct insertion of /r/ in Cliff Richards' *a*[r] *bachelor boy*, described by Trudgill (1983). Conversely, hyper-correction or overshoot in the use of NZE variants by NZ artists would lend further support to the assumption that own-accent singing involves initiative and intentional stylisation.

  In sum, these first two hypotheses state that NZ pop will be similar to US pop, with any reduction in aggregate being due to certain individuals intentionally and consciously avoiding SPMSS, rather than a slightly lower rate overall. This would provide some evidence for the claim that own-accent singing is a defiant and conscious phenomenon, likely to be done with some ideological intent.

[3]'Denaturalisation' in the language of Bucholtz and Hall (2005).

- Genre Hypothesis — Genre will be the primary structuring variable, with a high level of homogeneity in pop and more examples of own-accent styles in hip hop.

  Pop singers will use the SPMSS style, which is based on 'General AmE'. In pop, singers will be statistically indistinguishable from one another according to their place of origin and ethnicity. If there are gender differences in US pop, these will be mirrored by NZ artists. A small number of NZ pop artists are also expected to engage in 'own-accent' singing, though much fewer than in hip hop. These artists can be removed prior to tests of homogeneity. The prediction here is that *when* a NZ artist adopts SPMSS, they will be indistinguishable from their US counterparts.

  Counter-evidence to this part of the Genre Hypothesis would come from significant differences between NZ and US pop singers, or strong signs of ethnicity-based variation.

  Hip hop will also have a normative style, hip hop nation language (HHNL), which is based on African American English. However, many artists are expected to engage in 'own-accent' phonetic styles that represent their place of origin and/or ethnicity.

  While my analysis in this project is limited to mainstream pop and hip hop, I should be clear that I see genre as perhaps the most important factor structuring the phonetics of song. Not only will different genres have different phonetic traditions, they will also have different degrees of homogeneity. In the present analysis we will see hints of this through the predicted orientation of pop to 'General AmE' and of hip hop to AAE, and also of the prediction of greater homogeneity of phonetic styles in pop than hip hop. It should be kept in mind, however, that the types of music under analysis here represent only a very small subset of popular music, and an even smaller subset of song in its more general sense.

- Salience Hypothesis — High levels of *sociolinguistic salience* (e.g. variables such as BATH that act as stereotypical dialect markers) or *contextual salience* (e.g. instances of a variable that attract attention for reasons of auditory prominence or greater informativity) will allow NZ vocalists to enact their identity goals. Low levels of salience will lead to the use of whatever form has been encountered most often in similar contexts.

  This hypothesis is less clearly falsifiable, but will provide a framework for assessing the covariation between the variables for individual artists. There is likely to be individual variation with respect to the relative salience of different variables. However, this hypothesis would be supported if there are signs of an implicational scale across variables. I hypothesise that BATH will be the most salient variable for most singers, and will thus attract the most widespread use of the NZE variant. Non-prevocalic /r/ is also salient, though perhaps to a lesser extent due to its variability.[4] Those that completely avoid rhoticity (in

[4] An important distinction is made between cases of non-prevocalic /r/ in the NURSE lexical set, and cases in non-NURSE environments. The NURSE environment strongly favours the realisation of /r/ in a very large number of dialects of English. When there is partial rhoticity in a dialect, it usually includes rhotic NURSE. NZE is one such dialect, where NURSE words are frequently

Table 2.1: Predicted Salience Hierarchy. Schematic for a possible implicational scale relating to salience of variables. '+' indicates adoption of a NZE variant. Those most committed to such an identity are represented by the use of NZE variants in all variables (bottom row), while those less committed to presenting a NZ persona in song would only use NZE variants for variables to the left of the table, as in the top row.

	BATH	Rhoticity	LOT	GOAT
Commercial (SPMSS/HHNL)	–	–	–	–
	+	–	–	–
	+	+	–	–
	+	+	+	–
	+	+	+	+
'Authentic' (Own-accent)	+	+	+	+

+ denotes the use of a NZE variant

non-NURSE environments) should also avoid the SPMSS variant of BATH. LOT is expected to be less salient again. Those who use a NZE variant of LOT will thus be expected to also use NZE BATH and avoid non-NURSE rhoticity. Use of NZE GOAT is expected to be reserved to those with the strongest intentions to project an 'authentic'/NZE identity. The proposed salience hierarchy is summarised in Table 2.1.

It should be noted that a view of intentional and stylised SPMSS singing would predict the opposite outcome: greater use of SPMSS on more salient variables, and use of NZE in the absence of such higher order awareness.

Through this kind of logic, it may be possible to bootstrap our way to a better understanding of both the salience of each variable, and of the identity goals of the singers. This approach will be necessarily qualitative, but will be able to use the corpus data to formulate testable hypotheses for future studies. There is likely to be individual variability in these salience rankings, with some people being aware, for example, of NZE vs. SPMSS LOT, without being aware of the differences in rhoticity. One additional consideration: within any given variable, a NZE variant is more likely when a given token is in a contextually salient position (through auditory prominence or greater informativity) for those who show signs of wanting to express an 'authentic identity'. In this way, a NZ artist with authenticity goals may use NZE on a prominent instance of LOT but produce SPMSS in a non-prominent token.

I am aware that there is great potential for circular argumentation with this approach. My hope is that by outlining the hypothesis in this specific form, it then becomes falsifiable.

Before providing details about the constructing of the corpus, I present here one important caveat relating to the auditory analyses of the data presented throughout this chapter. All auditory analyses were conducted by myself, and while my own consistency was checked (through blind recoding) for some variables, my assessments

rhotic, particularly in Māori and Pasifika styles, as well as in Southland. When judging adoption of SPMSS, I therefore focus primarily on non-NURSE environments.

were not validated by a second rater. Given that I was familiar with the demographic background of the artists when doing the auditory analyses, there is scope for bias in my decision-making. I am confident that at a conscious level I was as rigorous and transparent as possible, but the potential for unconscious bias towards my hypotheses is an issue which exists for many of the results presented in this chapter.

Regarding the structure of this chapter, rather than presenting grouped methods, results and discussion, I split this chapter up into separate sections for each of the variables, and move through the background information, methods, results and discussion for each in turn. The discussion will therefore build gradually as more evidence is considered, leading to a brief restatement of the findings as they relate to these hypotheses at the end of the chapter (Section 2.9). Before beginning that process, however, it is necessary to introduce the PoPS corpus, and describe the methods by which it was built.

2.3 Introducing the PoPS Corpus

The PoPS corpus is made up of 190 vocal performances by 154 artists, with lyrics manually time aligned to the songs' audio at the utterance (or lyrical 'line') level, and then force aligned at the phoneme level. It is structured by country (NZ and USA), ethnicity (Pākehā and Māori/Pasifika in NZ, and European American and African American in the USA), genre (pop and hip hop) and gender (male and female in pop, but only male in hip hop since very few female hip hop tracks were revealed with the song selection methods described below). The number of songs and artists in each of these demographic cells is summarised in Table 2.2. A full list of songs is provided in Appendix A.

Table 2.2: Number of songs in each cell of the corpus, with number of unique artists in brackets.

Country	Ethnicity	Female Pop	Male Pop	Male Hip Hop	Total
NZ	Māori/Pasifika	20 (13)	17 (12)	19 (17)	56 (42)
	Pākehā	15 (13)	16 (11)	13 (10)	44 (34)
USA	African American	15 (11)	15 (10)	15 (15)	45 (36)
	European American	15 (15)	15 (15)	15 (12)	45 (42)
Total		65 (51)	63 (48)	62 (54)	190 (154)

2.3.1 Methods of song selection

Avoidance of selection bias was one of the primary motivations in developing the methodology for song selection, which proceeded systematically using the NZ singles charts maintained by Recorded Music New Zealand (RMNZ, available at `http://nztop40.co.nz/chart/nzsingles`, and shown in Figure 2.1). Setting up in advance a stringently defined set of rules to govern the selection of songs, I made myself as 'tasteless' (Brooks, 1982) as possible. That is, I did not allow my own judgements about the worthiness of a given song for study to guide selection decisions. Since the primary interest of this thesis is the music to which New Zealand listeners are

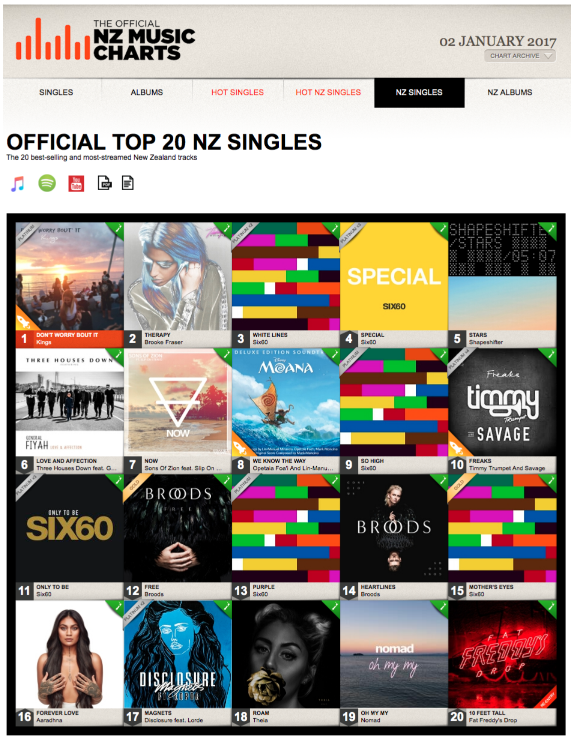


Figure 2.1: Recorded Music NZ (RMNZ) NZ top 20 singles chart (image used with permission from RMNZ).

exposed, the RMNZ charts were used to find the songs by both the USA and NZ artists, using a strictly defined set of inclusion criteria, which will be detailed below. The charts were searched manually, to determine which artists fit these criteria, and in the process, a database of information about 1786 songs was developed, including reasons for exclusion from the corpus. This information will be summarised once the inclusion criteria have been described.

2.3.1.1 Inclusion Criteria

- Place of origin — Artist must have grown up in NZ or the USA. There is debate about the critical/sensitive periods for language and dialect acquisition (Werker and Hensch, 2015), and I decided on what is probably a conservatively young age as the boundary for inclusion, excluding anybody who had moved to NZ/USA after the age of five.[5] For US artists, biographical information about their region of upbringing was also consulted.[6]

- Genre — The genre of the artist had to be either pop or hip hop/rap on the artist's page in iTunes. The decision to use iTunes genre was made for

[5]The corpus thus included, for example, Chelsea Jade (who moved from South Africa to NZ at age five) and Zaidoon Nasir from Times X Two (who moved from Iraq to NZ at age two) but excluded Young Tapz (who moved to NZ from Zimbabwe at age eight) and the rapper from Earth Tiger (Cruz Matthews, who moved to NZ from California at age seven).

[6]Regions were grouped into the following categories: West (including all west coast states, but excluding Arizona, treated as 'other'), South, Midwest (including towns as far east as Pittsburgh, Pennsylvania), and East (including towns in eastern Pennsylvania). Artists who moved between two different regions during childhood were treated as 'mixed'.

replicability and simplicity, since iTunes is rare in allowing only one genre label per artist. However, it is far from perfect. There are many artists with genres that might not be the best reflection of their music, and iTunes often applies different genres to individual releases.[7] Artist genres in iTunes are also relatively unstable. In cases where an artist's genre changed during the course of data collection to make them ineligible, any already collected data was maintained. In cases where an artist's genre changed making them eligible, songs were added.[8]

- Ethnicity — Artists belonged to one of four broadly construed ethnic groups: NZ Māori/Pasifika, NZ Pākehā, African American and European American. This was, by far, the most difficult information to determine. I discuss the details of this process below.

- Gender was treated as binary. There were no cases of artists whose gender appeared to be ambiguous in the context of a binary distinction, except for the Pākehā vocalist in the pop group Openside, whose track was encountered after the quotas for both male and female Pākehā pop had been met.

- Many songs are collaborations, labelled as, for example, X feat. Y. In such cases, both artists could be considered separately for analysis so long as they were easily distinguishable: in this case, the genre and place of origin were established for each artist, and if they were eligible for the study, their vocal sections of the song were analysed as separate performances. In cases where I could not reliably distinguish between an eligible and an ineligible artist, the song was excluded.

- In cases of a band or group, where there are multiple singers under the same name, the individuals were investigated. If all eligible individuals were of the same place of origin, ethnicity and gender, then the multiple vocal performances were treated as having been performed by a single artist. In cases where there were eligible vocalists of differing ethnicities, the vocal performances by members of the same ethnicity were grouped and treated as having been performed by a single artist.[9]

- Since individual high-profile artists often dominate the singles charts upon the release of a new album, a maximum of one song per artist per calendar year was selected, based on the chart date (not the release date).

- Analysis initially began from mid-2017 and worked backwards through the charts. For the NZ songs, all songs in the top 20 NZ singles and NZ Heatseekers Singles charts were assessed. For the US artists the Top 40 singles and Top

[7] For example, Lorde's genre as an artist is pop, as are her early releases. Her second album, however, is tagged as alternative. It is the artist genre to which I deferred in all cases except for the rare instance where two different artists were merged in error. For example, Savage does not have his own artist listing — he is mixed together with a dancehall group, and is assigned their genre. In such cases, I chose the genre most dominant amongst their releases.

[8] This latter case applied to both Lorde and Six60.

[9] For example, in the song 'Talking to you' by the 9-5ers, the Pākehā MC, Edgar, and the Samoan MC, Sabe, were analysed separately, whereas the two Pākehā vocalists in the song 'Brightside' by Mae Valley were grouped into a single analysis.

10 Heatseekers charts were used. The Heatseekers charts include the 'fastest-rising' titles outside of the Top 20/40 chart for a given week. The final stages of song selection occurred in early 2018, and up until that point, newer charts were considered as they became available.

- Songs reappearing in the charts that had originally been released more than ten years prior were excluded.

2.3.1.2 Identifying ethnicity

In an attempt to assess the independence of ethnicity from genre in phonetic patterns, I decided to demarcate a binary variable for ethnicity in each location. In New Zealand, it would have been highly unrealistic to require artists to have Māori ancestry on both their maternal and paternal sides, and it is customary (for example with the Māori electoral role) to treat any Māori ancestry as being significant in one's ethnic identity. For Pasifika peoples in NZ, it is also common to identify with more than one ethnicity. While distinct, there is significant overlap between Māori and Pasifika speech styles in New Zealand (Starks et al., 2015). In order to be able to fill the cells of the study design, it was decided that Māori and Pasifika artists would be grouped together for the purposes of the corpus, even though this kind of pan-ethnic grouping is problematic (Starks et al., 2015). Furthermore, anyone who claimed any degree of Māori or Pasifika heritage was included in this group. In the USA, due to the much larger music industry, and the availability of information about artists in the charts, it was much more feasible to find artists whose ethnicity was African American or European American with respect to both parents.[10]

Determining ethnicity and place of upbringing for NZ artists was a difficult endeavour, since many have only a moderately visible public profile. My method was to search online for interviews, wikis and other media sources for information about ethnicity and upbringing. This approach has obvious limitations, but resulted in relatively quick and acceptably reliable information for many artists. For those where no ethnicity information was found, a message was sent through Facebook (where possible, or by email to the artist's manager where Facebook messaging was unavailable) asking them which ethnicity/ies they identify with. More than half of those asked responded. These artists are marked with 'PC' in the ethnicity column of the list of songs provided in Appendix A. Internet searches for the American artists proved to be much more straight forward, since all US artists appearing in the NZ charts have a large online presence (see Section 2.9.8 for a discussion of this discrepancy between the NZ and US artists). The results of my internet searches no doubt provide essentialised and crude approximations of ethnicity, but for the purposes of structuring the corpus with minimal selection bias, these methods were deemed satisfactory.

[10]Artists with mixed ancestry were excluded, as were people with Hispanic ancestry. People with African-Caribbean ancestry were included as African American, so long as they had grown up in the USA (e.g. Jason Derulo has Haitian parents but grew up in Florida). In the US census, whether a person is Hispanic or not is treated as a separate question to ancestry, so a person can mark their 'race' (the term used in the US census) as 'White' or African American, and then choose whether or not they are also Hispanic. The 'White' category includes people with ancestry in Europe, the Middle East and Northern Africa. The definition of ethnicity used in the corpus creation was thus loosely based on the US census, including only non-Hispanic European Americans and African Americans.

One further note on ethnicity. The analysis of the corpus involves the singing and rap of Māori and Pacific peoples, and subjects those voices to an epistemology which is firmly rooted in Western schools of thought. I want to acknowledge this as a limitation. The thesis would have been richer had I engaged with mātauranga Māori (Macfarlane and Macfarlane, 2018) and Pasifika systems of knowledge.

2.3.1.3 Under-represented cells

The system of data collection outlined above emphasised objectivity and replicability, but perhaps sacrificed common sense at times. By selecting only from music charts, and by out-sourcing the defining of genre to iTunes, some cells of the design were extremely difficult to fill.[11] Some of the most prominent Māori and Pasifika female singers were not initially included in the corpus, for example, since their genre was almost always identified by iTunes as r&b/soul, rather than pop.[12] Presumably, iTunes is both reflecting and reinforcing stereotypes of ethnicity and genre in this regard. This situation was even more pronounced in the USA sample, where it was very difficult to find female African American singers labelled as pop, not r&b/soul. For both Pākehā male hip hop and African American female pop, the rules for song selection had to be loosened to find the remaining tracks.

After following the outlined method for choosing the NZ songs all the way back to the instantiation of the RMNZ NZ Singles chart (31 Oct 2011), there were still only seven tracks by Pākehā male hip hop artists. To amplify this demographic slightly without changing the method of selection too drastically, nominees for Best Urban/Hip Hop album at the New Zealand Music Awards (organised by RMNZ) were examined for instances of Pākehā male rappers. Three further tracks were added this way.[13] Another track featuring Pākehā rappers appeared in the charts after the corpus had been otherwise completed (‘On the Rark’ by Machete Clan), and this was also added bringing the number of individual artists in that cell up to ten. To find the remaining African American pop tracks by female artists, I turned to allmusic.com for genre definitions. If the genre/style definitions for an artist included some variant of pop (including pop/rock, alternative pop etc.), and, importantly, did not include hip hop, then the songs identified for that artist in my database were added to the corpus. I then also applied this genre requirement to the NZ female Māori/Pasifika pop cell of the corpus, which led to a better range of female Māori and Pasifika pop artists.

2.3.1.4 Summary of songs excluded from PoPS

As mentioned above, records were kept about the songs not selected for the corpus. These entries included basic information on the artist and song name, and whatever information was gathered about gender, genre, place of origin and ethnicity, and the

[11] As mentioned above, the most obvious gap in the design is female hip hop. Unfortunately, a study including female hip hop would need different song selection methods than those employed here. For example, the top-selling singles in a given year for each demographic cell could be identified to collect data otherwise missing from the charts.

[12] As described below, this problem was eventually overcome.

[13] Two of the tracks were by Tom Scott (one for Home Brew’s nominated album ‘Home Brew’, and one for @Peace’s album ‘Girl Songs’) and one by Jody Lloyd (for Dark Tower’s album ‘Canterbury Drafts’). The tracks were chosen on the basis of having a music video clearly showing which vocalist was performing which sections, and for having a good quantity of rap by the selected artist.

reason for exclusion. Any singles involving multiple eligible vocalists also needed to be analysed in terms of which individual(s) performed the vocals of the song. Information about an artist was gathered in whatever order was most forthcoming, and no more information was gathered once it became clear that a given song should be excluded. The statistics that follow are therefore rough, since only one inclusion criterion needed to be broken in order to exclude a song. In some cases, I knew immediately that an artist was Canadian or British, for example, so there would be no need to check genre, while the opposite situation could occur for other artists for which I knew the genre (bear in mind that the same artists appear repeatedly in the charts and so I began to be able to reject songs based on remembered information). Even though the figures which follow are thus approximate, I include them here to give an idea of the broader landscape of NZ chart music according to genres and artist origins.

- Exclusion on the basis of genre was most common, with 685 songs marked as excluded for this reason, from the following genres: 209 dance/electronic, 124 alternative, 116 r&b/soul, 52 rock, 45 reggae, 31 singer/songwriter, 16 country, 13 inspirational, 10 soundtrack, 9 world, 6 classical, and a handful of songs from blues, comedy, children, metal and other genres.

- Exclusion on the basis of artist origin was marked for 215 songs, with vocalists coming from a range of countries: 68 United Kingdom, 34 Australia, 14 Canada, 9 Sweden, 6 Norway, 4 Ireland and a range of others.

- Exclusion due to ethnicity was a rarer case, since ethnicity was harder to find out than genre and place of origin. Exclusions based on ethnicity were therefore generally songs which otherwise fit the criteria. Of the 44 songs excluded for this reason, most were by American artists who either had mixed ethnicity (16) or Hispanic ethnicity (16).

- Songs were also excluded once various types of quota were full. 217 songs were excluded because an artist already had a song included for the given year of charts. 315 songs were excluded on demographic grounds, once the relevant cell of the corpus design was full. For example, the Pākehā female pop cell filled quickly, and further songs fitting these criteria were dismissed.

- There were several other less common reasons for exclusion: songs not on iTunes (44), songs more than ten years old re-entering the charts (32), songs where there were few lyrics or no lyrics (37) and songs in te reo Māori (10).

As for the spread of time from which songs came, the large majority of songs came from the three years to which I applied the most stringent methods, with the hit rate decreasing as the quota for more cells in the design were met: 2017 (106 inclusions from 852 entries), 2016 (37 inclusions from 327 entries) and 2015 (18 inclusions from 148 entries). The reason the number of entries decreased was that I did not enter the same songs into the spreadsheet multiple times, and many songs stay on the charts for a long time. Thus, for 2017, I recorded most chart entries, and reduced the record-keeping to only informative new entries as I moved back in time through the charts. The remaining 29 songs were sourced from older charts (as well as a few songs from 2018), and from the New Zealand Music Awards, as described above.

2.3.2 Procedures for corpus management

Songs identified for inclusion were purchased through iTunes, converted to wav files and imported into Praat (Boersma and Weenink, 2019). Lyrics were downloaded from a range of different websites. They were then stripped of paragraph marks, punctuation and capitalisation and added to the 'lyrics' tier of a textgrid in Praat, which had a 'repetition' tier to mark repeated sections, and another tier to mark sections that needed to be excluded from analysis for a range of other reasons, such as sections performed by artists not under analysis. Lyrics were time-aligned to the soundfile at roughly one line intervals (the length of the interval varied and was determined by phrasing). The lyrics downloaded often contained errors, and I always went with my own interpretation of lyrics.[14]

Following previous studies e.g. (Coddington, 2004), repeated sections were excluded from analysis, and were defined as follows: whole repeated sections performed in a similar manner to their first occurrence in the song.[15] When multiple singers had overlapping vocals, the section was generally included unless it was clear that it would be difficult to analyse the production of the singer of interest. Call and response style improvisations in repeated choruses towards the end of a song were generally not included due to a high degree of overlap between vocal parts.

Audio files and Praat textgrids were uploaded to LaBB-CAT (Language Brain and Behaviour Corpus Analysis Tool, Fromont and Hay, 2012), where the corpus is stored and managed. This makes searching for variables of interest efficient, the results of which can then be exported as folders of audio and textgrid extracts, along with a spreadsheet containing contextual information for each token. The long-term aim of this corpus is to facilitate collaboration in developing a systematic dialectology of popular music. In total, the corpus includes 11 hours and 44 minutes of audio, including 4 hours and 35 minutes of time-aligned utterances. There are 36,109 word tokens of 3903 word types. Each of the 154 vocalists is tagged in LaBB-CAT with their gender, genre (which was assigned to performers rather than songs), the country they grew up in and their ethnicity. Each of the 190 vocal performances were tagged with the year and type of the music chart from which they were selected, and the artist name for the given performance (many tracks have multiple vocalists, so the artist name on the track is not always identical to the vocalist analysed). The transcripts were force aligned at the phoneme level using HTK (Hidden Markov Model Toolkit). Despite the fact that the vocals appear in the context of instrumentation, HTK alignment was impressively accurate, making it easier to locate variables within exported extracts.

[14]I transcribed the lyrics for some NZ hip hop songs that I could not find anywhere online, such as 'Little Did She Know' by Swidt.

[15]More specifically, repetitions of a line within one musical section were included in the analysis, e.g. a chorus which repeats the same line four times was included in its entirety on its first occurrence, but subsequent repetitions of that section were excluded. In cases where two different verses had some lines repeated and some new ones, both entire verses were included. Cases where a section was sung in a very different manner, e.g. with very different dynamics, at a different octave or with a different melody, were included. Note that while the phonetic equivalence of repeated sections is something to be tested, not assumed, the use of 'copy and paste' across choruses of commercial pop songs is prevalent enough to warrant this methodological choice.

2.3.3 Establishing lexical frequencies in song and speech through corpora

Exemplar models predict differing effects for high and low frequency words, and there is extensive evidence that lexical frequency is important in speech perception (e.g. Connine et al., 1993, discussed in Chapter 4). Following from this is the prediction that *ratios* of lexical frequencies can also affect language behaviour (Walker and Hay, 2011; Needle and Pierrehumbert, 2018). In the study of the production data below, then, I consider the frequency of the words used with respect to songs and in speech. Lexical frequencies for speech are widely available, and I took an average across three sources to determine a lexical frequency not overly dialect specific. These sources were Celex (excluding the written portion of the Cobuild corpus, as is discussed in detail in Section 4.3.2.5, the Buckeye Corpus (Pitt et al., 2005) and the Canterbury Corpus (Gordon et al., 2004). To establish lexical frequency in song, a Python script was developed by Robert Fromont to collect lyric data from the website lyricsplanet.com, using a series of nested loops that opened each artists' section of the website, and the title and lyrics extracted for each song. This resulted in 14.9 million tokens, from which frequencies for each wordform type were calculated. Lexical frequency in song was then divided by lexical frequency in speech to produce a 'songiness' ratio. Words that are 'songy' occur more often in songs than they do in speech. Examples of songy and 'speechy' words will be given in Section 4.3.2.5, but for the purposes of the production data below, the explanation given here should provide sufficient background.

2.3.4 Statistical methods: Dealing with small datasets

Throughout this thesis, I primarily employ mixed effects regression models for statistical analysis (with the occasional use of t-tests and chi-square tests to make more peripheral points about the data). Mixed effects models, or multilevel models, essentially involve regressions within regressions (Baayen et al., 2008). The fixed effects, also referred to as the Level-1 predictors, are the main variables of interest to the study. Beyond these objects of interest, however, we also have knowledge about other ways in which the observations are related to one another (their non-independence), for example when we have repeated measures for an individual. While we might be interested in generalities about macro-social categorisations of ethnicity or gender (which are thus included as fixed effects), we know that each individual will have idiosyncracies. These 'random' variations between members of the same ethnic group, for example, are modeled by Level-2 predictors — the random effects. These can be as simple as an intercept (α), allowing each individual to have a different base level with respect to the dependent variable, or they may involve slopes as well (or instead) (β), allowing each individual to behave differently with respect to one or more of the fixed effects. I will discuss the importance of random effects in more detail as the analysis unfolds, but I raise it here with particular reference to the several cases in the results below where the datasets are small.

In its entirety, the PoPS corpus is of a reasonable size (though still tiny in the scheme of sociophonetic corpus linguistics more generally, see e.g. Foulkes and Hay, 2015). The corpus affords a good number of tokens for even relatively low-frequency variables such as linking /r/. However, the cross-tabulation of such results according to place of origin, genre, gender, ethnicity, and phonological environment, can lead to small token counts in some cells. Before attempting to run a statistical model, I conduct exploratory data checking, and in cases where there appear to be major data sparsity issues, I abandon that analysis and regroup the data in such a way as to increase token counts. Occasionally, such data is simply presented and discussed in its raw form. I was particularly concerned during analysis about fitting random intercepts to datasets that included a large number of people with only one token. There is some evidence, however, to justify the use of mixed effects models even in such cases. In simulation analyses of mixed effects models with sparse data, Bell et al. (2008) found that while having a large proportion of singleton groups (that is, levels of a random effect with only one observation) can be problematic for the variance of the random effects structure itself, it does not interfere with the estimation of coefficients for the fixed effects. Since my analyses mainly concern the fixed effects, random intercepts for speaker (actually, singer/rapper, though I use the term 'speaker' throughout this chapter when discussing the random effects) were included in all final models.

The size of the dataset available for each analysis also has an impact on the model fitting procedures I employ. With large datasets, I use carefully defined backward modelling procedures (which will be described as they arise), but with small datasets, or small subsets of data, modelling is conducted in a more exploratory fashion. In such cases, a forward stepping procedure is used, in which the significance of variables of interest is tested one variable at a time, whilst keeping strong predictors in the model when they control for a large amount of variation.

2.4 BATH

The BATH lexical set provides something of a special case for this analysis of American dominance in singing accents due to the presence of the TRAP–BATH split in NZE and its absence in AmE (for a description of the process leading to this outcome, see Wells, 1982). It is of particular interest because this variation is relatively cognitively accessible to speakers/singers, and relatedly, because it involves a cross-dialectal difference at the level of the phoneme. Contrastiveness is one of the criteria for achieving marker status in Trudgill's 1986 discussion of salience in dialect contact settings, and another is a large phonetic difference between variants. BATH has both. In the discussion which follows, I will refer to the variable (BATH) and its two variants: SPMSS TRAP and NZE PALM.

Previous work analysing NZ singers has shown that vocal performers have particularly high levels of awareness for BATH (Coddington, 2004). Coddington found that five out of eight NZ singers, when asked about their singing accent, volunteered information about this variable. The artists in her study were also highly consistent in their realisation of BATH, in line with their stated intentions. In Australia, there is a strong genre distinction. O'Hanlon (2006) found 100% of BATH realised as TRAP (the SPMSS variant) in pop music but only 11% in hip hop, providing evidence for an own-accent style. O'Hanlon (2006) did not consider in detail whether the realisation of BATH with PALM (the variant for BATH in NZE and in most dialects of Australian English) in Australian hip hop was based on intentional avoidance of the American variant, or whether Australian English has actually become normalised in Australian hip hop. Note that ethnicity plays an important role in discourses of accent and authenticity in Australian hip hop (see Section 2.5.4 for further discussion of this). As we shall see, own-accent rap is less prevalent in NZ hip hop than it is in Australia. Amongst UK singers, Konert-Panek (2017a) found that Amy Winehouse realised BATH as TRAP in all occurrences across her two albums, while a particularly striking finding was presented in Konert-Panek (2017b). Adele went from not realising BATH as TRAP at all on her first album, to nearly 100% use of TRAP on her second and third albums, after achieving mainstream success. Additionally, an acoustic analysis of the tokens of BATH realised as TRAP had a higher F2 than sung tokens of TRAP from non-BATH words. This, along with the overall change in Adele's approach to BATH between albums provides strong evidence that she made an intentional shift to adopt SPMSS. Overshoot is a marker of stylisation, which in turn is a marker of intention. Adele's behaviour thus runs counter to the foundational assumption of this thesis.

With respect to the hypotheses outlined above, my expectations for BATH in the PoPS corpus are as follows:

1. Dominance Hypothesis: The American variant (realisation of BATH as TRAP) will be dominant in NZ music, alongside a minority of instances of the NZE variant, used by artists wishing to present an 'authentic identity'.

2. Accuracy Hypothesis: Testing this requires a consideration of individual artists' identity goals, and will thus be addressed once information on other variables has been gathered (Section 2.8).

   Any notable instances of phonetic overshoot in the realisation of BATH will be discussed. Overshoot or mis-realisation[16] of the American variant would imply intention and effort to perform the AmE style (counter-evidence to the foundational assumption of this thesis), while overshoot of the NZE variant would imply the opposite: the requirement of effort to produce the 'own-accent' variant (supporting the assumptions of the thesis). It should be noted here that the analysis of any instances of overshoot for this variable will be impressionistic rather than involving systematic acoustic analysis. A rigorous exploration of cues to stylisation, though it would be highly relevant, falls beyond the methodological scope of this project.

3. Genre Hypothesis: Genre should be the primary structuring variable, with the pop genre showing few signs of a singer's place of origin, gender and ethnicity. Hip hop, by contrast, is more likely to exhibit elements of the speech styles of the performers' speech communities. In this case, AmE is expected to be 100% consistent with respect to BATH, with all tokens using TRAP. Any differences between groups in the NZ pop data would thus provide evidence against this claim.

4. Salience Hypothesis: BATH is expected to be the most salient of the variables studied in this chapter. Conclusions about this will only be attempted, however, once we have information from the other variables.

[16]such as the use of /v/ in Cho's stylisation of Marlene Dietrich, based on a stereotype of German-accented English, despite Dietrich's use of /w/ (Bell, 2011).

2.4.1 BATH: Method

An auditory analysis was carried out for the 301 tokens of BATH that occurred in the corpus. Each token was designated as having either the phoneme TRAP (/æ/) or PALM (/aː/). In these broad terms, 254 tokens were realised as TRAP, and 47 as PALM. Initially, instances where BATH was realised as TRAP were also subdivided into those which were monophthongal (n=235) and those which were diphthongal (n=19, a realisation along the lines of [ei]). This [eɪ] variant of TRAP is grouped with the other TRAP tokens in the first analyses, and will then be discussed in its own right in Section 2.4.3.1.

Throughout the analyses of the PoPS corpus, I include function words in the datasets. Care was taken to exclude items realised as unstressed and having a reduced vowel, for all variables. Vowel reduction appears to be rarer in song than in speech, where each syllable has a rhythmic function. Given the limited size of the lexicon in pop songs (Murphey, 1992, though probably not in hip hop songs), function words are deemed to be an important part of the dataset, and any systematic variation that they exhibit will be controlled for with the inclusion of random intercepts on word in statistical models wherever possible.

2.4.2 BATH: Results

In the American data, BATH is near categorical, with 98% of the 167 tokens realised as TRAP. The TRAP–BATH split does not occur. In NZ songs, the realisation of BATH as TRAP is prevalent, with 67% of the 134 tokens using the American variant. Recall that realisation of BATH as TRAP is not attested in descriptions of NZE. It would be very rare in the speech of these performers. NZ pop is particularly strongly influenced by the American model, with 74% of the 88 tokens realised as TRAP. In NZ hip hop, this American influence is also strong, but to a lesser degree, with more than half of the tokens realised as TRAP (54%, n=46).

To examine the idea that BATH involves some kind of conscious identity decision, we can look at performers' consistency for the variable. Figure 2.2 provides two types of information. Firstly, it presents a histogram of the number of tokens of BATH that occurred for each individual. Secondly, it summarises each speaker's realisations of the variable, by grouping speakers according to whether they used PALM or TRAP categorically, or whether they used a mixture of the two variants. This highlights the fact that the majority of NZ performers were consistent in their realisation of BATH. There were 18 individuals with just one token, for whom, obviously, consistency cannot be assessed (14 out of 18 used TRAP). A further 18 vocalists had two tokens, and all of these individuals used the same realisation in each of the two occurrences (9 out of 18 used TRAP). Six of the remaining 17 artists used a mixture of realisations, nine used only TRAP, and two used only PALM. The 54 American artists reflect the pervasiveness of the TRAP–BATH split across US dialects, with just two people realising a total of three tokens of BATH as PALM.

Since some individuals (notably Justin Timberlake, who had 23 instances of BATH in the song 'Can't Stop the Feeling') have more tokens than others, the use of a random intercept for speakers in the modelling of this data is important. Rather than allowing Justin Timberlake's performance to dominate the results, the random intercept allows each individual to contribute equally to the model fit (see Section 2.3.4, above, for a justification of this approach despite the large number of individuals with just one token). The way to take the same approach in descriptive statistics is to show participant means, and also to show the mean of those participant means, when presenting the raw data (see Politzer-Ahles and Piccinini, 2018, for a detailed discussion of how best to represent raw data in a way that reflects mixed effects modelling). Looking at the NZ results again, if we take the mean of participant means instead of simply the mean of all tokens, the average rate of realising BATH as TRAP in NZ hip hop is 48% and the average in NZ pop is 78%. This process of taking the mean of participant means (which I will also refer to as the mean proportion in some cases), will generally be used whenever presenting or discussing aggregated raw data.

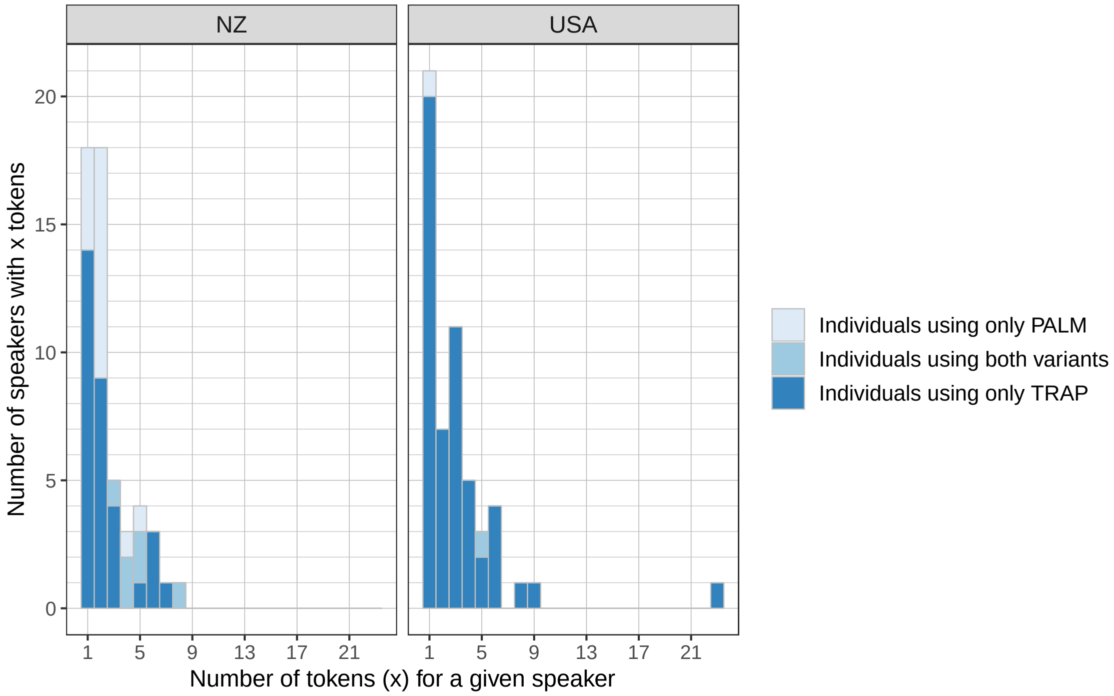


Figure 2.2: Histogram showing number of tokens per speaker, grouped by country, with colours summarising each speaker's realisation of BATH as either the SPMSS variant (TRAP), the NZE variant (PALM) or a mixture of the two.

Binomial mixed effects regression models were fit with the *lme4* package in R (Bates et al., 2015). Due to the small dataset, a forward modelling procedure was used, trying out main effects and interactions in simple models and then retaining those which reached significance. Significance of terms in the models were used as a guide, with p-values estimated using the *lmerTest* package (Kuznetsova et al., 2017). Final decisions about significance were, however, determined by log-likelihood comparison of minimally different models (conducted using the anova() wrapper function in R). As in all modelling procedures used throughout this thesis, the variance inflation factor (VIF) for each term in the model was calculated[17]. Models were initially fit with random intercepts for speaker and word, but the word intercepts had to be dropped to achieve convergence.

Independent variables which did not reach significance were gender and following

[17] using the vif.mer function, downloaded from `https://github.com/aufrank/R-hacks/blob/master/mer-utils.R`

environment (a two-level factor distinguishing between nasals and non-nasals). Additionally, a three-level factor combining gender and genre (female pop, male pop, male hip hop) was tested, but models with this term did not converge. Models were fit using either ethnicity (a four-level factor) or country (a two-level factor). While there appeared to be interesting differences between Māori/Pasifika and Pākehā in the former model, there were convergence issues, perhaps caused by the very low number of tokens of PALM amongst the US artists. The final model for the full dataset included an interaction of genre and place of origin (country), with a random intercept for speaker. Ethnicity will be explored in a second model, below. The output for BATH Model 1 is shown in Table 2.3. The model had a maximum VIF of 3.4, and the following structure:

BATH Model 1: TRAP ∼ genre * country + (1|Speaker)

Table 2.3: BATH Model 1.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.022	2.188	-3.209	0.001
genre=pop	15.797	3.428	4.608	<0.001
country=USA	18.758	5.292	3.544	<0.001
genre=pop:country=USA	-15.945	6.424	-2.482	0.013

Figure 2.3 shows the fitted interaction from the model, along with a summary of the raw data. In terms of the model predictions, both US and NZ pop artists realise BATH as TRAP categorically, as do US hip hop artists. NZ rappers however, are predicted by the model to realise BATH as PALM. Since speakers are largely consistent in their realisations, the model makes polarised predictions (near zero and one). Inclusion of the raw data shows that the variation is somewhat more nuanced, with a few NZ pop singers using PALM and several NZ hip hop artists using TRAP, along with six New Zealanders that use both variants. Note that the points on this and the following figures are plotted using the geom_jitter function within the *ggplot2* package (Wickham, 2016), which adds stochastic spread to the points, so that they are less overlapping. This makes it possible to determine the spread of the data, even in cases such as this where the majority of points are zero or one.

As mentioned above, an interaction of ethnicity and genre amongst NZ artists was found, but models of the full dataset that included this interaction did not converge. To pursue this, a second model was fit using just the subset of data by NZ artists. The interaction of genre and ethnicity was significant in a model which included a random intercept for speaker. Attempts to include an intercept for word led to non-convergence. The final model had a maximum VIF of 4.1. Table 2.4 shows the output of the model, which had the following terms:

BATH Model 2: TRAP ∼ genre * ethnicity + (1|Speaker)

Figure 2.4 shows, once again, the model fit in solid lines, the raw mean of participant means in dashed lines and the individual participant means in small crosses. The model predicts no differences between Māori/Pasifika and Pākehā artists in the pop genre. In hip hop, however, Pākehā artists are predicted to realise BATH as PALM, and hip hop artists to use TRAP.

The raw data reveals other trends not able to be modelled due to the small

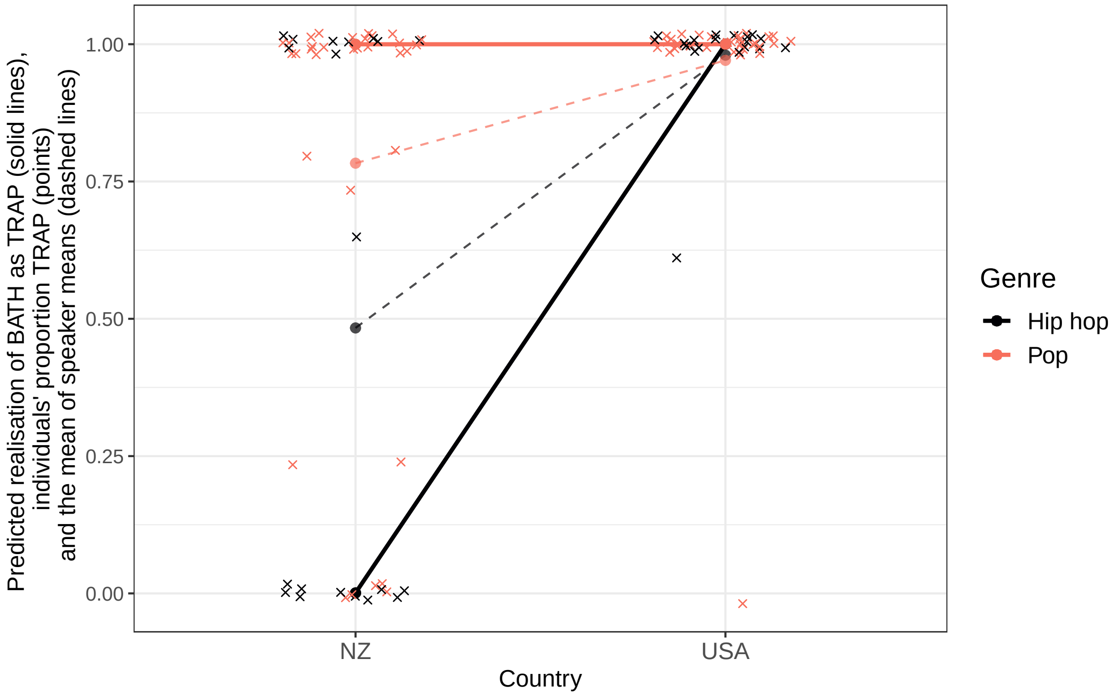


Figure 2.3: BATH Model 1: Predicted probability of realising BATH as TRAP according to genre and country of artist. Solid lines show the model fit, backtransformed to probabilities. Dashed lines show the mean of speaker means, and points show individual speaker means.

Table 2.4: BATH Model 2.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.734	2.562	2.628	0.009
genre=pop	1.342	2.713	0.495	0.621
ethnicity=pakeha	-14.154	4.277	-3.310	0.001
genre=pop*ethnicity=pakeha	15.484	5.109	3.031	0.002

number of tokens. Firstly, the difference between Māori/Pasifika and Pākehā rappers is actually small, with the mean of speaker mean percentage use of TRAP being 56% and 38%, respectively. It should be noted here that this genre by ethnicity interaction is only significant in a model that has a random intercept for speaker. This is because the main counter-examples to the trend happen to have a large number of tokens compared to other speakers. By including speaker intercepts, the model allows speakers to contribute to the model fit only once each. If we are to lump together all of the results into a single pool and then take means, the pattern is concealed, and in fact, the opposite genre by ethnicity interaction appears to exist in the hip hop data, with 68% of all tokens realised as TRAP by Pākehā hip hop artists, and just 42% of tokens by Māori/Pasifika artists. This difference between raw means and means of speaker means is driven by the performances of Māori/Pasifika artist David Dallas, who has five tokens, all realised as PALM, and the Pākehā vocalists in Machete Clan, who have six tokens all realised as TRAP.

Another trend in the raw data that is not captured by the model is that Māori/Pasifika

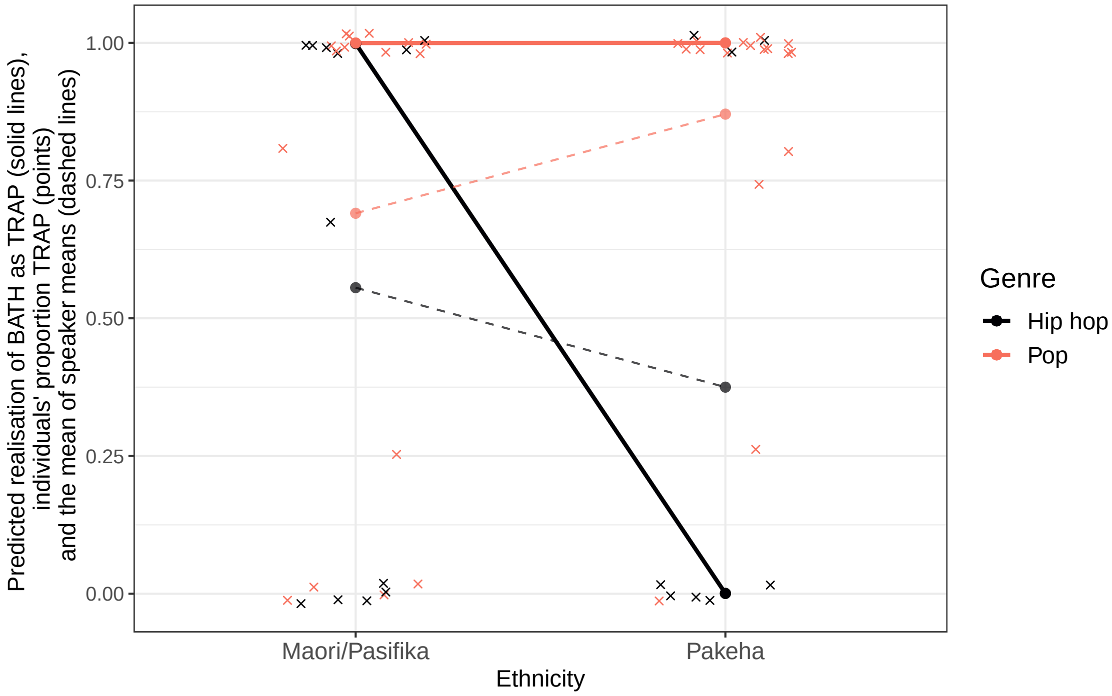


Figure 2.4: BATH Model 2: Solid lines show the predictions for the interaction of genre and ethnicity in the NZ data only. Dashed lines show the mean of speaker means and points show individual speaker means.

pop artists appear to use less TRAP than their Pākehā counterparts. This apparent difference is actually driven entirely by gender. The male Māori/Pasifika and Pākehā pop artists use TRAP 83% and 87% of the time, respectively, compared to 58% for female Māori/Pasifika pop singers, who use TRAP at a rate more consistent with the results for hip hop. These trends will be discussed below, as we consider how the results of this analysis fit with the hypotheses laid out earlier.

2.4.3 BATH: Discussion

The discussion begins by summarising the results in the context of the four hypotheses:

- Dominance Hypothesis — As hypothesised, the SPMSS variant (realisation of BATH as TRAP) was prevalent in NZ performances.
- Accuracy Hypothesis — No cases of qualitative overshoot (Gibson and Bell, 2010) were found in TRAP realisations, however there were some interesting examples of overshoot for realisation of BATH as PALM. NZE PALM is a front vowel [aː], while the PALM/FATHER vowel in AAE is backed [ɑː]. Several NZ artists appear to have chosen NZE at the phonological level, by aligning the BATH lexical set with the PALM phoneme, but then applied SPMSS (or AAE) phonetics to that phoneme, realising it as [ɑː]. One example of this comes from Edgar, an MC in the 9–5ers. His realisation of the vowels in BATH, START and LOT have similar F2 values. Despite having selected the NZE variant for BATH in a binary, phonemic, sense, Edgar applies a retracted realisation of the vowel which aligns with START and LOT, in the words *pass* (F2=1210Hz),

card (F2=1280) and *job* (F2=1230, see 2.7 for discussion of LOT). This reveals an intention to use his own accent at a phonemic level, but an adherence to SPMSS/AAE at the level of phonetic detail. This example is particularly salient given the involvement of these three words in a sequence of rhymes. To use fronted vowels in *pass* and *card* would have spoiled the rhyme with *job*. There are other instances of BATH realised with a US-like variant of PALM by various artists in the corpus, and together, these instances provide anecdotal evidence that NZ vocalists may have trouble performing the gestures associated with their own native speech style in the context of song.

- Genre Hypothesis — The model predictions supported the hypothesis that the artists' place of origin would be undetectable in pop, but the raw data provides counter-evidence to this claim. Most notably, there was an ethnicity difference amongst female pop singers, with Pākehā artists adopting the SPMSS variant at a much higher rate than Māori/Pasifika artists. The adoption of own-accent styles by female Māori/Pasifika pop singers was not predicted.

  Regarding diversity in hip hop, one of the key findings for BATH which warrants discussion is that Pākehā rappers were more likely to adopt an own-accent style than Māori/Pasifika rappers. As it turns out, the same ethnicity by genre interaction will be seen again in the realisation of non-prevocalic /r/, and will be discussed in more detail there (in Section 2.5.4), where I will interpret the finding as Pākehā rappers adopting NZE and Māori/Pasifika rappers adopting HHNL. Both styles are related to dimensions of hip hop authenticity (McLeod, 1999).

- Salience Hypothesis — Results (so far) were inconclusive with respect to the salience of BATH. This hypothesis will become more testable later in this chapter, once we have data from multiple variables in hand.

  There was some support for the idea that BATH forces artists to make and implement a conscious decision about their pronunciation. The majority of NZ artists did stick categorically to one variant or the other, but there were also cases of mixed usage. These cases provide evidence against the claim that NZ singers consistently enact a static identity goal whenever they encounter a BATH word, and show signs of conflict, uncertainty, or unawareness for these performers. There appeared to be a pattern in such cases, however, with the NZ variant occurring in more contextually salient environments such as at the end of an utterance, or in low frequency (more informative) words.

  This hypothesis will be considered again once we have data on other variables. It should be noted that categorical usage of TRAP does not provide evidence of high levels of awareness, since it is in line with the general adoption of SPMSS in NZ singing (Gibson, 2010b). Clear support for this hypothesis would have come from a case where all NZ artists were fully polarised in their usage — this was not the case.

The findings above provide mixed support for the hypotheses. A more nuanced methodology could go further — this auditory analysis of BATH was quite rudimentary. An acoustic analysis could test whether cases of BATH as TRAP in NZ performances are phonetically indistinguishable from those of US performers. This could also be tested by asking naive listeners to categorise extracts according to the origin of the vocalist. Would they perform at chance? Future work could also examine whether NZ TRAP vowels conform to AmE phonological rules, such as tensing and raising of TRAP prior to nasals. Gibson (2011) found that the NZ comedy duo Flight of the Conchords were sensitive to these patterns of allophonic variation in TRAP in a stylised performance of Barry White. Which rules are adopted by NZ artists could potentially be predicted and analysed according to Yang's Tolerance Principle (Yang, 2016).

Before finishing this discussion of BATH, I present a qualitative analysis of a passage from the song 'Hungover' by Māori/Pasifika pop act Sons of Zion. The word *dance* is realised in a way that, to my ears at least, gives away uncertainty in the singer's mind, about how to produce the vowel. While my binary auditory categorisation placed the vowel in the /aː/ phonemic category, the vowel begins somewhat raised and fronted then lowers and retracts slightly, as if the singer made an on-the-fly adjustment, aborting TRAP after the start of the vowel and rushing to PALM.

Sons of Zion adopt SPMSS most of the time, with occasional use of NZE that may signal a desire to project an 'authentic identity'. An example of this comes in the line prior to the instance of *dance* under scrutiny here. That line ends with a notably NZE-like realisation [ʉː] in the word *soon*. The fronted GOOSE token[18] occurs in an extended position at the end of a musical phrase. The word *soon* is actually highlighted in the final mix with the use of a delay effect that spans the musical rest that follows — it is 'on display'. This token is an excellent example of the kind of contextually salient environment proposed in the Salience Hypothesis to make it easier to enact identity goals. It is particularly interesting since GOOSE does not appear to have strong sociolinguistic salience in NZ singing, despite having a strong phonetic distinction between NZE and SPMSS. The NZE-like vowel realisation in *soon* is unusual in the context of the corpus. Due to factors such as the elongation of the vowel and the long musical rest which follows, this particular token is *contextually salient*, it 'juts out' from the local context. While GOOSE has limited *sociolinguistic salience* as a marker that can be used to project NZ identity in song, it has come to the attention of this Sons of Zion singer in this instance, and thus also becomes a chance to display his 'authentic identity'.

Back to the word *dance*: why does the word come out with the apparently conflicted realisation described above so soon after a notably NZE-like word? In between *soon* and *dance* come the lyrics 'and I can't pretend I'm not missing you', with SPMSS pronunciation of both *can't* and *not* (high frequency, low informativity lexical items in the middle of a phrase, thus having limited contextual salience here). Whatever the singer's reasons for switching back to SPMSS in that phrase, it is likely that both NZE and SPMSS would be cognitively activated for the singer at this point in time, be it in terms of competing phonological rules, clusters of episodic memories, or both. What is important about this token of *dance* is the direction of the change, beginning with the SPMSS variant and then quickly shifting towards the NZE one. This implies a shift from the automatic to the conscious — from a responsive to an initiative style. While this is merely an anecdote, it does — maybe

[18]Which also has a slightly fronting vowel trajectory, characteristic of spoken NZE, and in contrast to the backing and rounding that occur across the trajectory of the SPMSS variant of GOOSE, as described in (Gibson, 2010b, pp. 90–92).

— provide a glimpse into the executive functions underlying this singer's language production.

2.4.3.1 Variable pronunciation of the word *can't*

There is one more aspect of the BATH dataset worthy of discussion. As noted in the methods above, during coding of the BATH vowels a diphthongal [eɪ] variant was noted. After tracking this as a third auditory category alongside [ae] and [aː], a total of 19 such tokens were found, and as it turned out, they were all in the highly frequent lexical item *can't*. Diphthongal TRAP is a feature of AAE, and also Southern USA speech (Clopper and Pisoni, 2004). Analysis of the 98 tokens of *can't* that appear in the entire PoPS corpus (and which make up almost a third of the BATH dataset) revealed that [eɪ] is used almost exclusively by African American vocalists (12 of the 19 tokens) and a few European American artists from the South (3 tokens). There is no significant difference between pop and hip hop in its use. A chi-square test, excluding PALM realisations, comparing [æ] vs. [eɪ] realisations for African American vocalists vs. those of other ethnicities was significant (this constitutes 16% of African American tokens, vs. less than 3% of tokens produced by the other three ethnic groups. Post Malone, despite being white, can claim the authentic use of this variable because he spent his formative years in Texas, while white rappers from other parts of the USA avoid this variant entirely. It is used, however, by one NZ rapper: Name UL. These tokens are consistent with his stylised use of HHNL, drawing on the social capital associated with African American speech and rap styles, including the presentation of masculinity. The use of this variant by non-African American, non-Southern performers may actually carry with it a heightened risk of raising 'inauthenticity flags'. This variable may be an example of a linguisic form restricted to use by certain groups, namely African Americans and those from the South of the USA. While involving much lower stakes, the processes involved here may have some parallels to the avoidance of the 'the N word' by non-African American performers (Low, 2007; Cutler, 2014). Even though the sociolinguistic variable is a primary conceptual tool for analysing phonetic variation, this example reminds us that this variation takes on its meaning in the context of specific words. Lexical items have their own histories in memory. Phonemes are abstracted out of phonetic commonalities between words, rather than being cognitively prior to them. Finding evidence for lexical effects often requires a large amount of data. The next section, which looks at the highly frequent variable, non-prevocalic /r/, provides enough data to look for lexical effects in the phonetics of popular music performance.

The analysis of the BATH variable has shown us high rates of adoption of a SPMSS form by NZ artists. I have argued that this variable is special due to the united front presented by US artists, and its involvement in rhyme. Non-prevocalic /r/, while widely known as a US feature is less likely to cause a point of decision for singers, since it is variable. Deciding whether or not to realise an /r/ is a lower-stakes identity choice than deciding whether to realise BATH with PALM or TRAP.

2.5 Non-prevocalic /r/

With the exception of a small population in the south of the South Island (Villarreal et al., 2019), New Zealand is largely non-rhotic, however there are early signs of the emergence of rhoticity in multicultural Auckland, and in Pasifika communities (Gibson, 2016; Marsden and Holmes, 2014; Marsden, 2017), particularly in the NURSE lexical set. The USA, by contrast, is largely rhotic, with some exceptions in New England (Carmichael, 2017), the South, and in AAE. The existence of variation in both countries according to region and ethnicity means there is plenty of scope for testing the hypotheses laid out in Section 2.2 on this variable, particularly since it is highly frequent.

In sociolinguistic analyses of non-rhotic varieties such as NZE, a distinction is made between pre-vocalic and non-prevocalic /r/ (Hay et al., 2018). Studies of rhoticity in rhotic areas, on the other hand, tend to treat all cases of post-vocalic /r/ as belonging to the same variable, with conditioning based on following environment (Rácz, 2013). The former approach is taken here, beginning in this section with the analysis of non-prevocalic /r/.

2.5.1 Prior analysis of rhoticity in NZ and US music

In an unpublished conference paper (Gibson, 2010a), I presented an analysis of 3,352 tokens of potential non-prevocalic /r/, mainly in NZ and US rap/hip hop, along with a selection of songs from a range of other genres. The results presented in this section are essentially a replication of that study, but using the carefully balanced set of songs collected for the PoPS corpus, and much more careful methods for determining /r/ presence or absence.

In Gibson (2010a), NURSE environments favoured the realisation of /r/ across all groups (as is generally found in studies of rhoticity). In NZ and USA rap, rhoticity was in fact almost completely limited to the NURSE environment. NZ and US rap had 88% and 97% /r/, respectively, in NURSE words, and just 1% and 3%, respectively, in non-NURSE words. This systematic partial rhoticity was also found in my study of NZ Pasifika hip hop artists (Gibson, 2005). In the mixture of other genres analysed, rhoticity occurred in both environments, with NZ artists producing /r/ in 63% of NURSE words and 22% of other words, and US artists producing /r/ in 94% of NURSE words and 47% of other words.[19] There was one additional result presented in Gibson (2010a) which has not been replicated for the present thesis, and thus stands as an important piece of background information: analysis of radio and television interviews provided comparative information about the state of rhoticity in the *speech* of NZ rappers. There was no use of non-prevocalic /r/ in any non-NURSE environments. However, for the nine NZ rappers (including just one Pākehā artist, Tom Scott) for whom interviews were found, seven (including Scott) had at least some use of /r/ in NURSE in their interview speech style, at rates of between 20–100%. Four of the rappers whose interview speech was analysed in Gibson (2010a) are present in the PoPS corpus (PNC, Savage, Young Sid aka Sid Diamond, and Tom Scott of Homebrew).

There were several problems with the methods of this previous study:

[19] As a point of reference, in the UK context, Konert-Panek (2017a) found 30% coda /r/ on Amy Winehouse's first album, and 43% on her second.

- Genre and ethnicity were conflated, with most, but not all hip hop artists being African American or Māori/Pasifika in the hip hop sample, and with a more balanced range of ethnicities for the other genres.

- Artists were selected unsystematically, according to my own music listening taste and experiences. I chose artists I deemed to be important, influential or iconic, which in the case of hip hop meant that most tracks came from the 1990s and early 2000s, while release years for tracks in the other genre category spanned half a century.

- The auditory analysis was done in one quick sweep of the data with no information about lexical items or phonological environment (other than the NURSE vs. non-NURSE distinction) recorded, nor any checking of intra-rater reliability, nor any consultation with the spectrograms of the recordings. This may have led to confirmation bias, and the 'imagining' of /r/ in places where it tends to occur most often. This illusory perception of /r/ has now been robustly attested to occur, at least for untrained listeners (Hay et al., 2018), and to at least some extent for trained listeners too (Villarreal et al., 2019).

The results of Gibson (2010a) provide a baseline for the present analysis of the PoPS corpus. The initial motivation for developing the PoPS corpus was, in fact, to take a more systematic approach to the analysis of this variable, finding artists to fill each of the demographic cells, and conducting a more careful, albeit still auditory, analysis. The method section below outlines in detail how that analysis was conducted, but first, I present the expectations for this variable with respect to the four main hypotheses of this analysis.

- Dominance Hypothesis — Non-prevocalic /r/ will be present in NZ vocal performances, including in non-NURSE environments. Use of /r/ in NURSE words will not be taken as a clear marker of SPMSS style, but rhoticity in other environments will.

- Accuracy Hypothesis — By using the term SPMSS, I draw attention to the fact that the dialect hypothesised to be natively represented in the minds of NZ singers and rappers in the context of music is not any one dialect of American English as spoken in the USA, but rather a levelled variety derived from their exposure to recorded music. As such, the hypothesis that performers will be 'accurate' actually refers to accuracy with respect to an individual's specific exposure. If they have encountered certain words much more in music than in conversation, then the representation of the SPMSS variant should be more strongly encoded, with less competition from their experience of speech. This can be tested by calculating the relative lexical frequency of words in song lyrics and speech, and using that ratio to predict behaviour. The hypothesis that NZ artists will accurately replicate SPMSS/HHNL thus takes on a lexical dimension in the analysis of this variable: words that are over-represented in popular music as compared to speech, referred to here as *songy words*, should exhibit more use of non-prevocalic /r/ by NZ performers.

- Genre Hypothesis — Amongst pop artists: place of origin, ethnicity and gender should not contribute significantly to the structure of the data. If gender patterns do exist in the US, however, they should be paralleled in NZ. Pop singers in NZ will use /r/ in a similar way to their US counterparts, and to a similar degree.

Amongst hip hop artists: Many NZ and US hip hop artists will share a similar style, based on AAE, with a larger NURSE vs. non-NURSE distinction in rhoticity levels than will be seen for pop artists. Examples of non-rhotic singing/rap in the NZ dataset should appear mainly in hip hop, and come from artists that intend to project an 'authentic identity'. The hip hop data will also reflect the demographic characteristics of 'own-accent rappers'. For Pākehā rappers, this would mean overall avoidance of rhoticity, while realisation of /r/ strictly in NURSE environments would be a marker of a Māori/Pasifika identity (cf. Gibson, 2016).

- Salience Hypothesis — Rhoticity is expected to be a salient identity marker, but somewhat less salient than BATH. A subset of those NZ artists who were found to avoid the SPMSS variant of BATH are thus expected to also enact their goal of using NZE with this variable. This would mean the avoidance of /r/ in non-NURSE environments, where /r/ is a salient marker of SPMSS.

2.5.2 Non-prevocalic /r/: Method

An auditory analysis was conducted for 3659 tokens, along with visual inspection in Praat. A script was used to automate the mechanical aspects of the process, conducting the following tasks: opening soundfiles and textgrids; locating and playing the target words and their surrounding context; opening a dialogue in which to enter the analysis code, along with an option to insert additional comments about the token; inserting the annotations into a new point tier and saving the edited textgrid.

Extreme care was taken to provide a quality categorisation of the data into /r/ and /r/-less tokens. In recognition of the fact that /r/ is not a binary variable, but rather a very complex package of both temporal and spectral cues (see, for example, Villarreal et al., 2019), a relatively complex system was used to encode detailed information about each token. I will step through each of the pieces of information recorded here, to give a nuanced characterisation of the data collected, even though this is ultimately collapsed into a binary /r/ present vs. absent distinction. Of the 3659 tokens originally exported from LaBB-CAT, 58 were excluded due to the candidate token being followed by another /r/, or due to mistranscription. For the remaining 3601 tokens, nine main codes were used to denote the type of realisation. Six of these were for non-prevocalic environments (amounting to a total of 3242 tokens) and three were for linking environments (359 tokens).

2.5.2.1 Coding scheme for non-prevocalic /r/ environments

The codes for non-prevocalic tokens included one code to mark complete absence of /r/, and three to capture varying degrees of post-vocalic /r/ presence. These distinctions reflected the perceived degree of constriction and length of the /r/. There were 1976 tokens where I was confident that /r/ was completely absent, 156 tokens that were marked as having a subtle post-vocalic /r/, a further 214 tokens with a moderately strongly produced /r/, and 539 tokens where a strongly realised post-vocalic /r/ segment was perceived. In addition to these main categories, there were 324 tokens of rhoticised vowels, where more than half of the length of the vowel was perceived to be /r/-coloured. Many of these tokens did not have a post-vocalic consonantal /r/ segment, despite still clearly counting as examples of rhoticity. Finally, there were 33 tokens where a vocalic offglide gave me the initial impression of an /r/ segment, despite the absence of any actual rhoticity. For example, a non-rhotic FORCE vowel realised as [fɔːəs] can be initially misperceived by a non-rhotic listener as containing /r/ if care is not taken.

Ultimately, these six categories were collapsed into a binary code: the four categories denoting some degree of consonantal post-vocalic /r/ were grouped with the rhoticised vowel tokens, yielding 1233 instances of /r/-presence, while the non-rhotic offglide tokens were grouped with the no-/r/ tokens, yielding 2009 /r/-absent tokens.

2.5.2.2 Distinguishing non-prevocalic and /r/-sanhdi environments

Whether an instance of potential post-vocalic /r/ would be in a non-prevocalic or a linking environment could not be judged solely from the transcript, since pauses between words can occur in unexpected places, especially in the context of a song. The distinction between non-prevocalic and sandhi environments proved to be quite trivial to determine in most cases. There were some cases, however, where a linking /r/ had a glottal stop prior to the vowel. Such cases were treated as sandhi environments so long as the glottal gesture did not constitute the start of a new musical or prosodic unit, and was sufficiently short (the longest was 91ms). Cases where there was deemed to be a pause, or prosodic boundary between the /r/ and the vowel-initial word, were treated as non-prevocalic environments. The 359 tokens deemed to be sandhi environments were separated out as a distinct dataset, and will be discussed in section 2.6 below.

2.5.2.3 Blind re-analysis of difficult tokens

I was very much aware of the subjectivity involved in the identification of /r/, and there were many cases where the presence or absence of /r/ was not clear-cut. To deal with this, I kept track of my own uncertainty by adding a suffix code to tokens where my confidence in the code assigned was particularly low. Considering the full dataset (including sandhi environments), a total of 538 tokens were marked by an uncertainty code, while a further 70 tokens were noted to be difficult to assess due to being obscured by the instrumentation of the song. Another 235 tokens were marked with a manual comment of some kind, with the majority of these being a note about the identity of the following phoneme, in cases where the transcript alone would be misleading. Other tokens were marked as needing further assessment simply because the crucial stretch of audio was outside of the audio clip exported by LaBB-CAT.

Any token marked with either the low-confidence code or the obscured-by-music code was analysed again, with the tier of the textgrid containing the original code removed to allow for blind reanalysis. For this re-analysis phase, a five-way coding scheme was used, with a binary code for non-prevocalic /r/ environments (presence, absence) and a three-way code for sandhi environments (linking /r/, glottal, vowel hiatus). Agreement between the original analysis and the recheck was conducted after collapsing together the six original codes for non-prevocalic /r/ into a binary code. The script used to run the blind recoding process then re-appended the textgrid tier with the original code prior to saving the edited textgrid.

At each stage of blind re-analysis, a random sample of non-problematic tokens was included so as to keep a clear auditory reference point to my previous coding system. For these 150 non-problematic tokens, the check-recheck agreement rate of the two analyses (based on agreement according to the five-way coding system) was 97%. For the tokens marked as problematic, however, this reanalysis yielded a relatively low intra-rater agreement rate of only 74%. A third blind listen was conducted for those tokens where the first two analyses differed, and the majority code was then entered as final. Rechecking was also done for the handful of discrepancies found for the randomly selected non-problematic tokens. Any tokens that were marked as being obscured by the instrumentation on both the first and second pass were excluded from the dataset (n=16). The 235 tokens with additional comments were attended to manually, resolving a range of matters such as correcting the following phoneme.

2.5.2.4 Methods of statistical analysis

Since rhoticity is a frequent variable that yielded a relatively large dataset, a more systematic approach to statistical modelling could be taken, along with the assessment of a wider range of independent variables than was the case for either BATH above or LOT below. Five models are presented. Due to the absence of female hip hop in the corpus design, it was difficult to model all social variables at once without facing convergence issues. For this reason, the first model considers all of the data, focusing on the difference between NZ and US artists, along with genre and gender. The second uses just the subset of data for males, allowing a focus on ethnicity and genre, while focusing on the pop subset of data allows us to look at gender in the third model. The final models look at the role of region for vocalists from the USA.

The modelling process was as follows: to begin with, a binomial GLMER model with all main effects was run with intercepts for speaker and word. The VIFs were checked for this model and any explanatory variables with VIF > 15 were iteratively removed until the main effects were suitably non-collinear for the purposes of pruning non significant effects.

The dependent variable was a binary split of /r/ presence versus absence. The independent variables considered to have social meaning were as follows: ethnicity, genre, gender, chart type, and the preceding vowel. The lexical predictors were the frequency of the word in the LyricsPlanet website, and a binary factor to represent songiness of the word (see Section 2.3.3). Phonological factors considered included manner and place of following phoneme, position (treated as a 3 level factor: word internal, word boundary, and pre-pausal), and the length of the word. Additionally, to control for potential effects of self-priming (Clark, 2018), a measure of whether an individual's previous token of potential non-prevocalic /r/ had been realised with or without /r/ (which I will refer to as PrevR). Since GLMER drops all observations with missing data, the first token for each speaker was coded as 'unknown' for this variable, so that no observations would be dropped.

Phonological factors were only tested as main effects, not in interactions.[20] Social

[20]Non-significant phonological factors were initially tested in all 2-way interactions prior to removal, but this led to over-fitting and was ultimately rejected.

and lexical factors were tested in all 3 way interactions. Significance cut-off was p<0.05 (with one exception discussed below), and variables were removed in order of least significance/highest p-values in co-efficients, with all final decisions based on log-likelihood comparison of minimally different models. VIFs were tested at several points in the process, to ensure the avoidance of multi-collinearity in the model. Once a model was achieved where all fixed effects were significant, slopes for the remaining factors were added. Slopes accounting for least variance were then removed until the model converged. Anything not reaching significance after the addition of slopes was also removed. Finally, the model was tested again for multicollinearity, with a maximum allowable VIF of 10 for all final models.

This process was undertaken rigorously prior to making a few final changes to the dataset, which included the following: the addition of the final Pākehā hip hop track (Machete Clan's 'On the rark'), the reasons for which were described in 2.3.1.3); the exclusion of the written portion of Celex from calculations of lexical frequency in speech (see Section 4.3.2.5); the decision to include the control for self-priming, PrevR, and finally, the decision not to allow interactions for phonological predictors. The final models from the rigorous model fitting procedure were fit again after these changes were made. Some phonological factors became non-significant, and some slopes caused convergence issues and were removed. None of these changes affected the main results or interactions, or the resulting conclusions however.

2.5.3 Non-prevocalic /r/: Results

2.5.3.1 Rhoticity Model 1: All data

Table 2.5: Rhoticity Model 1, testing the binary distinction between NZ and USA artists (based on the full dataset for non-prevocalic /r/).

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.149	0.044	166.467	3.361	0.001
genre=pop	0.156	0.051	127.735	3.082	0.003
country=USA	0.254	0.060	138.468	4.246	<0.001
lexicalSet=nurse	0.596	0.062	124.137	9.625	<0.001
wordSonginess=songy	0.050	0.026	505.260	1.881	0.060
previousToken=Rpresent	0.078	0.016	3133.997	5.014	<0.001
previousToken=NA	0.013	0.029	3043.633	0.462	0.644
position=wordBoundary	-0.094	0.027	175.449	-3.535	0.001
position=prePause	-0.013	0.036	84.163	-0.376	0.708
genre=pop*country=USA	-0.167	0.073	131.864	-2.285	0.024
genre=pop*lexicalSet=nurse	-0.204	0.074	109.358	-2.745	0.007
country=USA*lexicalSet=nurse	-0.175	0.083	91.715	-2.107	0.038
country=USA*wordSonginess=songy	-0.054	0.030	3143.112	-1.846	0.065
genre=pop*country=USA*lexicalSet=nurse	0.235	0.104	100.979	2.245	0.027

In presenting the results of each of these models, predicted values are plotted along with raw data, calculating the mean rate of /r/ for each participant, along with the mean of participant means for each group. Rhoticity Model 1 examined the full dataset, focusing on the difference between US and NZ by including a binary variable for country of origin rather than a four-way factor for ethnicity. The fixed

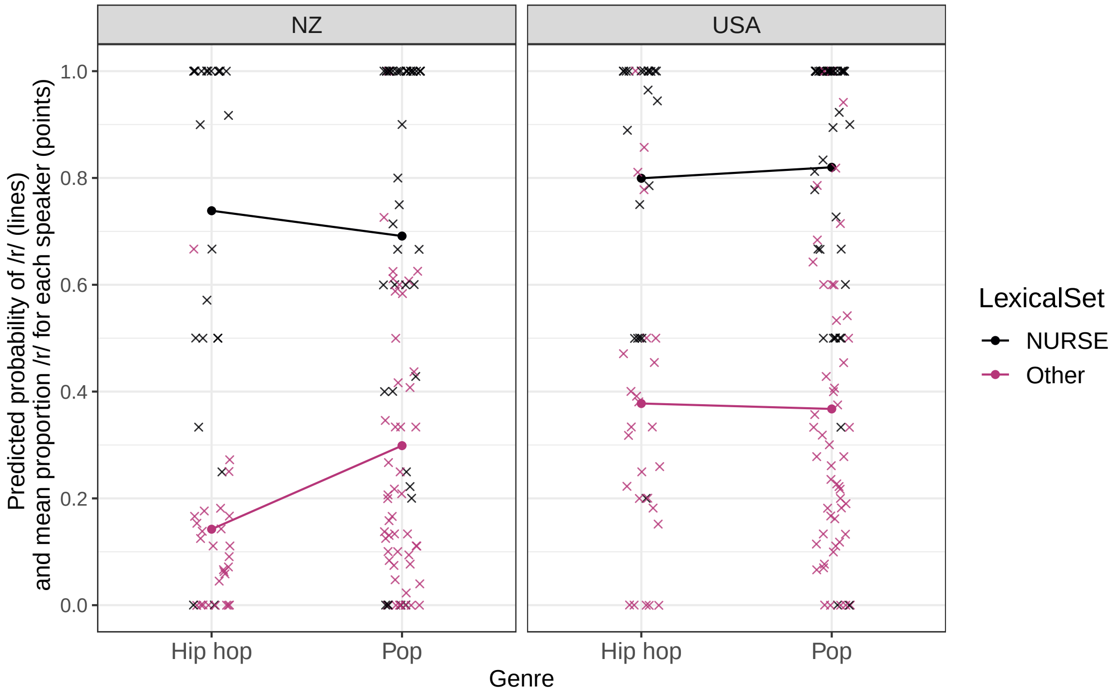


Figure 2.5: Rhoticity Model 1: Predictions from three-way interaction (lines) plotted with individpal speaker proportions of /r/ realisation (points). Each individual is represented by two points, one summarising their mean proportion of /r/-presence in NURSE environments (black) and the other their rate of /r/ when preceded by other vowels (purple).

effects are presented in Table 2.5. The model had a maximum VIF of 5.8, and had the following structure:

Rhoticity Model 1: r ∼ genre * country * NURSE + songy * country + PrevR + position + (1 + NURSE + position | Speaker) + (1 | word)

The two interactions in the model will be reported in detail below, but first I describe two other main effects. Firstly, /r/ was most likely to be realised when it occurred word-internally or when followed by a pause, and significantly less-likely in cases where the /r/ was word-final, and followed directly by the first consonant of the next word. Secondly, there was a greater likelihood of producing /r/ if /r/ had also been realised in the previous token for a given individual. This self-priming effect has been well documented with respect to syntactic structures (Bock, 1986; Branigan et al., 2000), and more recently in phonetics (Clark, 2018).

The three-way interaction between genre, country, and whether the word belonged to the NURSE lexical set or not is shown in Figure 2.5, along with participant mean rates of /r/. There are several findings captured in this graph which are worthy of note. The most important finding is that a very large majority of New Zealand vocalists are at least partially rhotic in their singing or rap. The predicted rate of rhoticity in NZ pop is around 10% lower than that for US music in both NURSE and non-NURSE environments, and there is no difference between genres in the US data. Words in the NURSE lexical set favour /r/ in all groups, but this effect is strongest in NZ hip hop, where there is much less use of /r/ in non-NURSE environments than there is for the other groups.

A second interaction is shown in Figure 2.6. The trend towards an interaction of

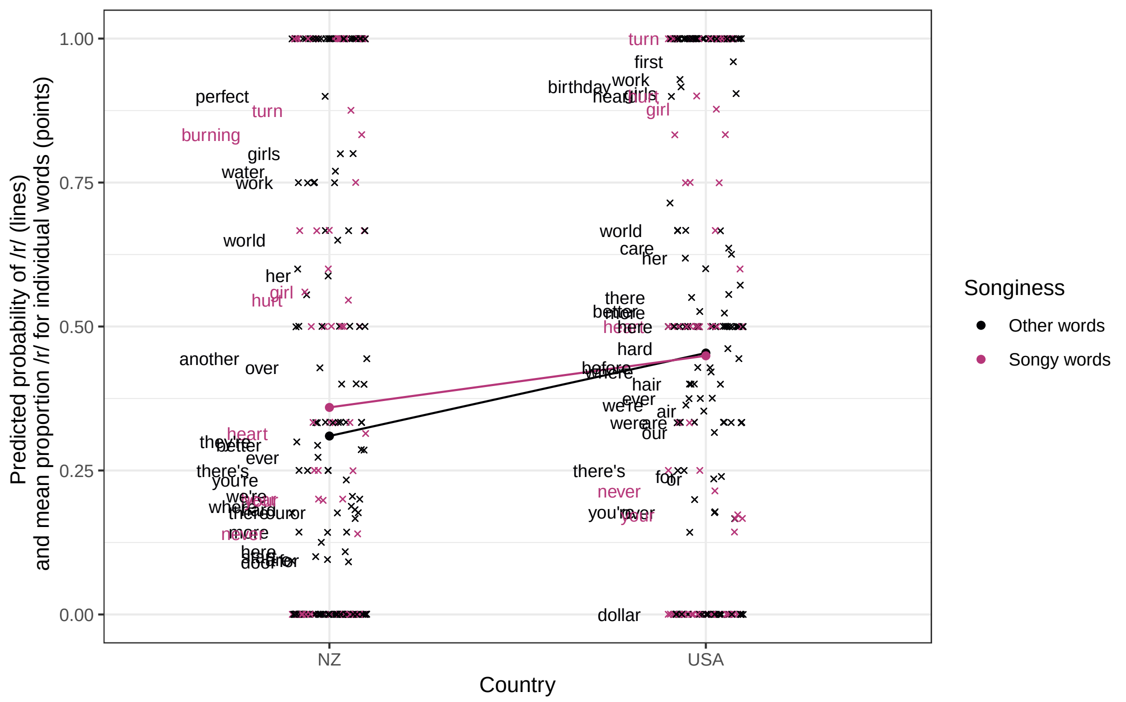


Figure 2.6: Rhoticity Model 1: Near-significant interaction (p=0.065) of word songiness with country (the upper tertile of words, when taking the ratio of song to speech lexical frequencies, are coded as 'songy'). Model predictions (lines) plotted with mean proportion /r/-presence for songy and other words (points). Word labels shown to the left of points that summarise 10 or more tokens. Points and word labels are horizontally jittered to improve readability.

songiness with the artist's country of origin was retained in the model even though it did not reach significance (log-likelihood model comparison p=0.065). The reason for keeping this trend in the model and presenting it here is that it speaks to one of the wider questions of this thesis: are cognitive representations of words structured according to the contexts in which they were encoded? This model predicts that NZ artists are more likely to be rhotic in words that occur disproportionately often in song lyrics as compared to spoken conversation. Figure 2.6 shows this trend towards an interaction in solid lines, and also plots the proportion of /r/ realised in songy and other words. For the purpose of exemplifying the identity of songy vs. non-songy words, some labels are included on the plot, for those items where the mean is based on at least 10 tokens. Of course this raw data is heavily influenced by whichever artists happened to produce the given words, so it is difficult to draw any pattern from the raw data itself. It is by holding constant the variation between speaker groups, individual speakers and individual words (through the random intercepts) that the model is able to extract this subtle frequency-based pattern.

Note that the original model fitting procedure used the ratio of song frequency to mean speech frequency (the average of Canterbury Corpus, Buckeye and CobS) in its original (very right-skewed) form. With this version of songiness, the interaction with country is highly significant (p=0.006), with NZ vocalists predicted to produce /r/ more in words that are more strongly over-represented in song lyrics. When this ratio is logged to produce a closer to normal distribution, however, the interaction fails to reach significance (p=0.13). This suggests that the interaction is being driven

Table 2.6: Rhoticity Model 2, based on male data only.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.215	0.367	-0.585	0.559
genre=pop	-0.310	0.459	-0.675	0.500
ethnicity=africanAmerican	-0.832	0.511	-1.629	0.103
ethnicity=pakeha	-2.577	0.587	-4.388	<0.001
ethnicity=maoriPasifika	-2.423	0.474	-5.117	<0.001
lexicalSet=nurse	2.480	0.516	4.806	<0.001
position=wordBoundary	-0.650	0.225	-2.890	0.004
position=prePause	0.033	0.309	0.105	0.916
followingPhoneme=fricative	0.357	0.169	2.110	0.035
followingPhoneme=sonorant	0.025	0.179	0.137	0.891
previousToken=Rpresent	0.431	0.144	2.989	0.003
previousToken=NA	0.386	0.259	1.493	0.135
genre=pop*ethnicity=africanAmerican	0.524	0.711	0.737	0.461
genre=pop*ethnicity=pakeha	1.354	0.774	1.748	0.080
genre=pop*ethnicity=maoriPasifika	1.657	0.674	2.457	0.014
genre=pop*lexicalSet=nurse	-0.780	0.734	-1.063	0.288
ethnicity=africanAmerican*lexicalSet=nurse	-0.034	0.695	-0.049	0.961
ethnicity=pakeha*lexicalSet=nurse	0.546	0.868	0.630	0.529
ethnicity=maoriPasifika*lexicalSet=nurse	1.978	0.669	2.956	0.003
genre=pop*ethnicity=africanAmerican*lexicalSet=nurse	1.493	1.031	1.448	0.148
genre=pop*ethnicity=pakeha*lexicalSet=nurse	0.228	1.121	0.203	0.839
genre=pop*ethnicity=maoriPasifika*lexicalSet=nurse	-2.155	0.989	-2.178	0.029

by the particularly songy words. As a compromise, the binary factor described above was created, separating the top third of the words in the dataset as ‘songy’ words, and leaving the remaining two thirds of words as ‘other’.

This is an important result, even though it is weak. Since it relates more closely to the themes explored in the second half of the thesis, I will save further discussion until the findings of both the production and perception components of this project are brought together in Chapter 5.

2.5.3.2 Rhoticity Model 2: Male data

In order to look more closely at ethnicity and genre, the second model includes only the male artists. In this way, the datasets for pop and hip hop are more balanced and can be better compared. After carrying out the same process of backward modelling as that described above (including refitting the final model after some changes to the dataset and consequently making minor adjustments to the model), the following final model was fit, which had a maximum VIF of 7.3 (for the interaction of genre with NURSE):

Rhoticity Model 2: r $\sim$ genre * ethnicity * NURSE + position + manner + PrevR + (1 + position | speaker) + (1 | word)

The output is shown in Table 2.6. As with Rhoticity Model 1, there are main effects for self-priming and the position of the token. Presence of /r/ in a given speaker’s previous token leads to higher predicted log odds of /r/ in the current token. Word-internal and pre-pausal /r/ is more likely to be realised than non-prevocalic /r/ at a word boundary. Additionally, this model found higher likelihood of /r/ when the following phoneme is a fricative than when it is a plosive or a sonorant. In addition to these simple main effects, there was a three-way interaction

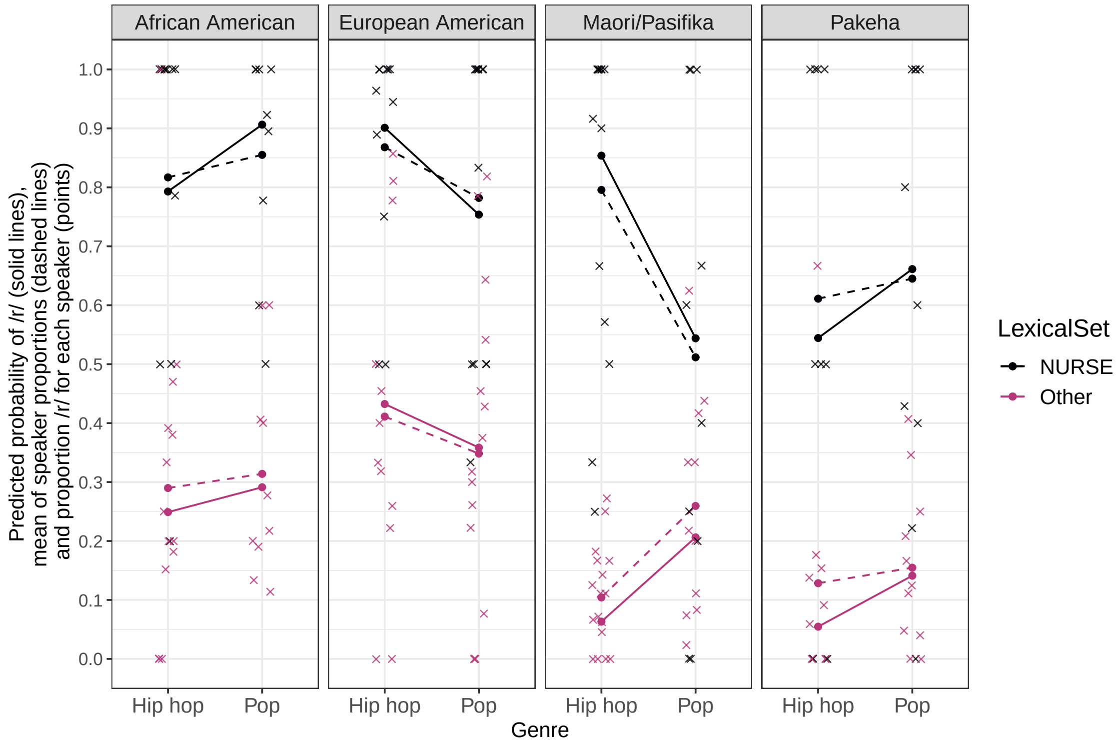


Figure 2.7: Rhoticity Model 2, using data from only male artists: Predictions from three-way interaction of genre, ethnicity and lexical set (solid lines) plotted with individual speakers' proportions of /r/-presence (points) and the mean of those speaker means (dashed lines). Each individual is represented by two points, one summarising their mean proportion of rhoticity in NURSE environments (black) and the other their rate of /r/ when preceded by other vowels (purple).

between genre, ethnicity and lexical set (NURSE vs. other). This interaction is plotted in Figure 2.7.

Looking at the interaction of genre, ethnicity and lexical set in the male data, we see that the strong separation between NURSE and non-NURSE words seen for NZ hip hop in Rhoticity Model 1 was actually driven to a large extent by the Māori/Pasifika rappers, while Pākehā rappers use less rhoticity overall than any other group. The inverse pattern is found for European American rappers, who are more rhotic than their pop counterparts, and also more rhotic than African American rappers, who in turn use *less* rhoticity than African American pop singers.

2.5.3.3 Rhoticity Model 3: Pop data

Rhoticity Model 3 excludes the hip hop data in order to look at gender in pop music. The model output is shown in Table 2.7. After a similar backward model fitting procedure, the final model had a maximum VIF of 9.3, for the interaction of gender with lexical set, and was fit as follows (the addition of slopes caused non-convergence):

Rhoticity Model 3: r $\sim$ gender * ethnicity * NURSE + lexicalFrequency + PrevR + position + (1 | speaker) + (1 | word)

Several of the main effects are now very familiar, with significantly less /r/ realised at word boundaries than within words or at the ends of utterances, and

Table 2.7: Rhoticity Model 3, based on pop data only.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.016	0.352	-0.046	0.963
gender=male	-0.529	0.437	-1.212	0.225
ethnicity=africanAmerican	-1.565	0.459	-3.411	0.001
ethnicity=pakeha	-0.209	0.438	-0.478	0.633
ethnicity=maoriPasifika	-1.739	0.468	-3.712	<0.001
lexicalSet=nurse	1.839	0.628	2.929	0.003
lexicalFrequency	-0.191	0.091	-2.098	0.036
previousToken=Rpresent	0.617	0.137	4.510	<0.001
previousToken=NA	0.095	0.250	0.382	0.702
position=wordBoundary	-0.489	0.197	-2.478	0.013
position=prePause	0.043	0.256	0.168	0.867
gender=male*ethnicity=africanAmerican	1.297	0.652	1.988	0.047
gender=male*ethnicity=pakeha	-0.960	0.656	-1.464	0.143
gender=male*ethnicity=maoriPasifika	0.968	0.657	1.474	0.140
gender=male*lexicalSet=nurse	-0.209	0.795	-0.263	0.793
ethnicity=africanAmerican*lexicalSet=nurse	1.901	0.777	2.448	0.014
ethnicity=pakeha*lexicalSet=nurse	-0.559	0.773	-0.723	0.470
ethnicity=maoriPasifika*lexicalSet=nurse	2.447	0.906	2.699	0.007
gender=male*ethnicity=africanAmerican*lexicalSet=nurse	-0.578	1.048	-0.552	0.581
gender=male*ethnicity=pakeha*lexicalSet=nurse	1.110	1.030	1.078	0.281
gender=male*ethnicity=maoriPasifika*lexicalSet=nurse	-2.698	1.147	-2.352	0.019

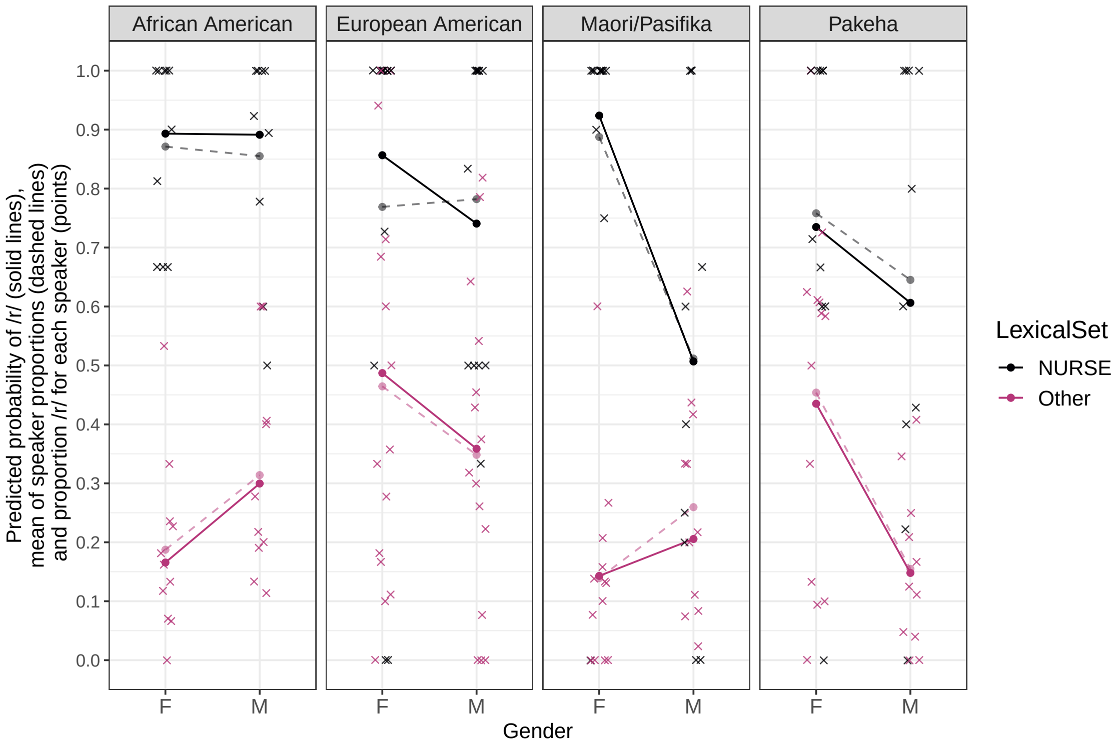


Figure 2.8: Rhoticity Model 3, using only the data for pop artists: Model predictions for three-way interaction of gender, ethnicity and lexical set (solid lines), plotted with proportion /r/ for individuals (points), and the mean of these speaker means in dashed lines. Each individual is represented by two points, one summarising their mean proportion of /r/-presence in NURSE environments (black) and the other their rate of /r/ when preceded by other vowels (purple).

a strong self-priming effect whereby the present token is predicted by the previous one. Additionally, there is an effect of lexical frequency (based on the LyricsPlanet website), where higher frequency words are less likely to be rhotic. This frequency effect approached significance in both of the previous models but was eventually dropped.

A three-way interaction (shown in Figure 2.8) was found in the pop data once again showing the way in which the ethnic groups behave differently for NURSE and non-NURSE tokens. Two trends in the NZ data are worthy of note. Firstly, female Pākehā pop singers use much higher rates of rhoticity than their male counterparts, and in non-NURSE environments, they use much more /r/ than any other sub-group of NZ performers, with the mean of speaker proportions being very similar to that of the European American female pop singers (45% for Pākehā and 47% for European American singers in non-NURSE environments, and 76% and 77%, respectively, in NURSE words). The second notable finding is that the female Māori/Pasifika artists strongly follow the Māori/Pasifika tendency to have high rates of rhotic NURSE and low rates of rhoticity elsewhere. This pattern is also evident for the female African American artists.

2.5.3.4 Rhoticity Models 4 and 5: USA data — the role of region

This section is different to those which precede it because it applies to variation amongst the US artists, rather than to differences between NZ and US artists. Non-rhoticity is a marker of both ethnicity and region in the USA, with lower rates of /r/ in African American speech, and in the South and New England dialect regions (see 2.3.1.1 for definitions of US regions in the corpus). There are enough tokens in the US dataset to consider whether these hip hop artists display their regional dialect through rhoticity, as we would expect given the fact that regional specificity forms a central theme in hip hop culture (Hess, 2009). The hypothesis relevant to this scenario is the Genre Hypothesis, regarding the prominence of genre in structuring variation. A modified version of the Genre Hypothesis for this analysis is as follows:

- Genre Hypothesis — In pop, there will be homogeneity. Artists from different regions and ethnicities will be indistinguishable. In hip hop, some artists will engage in own-accent rap, producing rates of rhoticity consistent with their place of origin or ethnicity. Specifically, European American rappers will be more rhotic than African American rappers and rappers from the west and mid-west will be more rhotic than rappers from the south and the east.

After a backward modelling procedure with the US data only, Rhoticity Model 4 was fit, containing the following terms:

Rhoticity Model 4: r $\sim$ NURSE + region*genre + boundary + (1 + boundary | speaker) + (1 | word)

The maximum VIF was 5.3, and the model output is shown in Table 2.8.

To check that this pattern was not related to the presence of female artists in pop and not hip hop, a model was run with only US male artists. The same region * genre interaction was significant (p=0.023) with the same overall pattern: no regional variation in pop, but a strong difference in rhoticity for hip hop depending on region of origin.

In the process of fitting Rhoticity Model 4, an interaction between ethnicity and lexical set was dropped in the late stages of the backward model selection process,

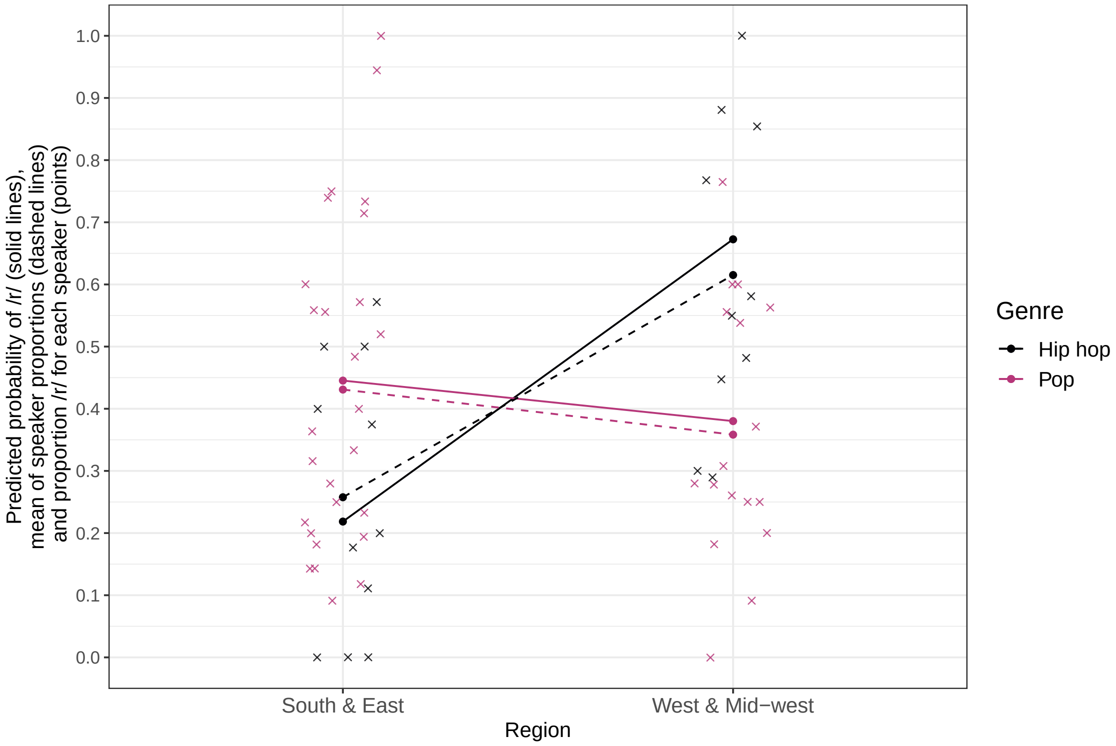


Figure 2.9: Rhoticity Model 4: Interaction of genre with region in USA data only.

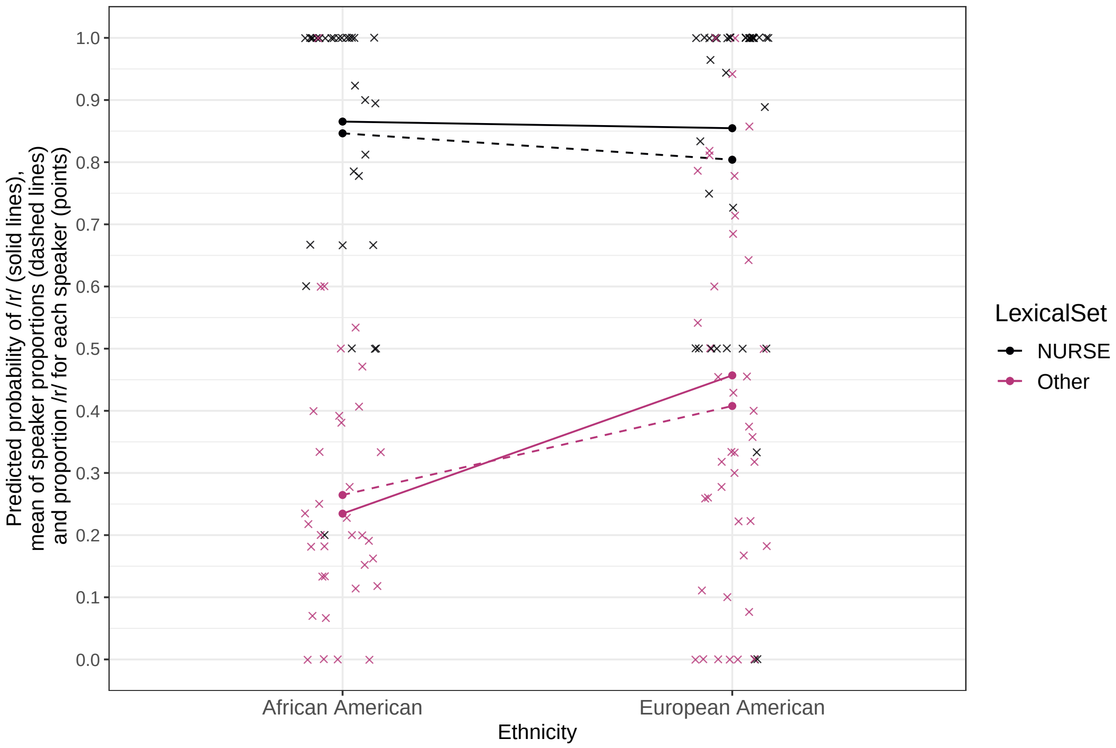


Figure 2.10: Rhoticity Model 5: Interaction of ethnicity with lexical set in USA data only.

Table 2.8: Rhoticity Model 4: based on USA data only.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.200	0.561	-2.140	0.032
lexicalSet=nurse	2.632	0.334	7.890	<0.001
region=west/midwest	1.995	0.703	2.836	0.005
genre=pop	1.055	0.623	1.694	0.090
position=wordBoundary	-1.047	0.218	-4.796	<0.001
position=prePause	-0.398	0.503	-0.791	0.429
region=west/midwest*genre=pop	-2.264	0.829	-2.732	0.006

because it failed to significantly improve fit (log-likelihood comparison p=0.119). This interaction is significant, however, in a model without the genre by region interaction.

Table 2.9: Rhoticity Model 5: based on USA data only.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.363	0.264	1.375	0.169
ethnicity=africanAmerican	-1.010	0.340	-2.970	0.003
lexicalSet=nurse	1.944	0.367	5.301	<0.001
position=wordBoundary	-0.972	0.221	-4.400	<0.001
position=prePausal	-0.105	0.481	-0.218	0.827
ethnicity=africanAmerican*lexicalSet=nurse	1.099	0.448	2.455	0.014

Rhoticity Model 5 shows this relationship. In this model, neither the addition of genre nor gender, nor any interactions with them, improve model fit. This final model is presented in Table 2.9, with the interaction plotted in Figure 2.10. This interaction shows that African American artists have a larger division between NURSE and non-NURSE environments, with less /r/ than European American artists in non-NURSE environments. This holds irrespective of their genre and gender. Rhoticity Model 5 was fit as follows, and had a maximum VIF of 1.7:

r ∼ ethnicity * lexicalSet + position + (1 + position | speaker) + (1 | word)

2.5.4 Non-prevocalic /r/: Discussion

Through five statistical models, analysing first the whole dataset and then three subsets of the data, several patterns of interest were discovered. Some of these relate directly to the hypotheses, and will be discussed first, followed by a discussion of various additional insights.

- Dominance Hypothesis — The large majority of NZ artists are substantially rhotic in their singing/rap, supporting the hypothesis that SPMSS remains the dominant norm in NZ popular music.
- Accuracy Hypothesis — There was a near significant interaction of word songiness and place of origin. This provides some evidence, albeit tentative, that these performers have phonetically detailed probabilistic knowledge about the

phonetics of popular music. Words heard more in song than speech have also been heard with high rates of non-prevocalic /r/. This fits with evidence that speech production is affected by such associations (Foulkes and Hay, 2015; Hay and Maclagan, 2012; Seyfarth, 2014).

Further support for this hypothesis comes from Pākehā NZ pop singers, who closely follow the rhoticity levels of their US counterparts, including the greater realisation of /r/ by female than male singers. Whether these groups were statistically indistinguishable, however, is less clear from the models presented, since the three-way interactions involve differences across multiple dimensions.

There are also, however, several pieces of counter-evidence to this hypothesis. Once again, Māori/Pasifika female pop aligns more with Māori/Pasifika hip hop, and in this case, with Māori/Pasifika speech styles more generally. This was not expected. There is a similar distinction between males and females within the US pop data, with females having lower rates of non-NURSE /r/, that is, a style more typical of AAE. Note however, that while US pop music is not homogeneous across gender and ethnicity, there appear to be parallels between NZ and US styles on both of these dimensions. European American and Pākehā female pop singers have more /r/ than their male counterparts. African American and Māori/Pasifika female pop singers have a larger distinction between NURSE and non-NURSE tokens than their male counterparts (a feature of both African American and Māori/Pasifika speech styles). While the situation for African American and Māori/Pasifika is complicated by overlap in the speech styles of these communities, the results for Pākehā singers suggest a more clear-cut adoption of the SPMSS by NZ artists, complete with gender-based norms.

- Genre Hypothesis — As just discussed, there is clear evidence against the hypothesis that artists' demographic backgrounds will be undetectable in pop music, particularly with respect to ethnicity. There is, however, some evidence for homogeneity in pop in the lack of any regional variation amongst the US pop artists.

  Regarding diversity in hip hop, there is plenty of evidence for own-accent rap in both the NZ and US contexts. For New Zealanders, we see a strong separation of NURSE and non-NURSE environments for Māori/Pasifika rappers, but not for Māori/Pasifika (male) pop singers. We see Pākehā rappers avoiding rhoticity in all environments, suggesting the conscious use of an own-accent style for both of these groups. The large gap between NURSE and non-NURSE words for Māori/Pasifika could also be interpreted as a case where they are adopting HHNL as represented in iconic hip hop (recall the findings of Gibson, 2010a). If the latter interpretation is adopted, then this difference between Māori/Pasifika and Pākehā rappers is in line with the findings for BATH. Since Māori/Pasifika and AAE styles overlap here, both of these interpretations can actually be considered to be true. It is notable, though, that these Māori/Pasifika rappers are definitely *not* adopting the patterns of rhoticity used by the commercial US rappers in this corpus. The results also support the hypothesised adoption of own-accent styles by US rappers, evidenced by adherence to regional dialect norms. There is thus strong support for the hypothesis that patterns of rhoticity in hip hop would reflect the artists' demographic backgrounds.

The one big surprise in the results was the high overall use of non-prevocalic /r/ by African American rappers. This result is completely different from the findings of Gibson (2010a) discussed above. This adoption of rhoticity in African American rap may reflect a change in progress in AAE itself — the songs analysed here are roughly 15 years newer than those studied in Gibson (2010a) — but the most obvious explanation for this difference is the massive commercialisation of hip hop that has occurred in the intervening years. Where there used to be a huge divide between hip hop and pop, both stylistically and in terms of audiences and subcultures, hip hop production styles and rap have now diffused through commercial music. Examinations of underground hip hop scenes (e.g. Taylor, 2011; Williams, 2017) may still find much lower rates of /r/ in non-NURSE environments.

- Salience Hypothesis — With the results of both BATH and rhoticity in hand, we can begin to test this hypothesis (which predicts that identity goals will be enacted when there is greater sociolinguistic or contextual salience), by looking at the relationship between the use of these variables for individual artists. Amongst the 15 artists who completely avoid the use of SPMSS BATH, the mean rate of rhoticity in non-NURSE environments (the only environments that can be treated as markers of SPMSS across all ethnicities) is 11.7%, with 6 artists avoiding non-prevocalic /r/ in these environments altogether. There are six artists who produced both variants of BATH. They all use some non-NURSE rhoticity, with a mean of 29.1%. The mean non-NURSE rhoticity amongst the 31 artists who consistently use SPMSS TRAP is 26.2%. The difference in rhoticity between those who consistently use PALM and those who consistently use TRAP is significant in a 2-tailed t-test comparing the speaker mean rhoticity rates of each group (p=0.046). Of those who always realised BATH as TRAP, there are two who avoid non-NURSE rhoticity, contrary to the Salience Hypothesis.[21] Overall, this comparison provides early support for the Salience Hypothesis, that NZ variants are used as an act of identity, which will involve the adoption of whichever markers of NZ-ness are salient to the artist in question. The realisation of BATH and rhoticity, which have acted as the dependent variables of the analysis thus far, now gain explanatory power, and can be used to predict results in other variables. Artists who use NZE variants for both BATH and /r/ are assumed to do so on purpose. If they use SPMSS on other variables, this may be a sign that they do not have awareness and/or control of those variables.

Māori/Pasifika hip hop stands out as having the pattern originally described for Pasifika NZ hip hop artists in Gibson (2005) and Gibson (2010a), a particularly large distinction between NURSE and non-NURSE environments. This was initially thought to be strongly influenced by US hip hop. In the results presented here, a

[21] These are Iva Lankum, for whom this is only based on 3 tokens and can therefore be dismissed. The other artist is Lukas, whose average is based on 26 tokens, and thus a clear exception to the proposed implicational scale introduced in Table 2.1. There is a strong case for the active avoidance of rhoticity in Lukas' performance, which makes his use of SPMSS BATH problematic for some of the central assumptions of this thesis.

large separation between NURSE and non-NURSE did not appear to be a feature of US hip hop so much as a feature of African American styles, irrespective of genre. This is counter to the Genre Hypothesis, that genre will structure variation more strongly than speaker characteristics. It appears that lack of rhoticity in non-NURSE words is a feature of African American speech that carries over into some popular music performances, at least for females. This is the first clear evidence that the adoption of a normative pop music style may not operate in the USA in the same way that it does in NZ. The strong separation between NURSE and non-NURSE in Māori/Pasifika hip hop may reflect the importance of this phonological rule in the projection of Māori/Pasifika identities in the NZ context. Disentangling local and global influences on this variable requires a deeper ethnographic investigation of the identities and stances being portrayed by these groups, in a range of styles, including but not limited to rap.

Before closing this discussion of non-prevocalic /r/, I return to the finding in both BATH and rhoticity that Pākehā rappers were predicted by the models to use an own-accent style more than Māori/Pasifika rappers. This result requires a consideration of authenticity. Alim (2002), shows that language is used by hip hop artists to construct an identity of 'street-consciousness', and that there is a greater presence of non-standard (including AAE) grammatical and phonological features in the rapping of hip hop artists than there is in their speech. In 'White Sunday Prelude' on Deceptikonz' debut album, NZ Samoan rapper Mareko challenges the assumption that it is 'fake' to use HHNL features in rap: 'you're probably not listening anyway because of my fake American accent'. He criticises those who would require him to match location with rapping style, perhaps in line with the perspective of Aboriginal rapper, Wire MC, quoted by Pennycook and Mitchell (2009, p. 37):

> As for the whole Aussie accent thing, man, I have a struggle going on with that one personally ... having white boys come up to me and saying 'you know, maybe you should rap a bit more Aussie'. And I'm like 'What?! Are you trying to colonize me again dude?! Stop it. Stop it'.

This quote brings up a very important, and easy to overlook, question around the national identity assumptions that are made around e.g. Australian English/NZE. These rappers do not necessarily associate with the mainstream identities and language styles of the countries in which they reside. They may, however, feel a strong sense of belonging within the HHN, and with African American culture more broadly. Zemke-White (2008, p. 109) quotes an Auckland rapper, Coco Solid, who stated this perspective plainly during an interview: 'We [Pacific peoples] feel we resemble them [African Americans] physically and socially, with parallels of oppression and colonisation. Pasifika communities want to represent these movements on the other side of the planet'. Māori and Pasifika hip hop artists have a licence to employ features of HHNL that Pākehā rappers do not (cf. Cutler, 2014; Sweetland, 2002).

McLeod (1999) outlined a range of dimensions by which hip hop artists can claim an authentic belonging to hip hop culture. *The Street, Black, Hard, Underground*, are in opposition to *The Suburbs, White, Soft, Commercial.* Cutler (2014) has explored questions around authentication for white rappers in great depth. Discussing her work, Pichler and Williams (2016, p. 562) state that 'some authenticate by highlighting closeness to African-American street culture, others authenticate by signaling honesty about their own (white, middle-class) background'. This may be exactly what is happening in the greater use of NZE by Pākehā rappers than Māori/Pasifika rappers. My discussion here may be guilty of the tendency for studies of authenticity in hip hop 'to be framed through notions of essential blackness, and critical interrogations of white hip hop legitimacy' (Harrison, 2008, p. 1783). Harrison encourages studies of authenticity in hip hop to emphasise 'hip hoppers who fall outside the black–white racial binary' (p. 1783). Through my methodological erasure of such individuals for the purposes of systematicity, I acknowledge that the research presented here has failed outright to take up that challenge.

2.5.4.1 Non-prevocalic /r/ in Pasifika varieties of NZE

The results for non-prevocalic /r/ supported a small body of evidence that rhotic NURSE is a feature of Pasifika youth styles of English in NZ (Marsden, 2017; Gibson, 2016), though very little is known about the origins, extent or indexicalities of this variant. It is possible that hip hop is partially responsible for the use of this feature in youth Pasifika Englishes in New Zealand, used by rappers and non-rappers alike in the day-to-day presentation of self, where hip hop affiliation is an important aspect of identity. We know that media can affect language use especially when the consumer is emotionally engaged with the mediated speech (Stuart-Smith et al., 2013). Hip hop listeners are also likely to be highly engaged, in a number of ways, with the rap to which they listen. The debate around the role of media in language change will be discussed in Section 5.4. Another, perhaps more likely, reason for rhotic NURSE in, for example, the NZ Samoan community, is the existence of strong family ties and ongoing migration between NZ, Samoa, American Samoa and the west coast of the USA. Pasifika English in NZ is the result of language contact and ongoing language shift. While NZE plays a primary role in the development of Pasifika Englishes, even in the Pacific Islands (Biewer, 2015), AmE — or perhaps more specifically, AAE, is also strongly present as an input variety. Young Pasifika people in NZ are likely to be meaningfully exposed to a range of English varieties, and draw on linguistic resources from multiple sources in their construction of identity through speech.

2.6 Linking /r/

Linking /r/ is a morpheme-final historically-present /r/ produced before a vowel. It is only genuinely 'linking' in non-rhotic varieties, where post-vocalic /r/ is not realised prior to consonants or pauses, but is realised in syllable onset, and intervocalically (where it is generally resyllabified to the onset of the following syllable). The lack of non-prevocalic /r/ allows for an alternation at morpheme boundaries where /r/ is variably present according to whether it is followed by a vowel. In rhotic varieties, 'linking' /r/ is functionally equivalent to other instances of post-vocalic /r/. In non-rhotic varieties, however, it has a specific function — to resolve vowel hiatus. In English, vowel hiatus is avoided, especially at important boundaries (e.g. at the start of prosodic units). Linking /r/ is thus both a vestige of rhoticity and a part of a larger hiatus-breaking system in English. In the latter sense, it is related to the determiners:

- 'a tree' vs. 'an apple'

- ‘[ðʌ] tree’ vs. ‘[ðij] apple’
- ‘[kaː] door’ vs. ‘[kaːr] alarm’

Unlike the two variables studied thus far, linking /r/ does not distinguish SPMSS from NZE. Similarly high rates of linking /r/ would be expected in both Pākehā NZE, and in European American English. There are ethnicity based differences in both countries, however: linking /r/ rates are lower in African American English, with glottal stop insertion [ʔ], and are also lower in Pasifika varieties of NZE, with vowel hiatus (for which I will use the shorthand [ø]) (see e.g. Kennedy, 2006; Bell and Gibson, 2008; Gibson and Bell, 2010; Gibson, 2016). This tolerance for vowel hiatus may be a marker of Pasifika NZEs more generally (see Pollitzer, 2009, for the frequent use of vowel hiatus in pronunciation of the determiners in Pasifika NZE). Linking /r/ also differs from BATH and rhoticity because it is strongly influenced by its prosodic environment. I provide context for both the social and prosodic factors influencing the realisation of linking /r/ in the next two sections, and will then describe how it relates to the four hypotheses applied to each of the variables analysed in this chapter.

2.6.1 Social factors affecting /r/-sandhi: Prior variationist studies of linking /r/

Foulkes (1997) presented the first systematic sociolinguistic study of /r/-sandhi in conversational speech in Newcastle upon Tyne in the UK. He found that linking /r/ showed structured heterogeneity with rates ranging from between 37% in young working class speakers to 79% in older middle class speakers.

In the NZ context, Hay and Sudbury (2005) describe the decline of rhoticity in early NZE, and the rise of intrusive /r/. They found that the presence of /r/-sandhi is favoured by a range of predictors: back vowels in either the preceding or following syllables; common collocations in the corpus; higher levels of speaker rhoticity; lower frequency following words; and male speakers. All of these independent variables will be examined in the modelling of linking /r/ below. Hay and Maclagan (2012) investigated the use of linking /r/ in 27 New Zealanders born between 1900 and 1935, and found compelling evidence for the storage of phonetic detail in the lexicon — words that occur more often before vowels also have both a higher rate of linking /r/ use, and also stronger /r/ constriction (as measured by F3 minima). These results are relevant to the overarching questions of this thesis, but for now, it is the specific rates of linking /r/ usage that provide a useful context for the corpus analysis. Taking ‘a very conservative approach regarding what “counts” as an /r/’ (p. 750), Hay and Maclagan found that /r/ was realised in 82% of the potential environments for linking /r/ across word boundaries (which is the environment to be studied in the corpus data below). This is similar to that found by Hay and Sudbury (2005) in the earlier period of NZE (83%). Once again, males were found to be more likely to produce linking /r/.

More recently, Hay et al. (2018) analysed potential occurrences of /r/-sandhi in the speech of 107 Canterbury Corpus speakers with birth dates ranging from 1900–1978. The analysis included the data reported by Hay and Maclagan (2012). 65% of tokens were realised with /r/ (n=2216, note that the dataset include 158 environments for potential intrusive /r/, which receive lower rates of /r/ production).

Analysis of year of birth showed a slight decrease in rates of linking /r/ over time, explaining the decrease in rates of /r/-presence across the two analyses. There was no further analysis of the cases where /r/ was absent, so rates of glottal stop insertion and unresolved vowel hiatus are unknown.[22]

The above figures are based largely on the speech of Pākehā New Zealanders. There are several small studies which suggest that linking /r/ is doing identity work in Māori and especially Pasifika communities. In a small qualitative analysis of a young Samoan male in a Pasifika Languages of Manukau Project interview, Bell and Gibson (2008) analysed 21 potential linking /r/ tokens, and found that 76% had no consonant (i.e. had vowel hiatus [ø]), 19% had [ʔ] and 5% had [r]. Similarly, in an analysis of 23 tokens produced by a young Samoan male in the QuakeBox Corpus, Gibson (2016) found 78% [ø], 4%=[ʔ] and 17%=[r]. Results in the *bro'Town* study presented in Gibson and Bell (2010) suggested (again, in a qualitative analysis of less than 20 tokens per speaker) that absence of linking /r/ was associated with the tough, streetwise character Valea (26% [r]) rather than his studious twin Vale (53% [r]), suggesting that avoidance of linking /r/ in the NZ setting is not just a marker of ethnicity, but also has semiotic value as a marker of toughness. Studying the speech of primary school students in many locations around NZ, Kennedy (2006) found the lowest rates of linking /r/ in the one school which had a majority of Pasifika students (in South Auckland). Schools with high proportions of Maori but not Pasifika students had intermediate levels of linking /r/. Even though these studies all include small numbers of speakers and tokens, they do show drastically lower rates of linking /r/ usage amongst Pasifika speakers than the rates reported above from the Pākehā speakers typical of the Canterbury Corpus.

2.6.2 Phonological factors affecting /r/-sandhi

The phonological literature also has insights about the rhythmic aspects of /r/-sandhi, and cross-linguistic patterns regarding the resolution of vowel hiatus. Uffmann (2007, p. 458) argues that 'Glottal stops are found epenthetically in onsets of initial or stressed syllables, that is, in prominent positions. They are, however, not found as hiatus breakers before an unstressed syllable ... Glides, on the other hand, are typical hiatus breakers, occurring intervocalically in a large number of languages.' Linking /r/ is functionally acting as a glide in cases where /j/ or /w/ cannot be used as hiatus-breakers (i.e. after non-high vowels). When /r/ is not realised, then, the choice between an unbroken vowel hiatus and glottal stop insertion relates to boundary strength, and the prominence of the second syllable in the pair, or perhaps the *relative* prominence of that syllable with respect to the syllable ending in /r/. With these considerations in mind, it is clear that in a discussion of linking /r/ in music, rhythm is especially important.

With this social and phonological background in mind, I present here the hypotheses for this variable, which are rather different for this variable than they were in the context of BATH and rhoticity.

- Dominance Hypothesis — Since all groups use linking /r/ to some extent, this hypothesis is not applicable here.

[22]While intrusive /r/ presents very interesting evidence about the mental representation of /r/-sandhi, the potential environments for intrusive /r/ are so infrequent that I decided to exclude it from analysis and discussion of the PoPS corpus data.

- Accuracy Hypothesis — NZ pop artists are expected to follow whatever the model for SPMSS happens to be, as will be measured through the performance of European American and African American pop singers. If NZ artists overshoot the rate of linking /r/, this would provide evidence that they are trying to display rhoticity, in contrast to the Accuracy Hypothesis, which predicts NZ singers to perform SPMSS without signs of intentional stylisation.

- Genre Hypothesis — In pop music, the patterns of usage which characterise ethnicity should be levelled. All pop singers will adopt the same pattern of linking /r/ realisations. In hip hop, there should be instances of own-accent rap, where realisation of linking /r/ lines up with the speech community underlying each of the demographic groups. This would mean high rates of [r] for both European American and Pākehā rappers, high rates of [ʔ] for African American artists, and high rates of [ø] for Māori/Pasifika rappers.

- Salience Hypothesis — Linking /r/ presents two quite distinct possibilities with respect to salience. It appears to function more as an indicator than a marker in non-rhotic communities, having a low level of salience (perhaps with Pasifika NZEs as an exception). However, we know that post-vocalic /r/ is itself a salient marker of SPMSS. It is therefore possible that singers will treat linking /r/ as simply another environment in which to enact their typical approach to post-vocalic /r/ more generally. This is reasonable given the lack of distinction in US dialects between rates of /r/ in non-prevocalic and intervocalic positions. Due to this uncertainty, linking /r/ was left off the salience hierarchy proposed in Table 2.1. Under the former analysis (low salience), it would be placed to the very right hand side. Alternatively, it could be grouped with rhoticity, at least for those singers who have not abstracted rules about the allophonic conditioning of these variables. A NZ artist would require a very sophisticated meta-linguistic handle on their own spoken phonology to both avoid non-prevocalic /r/ *and* produce high rates of linking /r/. The reproduction of this pattern, if accompanied by a full range of NZE-like vowels, would be evidence that the vocalist had successfully eschewed the influence of SPMSS on their performance style.

- Additional hypothesis — There will be a strong effect of prosodic context on the realisation of /r/-sandhi, with [ø] being used where the prosodic boundary is weakest (as the least fortis option), [r] being used to resolve hiatus at moderately strong boundaries, and [ʔ] being used to mark a strong boundary, such as the start of a new foot or prosodic unit, or perhaps a jump in pitch.

2.6.3 Linking /r/: Method

Previous analyses of linking /r/ have treated it as a binary variable: /r/ presence vs. absence. This masks important gradience, and as we shall see, these gradient realisations may be quite significant with respect to both prosodic constraints and social meaning. The present analysis thus uses a three-way auditory coding system, as described below.

2.6.3.1 Dependent variables

Linking /r/ tokens were coded in the same pass of the data as the analysis of rhoticity, described above in Section 2.5.2. Four different variants were coded:

ø = vowel hiatus (i.e. no consonant inserted)

ʔ = glottal stop inserted

r = canonical linking /r/

rʔ = [r] followed by short glottal stop not deemed to be a pause

The fourth category was a surprising one, which occurred on 39 tokens – these are cases where a linking /r/ was produced prior to a short glottal stop. That is: [VrʔV]. The longest of these pauses was 91ms. Note that any token where there was deemed to be a genuine pause was treated as part of the non-prevocalic /r/ dataset, and included there as a pre-pausal token. If the pause was very short, it was deemed to be a linking /r/ environment despite the glottal stop. For the purposes of the main analyses below, rʔ is treated as /r/.

2.6.3.2 Measurement of prosodic patterns

During the auditory analysis, it became intuitively clear that the relative stress of the syllables before and after the potential /r/ was important, as hypothesised. After completion of the main analysis, I went through all tokens again and coded the prominence of the musical beat for the syllable on either side of each potential token of linking /r/. Three levels of prominence (0, 1 or 2) were established, with 2 corresponding to strong beats (beats 1 and 3 of a 4|4 bar of music), 1 to minor beats (beats 2 and 4) and 0 to syllables falling on the eighth note divisions between these beats.[23]

To make a continuous scale, the prominence of the post-/r/ syllable was subtracted from that of the pre-/r/ syllable, creating a 5-point scale from -2 (weak–strong) to +2 (strong–weak). The results of this process are presented in Table 2.10, along with raw results of the auditory analysis of linking /r/. Once the data was reviewed with the stress patterns assigned in the above manner, it became clear that tokens with strong–strong or weak–weak patterns (that is, tokens with a stress pattern of 0 on the scale) were qualitatively different and should not be grouped together, and that it was trivial to determine which of the two syllables had greater stress. These 47 tokens were thus listened to again and coded at half time to allow the tie to be broken.[24] In the statistical models, the resulting stress patterns were collapsed into a two-way distinction between strong–weak, and weak–strong.

After establishing the codes for rhythmic prominence, each phrase was spoken aloud to check for the sentential and lexical stress patterns, as they would be likely to occur in speech. In all but 35 out of the 359 tokens, the rhythmic pattern as measured in musical terms aligned with the spoken stress pattern of the phrase.

[23]Whether to establish a half-time or double-time pulse is subjective, I went with my first musical intuition.

[24]Triplet patterns are common, especially in rap. Syllables falling on the second and third beats of a triplet were initially given code 0, but in order to break ties for tokens with equal prominence on either side of the /r/, the three beats were given descending prominence (2,1,0).

Cases of rhythm/stress mismatch often involved adaptations not just to sentential stress, but also lexical stress.[25] Since the rhythmic patterns largely overlap with stress, I refer to this variable as 'stress pattern' in the models and discussion below.

Table 2.10: Distribution of variants at potential linking /r/ environments, according to stress pattern. Lower numbers denote weak–strong patterns (e.g. *her eyes*) between the pair of syllables, higher numbers denote strong–weak patterns (e.g. *car and*).

Stress Pattern	ʔ	ø	r	Total
-2	10	19	1	30
-1	32	64	17	113
0	5	34	8	47
1	22	58	30	110
2	8	19	32	59
Total	77	194	88	359

In addition to this analysis of the rhythmic pattern, all tokens were coded with respect to whether or not there was a pitch change between syllables. A three way factor was coded, depending on whether the passage was rapped or sung (this code was also included in modelling and can differ from genre, though most rap occurs in hip hop, and the majority of singing occurs in pop). The token counts for each of these combinations is shown with a summary of the coding scheme below:

- Rap:

  0 = little or no pitch change (n=97)

  1 = part of a speech-like pitch contour (n=35)

  2 = abrupt pitch jump (n=6)

- Singing:

  0 = same note (n=94)

  1 = 1 or 2 semitones pitch change between syllables (n=77)

  2 = 3+ semitones pitch change (n=50)

2.6.3.3 Methods of statistical analysis

For statistical modelling, since multinomial and ordinal models can be difficult to interpret, it was decided that two binomial models would be fit: the first includes the full dataset of 359 tokens, predicting the log-odds of /r/ presence vs. absence (thus grouping [ʔ] and [ø] together, as most previous studies of /r/-sandhi have done); the second model includes only those tokens where /r/ was absent (n=165), and the dependent variable is the log-odds of insertion of [ʔ] vs. [ø] (the use of an unbroken vowel hiatus). Tokens of [rʔ] were treated as /r/ in the models (i.e. included as

[25] For example, Lorde's song 'Yellow Flicker Beat' includes the phrase *'they're sil'ver and 'gold*, changing the lexical stress of *silver* to fit the rhythmic demands of the song.

/r/ in Linking /r/ Model 1 and excluded from Linking /r/ Model 2), and will be discussed briefly at the end of the results section.

I summarise here the independent variables that were tested in the model fitting procedure for each of the models presented below. The following variables were significant in at least one of the final models:

- Gender-Genre — a 3-level factor: female pop; male pop; male hip hop
- Ethnicity — a 4-level factor: African American, European American, Māori/Pasifika , and Pākehā
- Rhoticity level — a continuous predictor ranging from 0–1: proportion of all tokens of non-prevocalic /r/ that were realised with /r/ for a given speaker
- Stress pattern — a binary factor distinguishing between strong–weak and weak–strong prosodic structures

The following variables were tested, and found not to be significant in either final model:

- Lexical:
  - frequency of words pre- and post- /r/: continuous
  - songy vs. speaky words
  - frequent collocation: binary factor (any two-word pair that occurred more than once in the dataset of 359 tokens was treated as frequent)
- Phonological/Rhythmic:
  - preceding and following vowel backness: binary factor
  - length of pre-r word: continuous
  - speaker rate (syllables/sec): continuous, based on all data for that speaker
- Musical:
  - note change — 3-level factor
  - whether beats and stress were mismatched or not

In both models, the relative prominence of the two syllables was highly significant. As shown in Table 2.10 above, strong–weak patterns favour vowel hiatus, while weak–strong patterns strongly favour glottal stop insertion. Tokens of linking /r/ are intermediate. This finding supports the additional hypothesis above, and will become important as we consider the social factors conditioning realisations of linking /r/. This also means that in Linking /r/ Model 1, predicting /r/ presence vs. absence, the two non-/r/ realisations are pulling in different directions, despite being grouped together.

2.6.4 Linking /r/ Results

To summarise the raw results, linking /r/ was present in 54% of the 359 tokens, [ʔ] occurred in 25% of tokens and the vowel hiatus was unbroken [ø] in 21% of tokens. After a backwards modelling procedure testing the variables listed above, the final model for /r/ presence vs. absence, shown in Table 2.11, had a maximum VIF of 1.3 and included the following terms:

Linking /r/ Model 1: r ∼ genre * ethnicity + speakerRhoticity + stressPattern + (1|Speaker) + (1|word)

2.6.4.1 Linking /r/ Model 1: Presence vs. absence of linking /r/ in all data.

Table 2.11: Linking /r/ Model 1: presence vs. absence of /r/ in all data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.433	1.043	-1.374	0.169
genre=pop	-0.166	1.017	-0.163	0.870
ethnicity=europeanAmerican	0.054	1.178	0.046	0.964
ethnicity=maoriPasifika	0.392	1.134	0.346	0.730
ethnicity=pakeha	1.470	1.271	1.157	0.247
speakerRhoticity	3.993	1.283	3.111	0.002
stressPattern=weakStrong	-0.957	0.381	-2.512	0.012
genre=pop:ethnicity=europeanAmerican	1.541	1.460	1.056	0.291
genre=pop:ethnicity=maoriPasifika	2.885	1.402	2.058	0.040
genre=pop:ethnicity=pakeha	1.035	1.506	0.687	0.492

There are two significant main effects that are not involved in an interaction. Speaker rhoticity level and the stress relationship between the pre- and post-/r/ syllables are both highly significant. The more rhotic an individual is, the more likely they are to use linking /r/. The strong–weak stress pattern favours /r/, where the prosodic boundary between the two words is minimal. The weak–strong pattern disfavours /r/, but as already discussed, this is largely because it *favours* glottal stops, which are grouped here with [ø].

There is a significant interaction between ethnicity and genre. African Americans have low rates of linking /r/ in both pop and hip hop. In the other three ethnicities, pop has higher rates of linking /r/ than hip hop. This distinction is especially strong amongst Māori/Pasifika performers, who have low rates of linking /r/ in hip hop, but much higher rates in pop. The distinction between pop and hip hop is much less evident for Pākehā vocalists, where hip hop has high rates of linking /r/, following NZE. The only term in this interaction which reaches significance is the comparison of Māori/Pasifika pop to the reference level for both ethnicity and genre: African American hip hop. Neither Māori/Pasifika nor Pākehā have significantly different rates of linking /r/ from European American pop. The interaction is plotted in Figure 2.11, along with individual speaker mean rates of linking /r/ (points) and the mean of those means (dashed lines). For NZ hip hop, artist names are also included in the figure, since they will be relevant to the discussion of the hypotheses below. Note that the mean proportion /r/ is very similar in pop

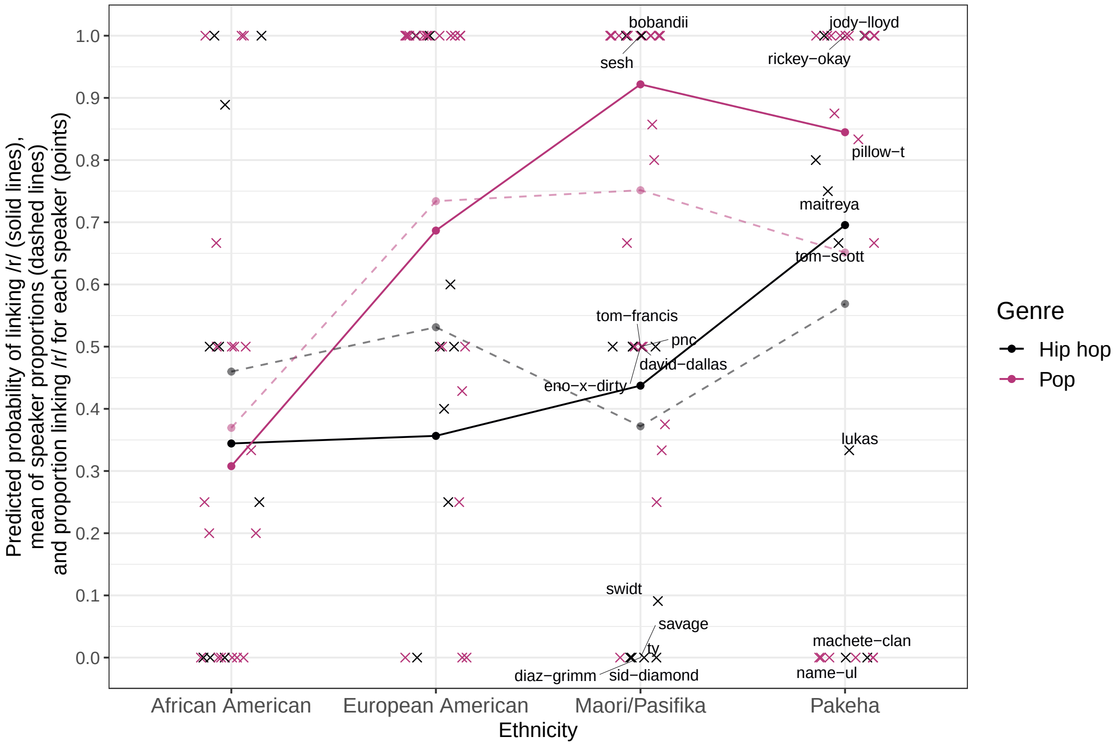


Figure 2.11: Linking /r/ Model 1: Interaction of ethnicity and genre, showing predicted presence (vs. absence) of /r/ along with raw data.

music for European American , Pākehā and Māori/Pasifika singers but much lower for African Americans. Though gender was not significant in models that included genre, female pop singers (68.5% mean of speaker means) used significantly more linking /r/ than male artists (52.8%) in a model that just included stress pattern, speaker rhoticity, gender and random intercepts for speaker and word (Log odds of linking /r/ for males vs. females: -1.21, p=0.03). The mean of speaker means for male pop singers was 58.1%, thus falling in between female singers and male rappers (47.3%).

The raw data suggests that Pākehā females use more linking /r/ than any other group, but gender was not significant in Linking /r/ Model 1. I briefly analyse this result here, since it has important implications as a potential site of stylisation of SPMSS by NZ artists, which would run counter to the foundational assumptions of the thesis. The mean rate of linking /r/ by Pākehā females was 91%, compared to 61.5% for European American artists. This difference in raw data is what drew my attention. However, once we look at the mean of speaker means, rather than allowing certain individuals to dominate the analysis, the differences collapses substantially, with 79% /r/ for Pākehā and 72% /r/ for European American singers. A simple mixed effects model was fit to the binary distinction between /r/ presence and absence for just the 58 tokens observed for these two groups. The model included just a main effect of gender and speaker rhoticity, with a random intercept for word. Models with an intercept for speaker did not converge, so in this statistical analysis, Pākehā female singers with a higher token count for linking /r/, and European American singers with a high token count for linking /r/ absence, contribute more to the model than they should. The model thus reflects the overall means of the results rather than the mean of participant means. With these caveats in mind, the model found that Pākehā female pop singers were significantly more likely to produce linking /r/ than their European American counterparts (intercept = 0.55; estimate for European American singers as compared to Pākehā singers: = -2.43, p=0.013; speaker rhoticity: estimate = 4.49, p=0.038).

2.6.4.2 Linking /r/ Model 2: [ʔ] vs. [ø] in non-/r/ tokens.

The next model was based on the subset of tokens where /r/ was absent, modelling the log-odds of glottal stop insertion [ʔ] as opposed to vowel hiatus [ø]. The output of the model is shown in Table 2.12, had a maximum VIF of 2.8, and the following terms: Linking /r/ Model 2: [ʔ] ∼ genderGenre + ethnicity + stressPattern + (1 | Speaker)

Table 2.12: Linking /r/ Model 2: [ʔ] vs. [ø] in non-/r/ tokens

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.824	0.937	0.880	0.379
gender/genre=maleHipHop	-2.563	1.116	-2.298	0.022
gender/genre=malePop	-3.175	1.257	-2.526	0.012
ethnicity=europeanAmerican	-2.520	1.182	-2.132	0.033
ethnicity=maoriPasifika	-0.778	0.943	-0.825	0.409
ethnicity=pakeha	-0.027	1.001	-0.027	0.979
stressPattern=weakStrong	3.801	1.022	3.718	<0.001

To summarise the findings of this model, for tokens without [r]: the weak–strong stress pattern favours [ʔ]; female pop singers use the most [ʔ], though recall that they also use more /r/ than any other group so this is based on a small subset of their tokens; log-odds of [ʔ] are highest for African American artists, followed closely by Pākehā performers. Māori/Pasifika had less use of [ʔ] again, while the European American group are the least likely to use [ʔ] when they avoid linking /r/.

This model only converged with the inclusion of random intercepts for either speaker or word, but not both. The pattern of significance and coefficients is similar in both models, with the exception of the Māori/Pasifika data, where the level of [ʔ] is near-significantly lower than it is for African American vocalists in the model with the word intercept only (estimate=-1.3968; p=0.063). Since it has been my standard practice throughout this chapter to ensure individuals with a higher token count do not influence the results too strongly, the Speaker intercept was used in the final model, making the difference between Māori/Pasifika and African American non-significant (p=0.41).

2.6.4.3 A fourth variant: [rʔ]

For the 194 [r] tokens: 19% were realised as [rʔ]. This occurred at a higher rate in pop (24%, n=138) than hip hop (7%, n=56). This distinction between genres was significant (p-value for genre = 0.02) in a simple binomial GLMER model, fitted as follows:

log-odds of rʔ (vs. /r/) ∼ genre + (1 | Speaker) + (1 | word).

There were no significant country or gender differences, but the raw data suggests a slight ethnicity trend, with greater use of [rʔ] by both both European American and Pākehā artists.

2.6.4.4 Māori vs. Pasifika rappers

In the analysis of this variable, the grouping of Māori and Pasifika artists is particularly problematic, since low rates of linking /r/ may be one of the features that distinguishes Pasifika NZEs from Māori English. A small number of artists could be clearly identified as *either* Māori or Pasifika, that is, not having both Māori *and* Pasifika ancestry. The raw data for these two groups was pooled: Māori had 63% [r] (n=16) while Pasifika had 39% [r] (n=18). Both groups used mainly vowel hiatus on the non-/r/ tokens, however. Obviously this is far too few tokens to draw any robust conclusions: but does support the prior studies which also had very small numbers of tokens.

2.6.5 Linking /r/: Discussion

To summarise the results: the weak–strong stress pattern favours [ʔ]. Boundary strength may be particularly important for sandhi phenomena, with different strengths of epenthesis applied to different strengths of boundary, with the consonants from strongest to weakest: [ʔ] > [r] > [ø]. More prominent syllables have a stronger preference for a consonantal onset, be it [r] or [ʔ]. The traditional analysis of /r/ presence vs. absence may actually be ill-advised given the conflicting boundary marking functions of glottals and vowel hiatus. Additionally, many independent variables that were found to be important predictors in previous studies were tested, and did not emerge as significant here, with the notable exception of the correlation between rhoticity and linking /r/. The lack of significance of other predictors is likely in part due to the small size of the dataset, but may also be related to differences between singing/rap and speech for this variable. This appears to be the case for gender. Males were found to use significantly more linking /r/ than females in all the NZ studies described above, but in song, there is a significant effect in the opposite direction.

The results are now considered with respect to the hypotheses:

- Dominance Hypothesis — Not tested.
- Accuracy Hypothesis — The finding that female Pākehā pop singers have significantly more linking /r/ than their European American counterparts, who are assumed to set the norms for SPMSS, constitutes evidence against the Accuracy Hypothesis, and thus also against the foundational assumption of this thesis that the production of SPMSS by NZ artists does not involve stylisation. It appears to be a case of quantitative overshoot, with female Pākehā pop singers producing more /r/ in this environment than either European American pop singers do, or than they themselves do in speech (at least according to the figures presented by Hay et al., 2018). Following the logic I have presented elsewhere in the context of this hypothesis, overshoot is a sign of stylisation, and stylisation involves intention and awareness. While not as extreme, this result is akin to the hyper-correction argued for by Trudgill (1983).

The relationship between linking /r/ and non-prevocalic /r/ will be discussed with reference to the Salience Hypothesis, below, but this finding for linking /r/ also calls into question female pop singers' unmotivated use of rhoticity, as it might imply a desire to *display* post-vocalic /r/, regardless of following environment.

There is another possible challenge to this hypothesis, at the other end of the USA–NZ identity spectrum, in the linking /r/ results when considered alongside the results for non-prevocalic /r/. Amongst the Pākehā rappers, there are two individuals who appear to have a highly sophisticated understanding of NZE phonology: Jody Lloyd and Maitreya both avoid non-prevocalic /r/ and produce linking /r/ near-categorically, the pattern seen in NZE. The rap of these artists challenges the Accuracy Hypothesis which expects to see signs of stylisation in own-accent performances — that is, these rappers are *too accurate to NZE* to support this hypothesis. The rap of these individuals, with respect to post-vocalic /r/, suggests that they have overcome the influence of their exposure to music, and found a way to fully transport their spoken sound structures into the domain of rap.[26]

- Genre Hypothesis — There is strong evidence against the hypothesis that pop music is homogeneous, masking the demographic background of a speaker, at least in US music. African American performers are highly consistent in realising a low rate of linking /r/, independent of genre. Some support for the hypothesis comes from the lack of significant differences amongst the other three ethnic groups in pop. Unlike the findings for rhoticity and BATH, Māori/Pasifika artists do not adopt an own-accent style in pop, using a high rate of linking /r/. It is possible, then, that the reports of low rates of linking /r/ in Māori/Pasifika communities are actually related more specifically to the subset of Māori/Pasifika speakers invested in hip hop culture.

  With respect to diversity in hip hop, in analysing rhoticity, we found European American rappers diverging from African American rappers and adopting patterns that reflected their region and ethnicity. In linking /r/ this is not the case. European American hip hop artists adopt a rate of linking /r/ that is similar to that produced by African American artists. Māori/Pasifika rappers have a similar rate of linking /r/, which at the outset might also seem like adoption of the African American style. However, by analysing not just presence and absence of /r/, but the distinction between epenthetic glottal stop and vowel hiatus, we see that Māori/Pasifika rappers are indeed adopting an own-accent style in their performances, using [ø] where African American rappers tend to use [ʔ]. There may also be a difference between Māori and Pasifika artists, with Pasifika artists using the lowest rates of linking /r/. We also find support for the hypothesis that hip hop would involve own-accent rap amongst Pākehā performers. As predicted, Pākehā rappers use more linking /r/ than Māori/Pasifika rappers.

[26] This interpretation follows the assumption that performance of NZE in song/rap requires awareness. An alternate interpretation here, in line with (Trudgill, 1983), is that these rappers are not trying to follow US norms, and that all the other artists in the corpus *are* trying to do so. Given the weight of evidence across the corpus as a whole, the former interpretation seems the most plausible to me, though these exceptions warrant further analysis.

- Salience Hypothesis — The results for linking /r/ are quite different from those for BATH and rhoticity. The variable is most strongly determined by the prosody of the phrase in which it occurs, perhaps relegating social meanings to a lower level of salience. Also, even though it makes sense for a non-rhotic linguist such as myself to separate rhoticity from sandhi phenomena, we cannot assume that such a division exists in the minds of language users from either dialect region. We can expect linking /r/ to simply be a subcategory of postvocalic-/r/ for rhotic speakers (Hay et al., 2018), and this is also possible for singers and rappers from other dialects. If there exists a motivation to 'sound American', then linking /r/ would provide a phono-opportunity (Coupland, 1985) to display post-vocalic /r/ in an environment where these singers also use the variable in their native dialect. This could even suggest an additive effect of their experience with singing and their own speech.

  Some of the results are difficult to interpret with respect to the Salience Hypothesis. In the NZ data, though, there are clear signs of own-accent rap, suggesting this variable may have greater social salience for NZ hip hop artists, potentially doing the social work of marking Pasifika speech styles. The strong effect of prosody on the realisation used at /r/-sandhi environments relates closely to contextual salience (prominence/informativitiy): if language users have well-established expectations that they will hear [r] or [ʔ] in weak–strong environments, then a single token of vowel hiatus in this environment will attract contextual salience. That contextual salience, once mapped repeatedly onto a social category such as 'Pasifika' (or 'South Auckland' or 'rapper' or 'tough guy'), will begin to accrue sociolinguistic salience, and move to an n+1th order of indexicality — from an indicator to a marker. Regarding the Salience Hypothesis, one thing is clear. There are some individuals who *do* have a handle on the difference between rhoticity and linking /r/. Two Pākehā rappers (Jody Lloyd and Maitreya, labelled in Figure 2.11) completely avoid non-prevocalic /r/ (and realise all instances of BATH with PALM), and consistently use linking /r/. This is the one place where the significant positive correlation between rate of rhoticity and linking /r/ breaks down, showing instead a strong inverse relationship. These are clearly cases of own-accent performance. It would be very tidy to say that these performances come from heightened awareness on the part of these two rappers. However, if they are fully consistent in their use of NZE, how are we now to prove that AmE phonetics/phonology has any sway over them at all? To answer these questions would require us to delve much more deeply into the phonetic detail of these individual artists, as well as having conversations with them about their awareness of their accent in rap. Whatever mechanisms have led to these performers reaching the bottom-right hand corner of the schematic in Table 2.1, they remain the exception to the rule. Most NZ artists are strongly influenced by the recorded music they have listened to.

In sum, the results for linking /r/ are complex, and it is difficult to say whether they support or refute the hypothesis that greater salience results in greater adherence to identity goals. There appear to be multiple social meanings associated with the variants used at /r/-sandhi environments, some of which are bundled up with the meanings of rhoticity, whilst others may be quite salient, clearly demarcating social groups.

The variants of linking /r/ can be viewed as existing in an indexical field (Eckert, 2008), with a wide range of interconnected linguistic and social meanings that cannot be understood with the kind of static approach to variation I have broadly adopted in this chapter. Different social and prosodic contexts will highlight different indexicalities for different individuals. The analysis above only scratches the surface of how these indexical relationships might be structured. In terms of social meanings, for example, [ʔ] is associated with both female *and* African American styles in music. This is an unusual pairing of indexicalities. In other variables, and in broader discourse, maleness and hip hop are linked, while femaleness and pop are linked. This may be a case of the same phonetic realisation having different social meanings depending on the speaker and context (Campbell-Kibler, 2012). Glottal insertion could signal both ‘clarity’ (prosodic boundary marking) for female pop singers, and ‘toughness’ for African American rappers.

2.7 LOT

While rhoticity and BATH were treated as involving relatively discrete boundaries between variants (presence vs. absence in the former, and a phonemic split in the latter), LOT is much more amenable to analysis of the variation as gradient, partly because the difference between dialects is on multiple dimensions. In NZE (and BritE) it is a short, rounded, mid back vowel. Being a back vowel with lip-rounding, it has lower F2 than the unrounded and (usually) fronter SPMSS variants. And being mid to open-mid in height, NZE LOT has lower F1 than SPMSS variants which are generally open. The dialectology of LOT in the USA is discussed in detail in Labov et al. (2006, Ch.2). In an acoustic analysis of spoken and sung LOT in recordings of Joe Elliott from the hard rock band Def Leppard, Konert-Panek (2018) found all spoken tokens to be rounded (mean F1=569Hz, mean F2=939Hz), and both rounded (mean F1=795Hz, mean F2=1293Hz) and unrounded (mean F1=1078Hz, mean F2=1674Hz) variants were found in singing. I report these values here since formant measures of sung vowels are very rare in the literature, and provide a comparison for the data presented below.

Before discussing the methods for measuring LOT, I summarise the hypotheses.

- Dominance Hypothesis — Instances of open, fronted, unrounded LOT will be prevalent in NZ singing.

- Accuracy Hypothesis — When SPMSS is adopted, it will be accurate. Since LOT will be measured acoustically, evidence for this hypothesis would come from the absence of a statistically significant difference between NZ and US artists for both F1 and F2. Relatedly, I hypothesise the absence of any cases of overshoot in the quality of SPMSS variants of LOT produced by NZ artists. Such cases would stand out from the distribution of US variables, by having an even higher F1 and F2 than that adopted by US artists. Cases of overshoot in the direction of NZE would provide evidence that it is NZE, not SPMSS, that is stylised.

- Genre Hypothesis — Homogeneity in pop and own-accent rap: In pop, there should be a SPMSS which is adopted by most artists, whilst in hip hop there should be more evidence of own-accent styles. In this case, there are no clear predictions around ethnicity or gender, but the distinction between NZE and SPMSS is strong.

- Salience Hypothesis — While still a member of the USA-5 (Simpson, 1999), LOT is considered to be the least salient variable considered thus far (with the exception of linking /r/, for which the level of salience is likely to by highly variable). Artists consistently using a NZE variant of LOT should be the same individuals who used NZE variants for BATH and non-prevocalic /r/. Artists with mixed pronunciations across variables, showing some desire to use NZE but not full commitment to an own-accent style, will be more likely to use NZE LOT in contextually-salient positions. Given the small dataset, examination of this latter part of the Salience Hypothesis will be qualitative.

2.7.1 LOT: Method

The PoPS corpus was searched for cases of LOT in stressed syllables (having no unstressed entries in Celex), for both male and female artists. Instances of LOT following /w/ or preceding a tautosyllabic lateral were excluded.[27] A csv file with a range of contextual information was downloaded along with audio and textgrids for each token. A Praat script was run to open each soundfile with its relevant textgrid along with a dialogue showing the target word. For each token, I entered an annotation code. Code 1 was entered for valid tokens, for which I placed the cursor in the best part of the LOT vowel for formant measurement.[28] This was roughly the middle of the vowel in cases where the whole vowel was correctly tracked. In cases where instrumentation interfered with formant tracking in part of the vowel, a point was selected that was deemed to be most representative of the vowel realisation, based on auditory analysis and inspection of spectrogram. Code 2 was reserved for cases where the word was unstressed, or was for some other reason mis-selected by the search, such as the lexical item *momma* which is deemed as having a STRUT vowel. Code 3 was used for the many instances where I was not confident of the accuracy of Praat's formant tracking. Code 4 was used to exclude all cases of falsetto or very high pitched singing, due to its effect of raising all formant frequencies. The formant values and time codes from which the measurement were taken were extracted by Praat and merged with the csv downloaded from LaBB-CAT in R, in preparation for statistical analysis.

The vowel measurements were not normalised. Having acoustic data for only a single vowel meant that 'vowel-extrinsic' methods of normalisation (such as, the most widely-used, Lobanov and Nearey methods) were out of the question (for a review of normalisation methods, see Adank et al., 2004). The absence of robust information about the upper formant structure of the vowels, due to the musical

[27] The main reason for excluding LOT after /w/ relates to the fact that SPMSS LOT is unrounded while NZE LOT is rounded. Beginning the vowel with the rounding gesture involved in producing /w/ would lead to problematic coarticulation.

[28] The data for males and females were analysed separately to allow for differing formant settings, which were as follows: Maximum formant=5000Hz for males and 5500Hz for females; Number of formants=5 for males and 6 for females.

accompaniment, also ruled out vowel-intrinsic methods of normalisation. The lack of vowel normalisation means that the male and female data need to be considered separately, though comparison across speakers, even within speaker-sex, still remains problematic.

To verify the usability of the formant measurements, given this lack of normalisation, I went through the dataset a second time and conducted an auditory analysis of lip rounding, using three categories (unrounded, somewhat rounded, rounded). This measure strongly predicted F2, providing at least some support for the validity of the acoustic measures used. In addition to this checking procedure, the values were compared to unnormalised values reported in the literature for both NZ and US speech. This revealed across a wide range of sources that the distributions of NZ and US LOT in speech are largely non-overlapping, with NZ LOT vowels having consistently lower F1 and F2 than the wide range of LOT vowel qualities attested in dialects of the USA (Bigham, 2010; Clopper and Pisoni, 2004; Peterson and Barney, 1952; Labov et al., 2006; Schneider and Kortmann, 2004; Gibson, 2010b; Hay et al., 2008).

2.7.2 LOT: Results

The means of participant means for each gender, country and genre group are shown in Table 2.13. The raw data showing speaker means are shown in Figure 2.12 for males, and Figure 2.13 for females. New Zealand and American performers are largely overlapping in their realisations of LOT, contrary to the comparison of unnormalised values reported in the literature for NZ and US speech. The values for US male performers are also similar to the sung tokens presented in Gibson (2010b), which were demonstrated to be significantly opener and fronter than spoken equivalents by the same three individuals.

Table 2.13: Raw data for LOT: Means of speaker means for each combination of gender, genre and place of origin.

Gender	Country	Genre	F1MeanOfMean	F2MeanOfMeans
Female	NZ	Pop	892	1685
Female	USA	Pop	959	1708
Male	NZ	Hip Hop	782	1438
Male	NZ	Pop	819	1516
Male	USA	Hip Hop	803	1478
Male	USA	Pop	832	1584

Four simple LMER models (two for F1 and two for F2) were fit to the data to test for the effects of country, ethnicity and genre. Two models included all the data, and included a main effect for gender, which along with speaker intercepts, goes some way towards normalising the data. These models could not, however, include genre. Since pop includes females and hip hop does not, such models would spread the variance caused by gender across the two collinear predictors, gender and genre. For this reason, two more models were fit with just the male data, to test effects of genre. All four models had random intercepts for speaker and word, with no slopes.

Gender was, of course, highly significant in both of the models that included the female data. In addition, country approached significance for F1 (output shown in

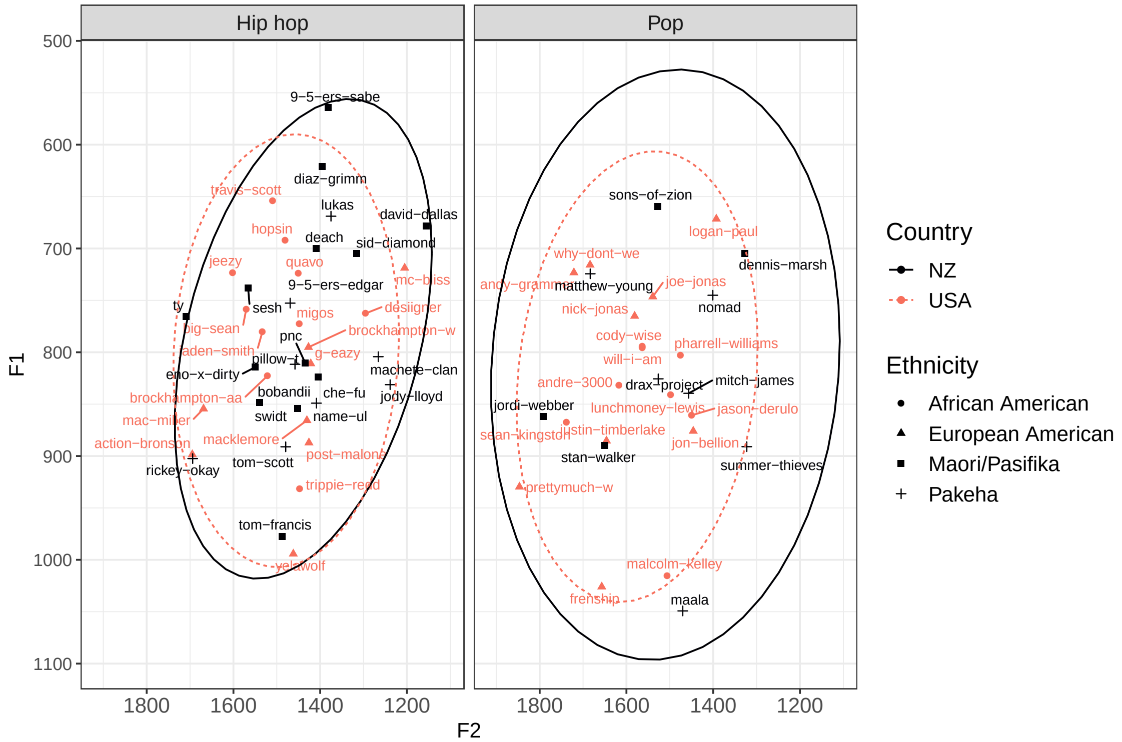


Figure 2.12: Speaker mean F1 and F2 in LOT for male hip hop and pop data. As a reference, note that the mean values for spoken NZE reported in Gibson (2010b) were F1=529Hz and F2=1057Hz.

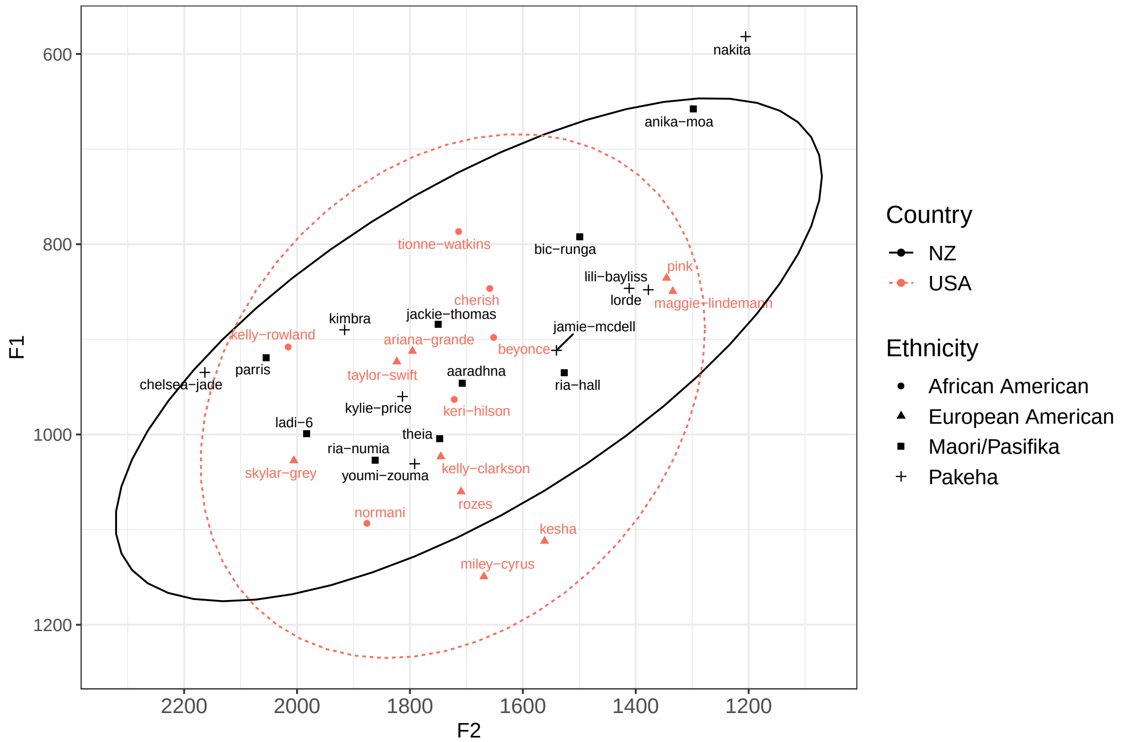


Figure 2.13: Speaker mean F1 and F2 in LOT for female pop data.

Table 2.14: LOT Model 1: F1, with main effects for country and gender, based on all data.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	919.691	21.808	117.922	42.173	<0.001
CountryUSA	34.393	19.586	82.688	1.756	0.083
GenderMale	-125.360	21.389	92.898	-5.861	<0.001

Table 2.14, with US artists having near-significantly more open vowels (Intercept = 920Hz; country=USA estimate=34Hz, p=0.08). A model with ethnicity did not fit as well, based on log-likelihood comparison, but suggested the difference between US and NZ artists was driven by European American artists having particularly high F1. In the model for F2, a model with ethnicity showed a trend for Pākehā artists to have lower F2 than African American artists (Intercept = 1693Hz; ethnicity=European American estimate = -21Hz, p=0.64; ethnicity=Māori/Pasifika estimate=-37Hz, p=0.4; ethnicity=Pākehā estimate=-85Hz, p=0.06), with European American and Māori/Pasifika artists having intermediate values. There was no trend for country.

In the two models fit with just the male data, there was no significant main effect of genre, ethnicity or country when modelling F1. In the F2 model, however, country was significant, with New Zealanders having lower F2 (output shown in Table 2.15).

Table 2.15: LOT Model 2: F2, with main effect for country, based on male data only.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1409.796	24.797	87.409	56.854	<0.001
CountryUSA	68.034	30.419	62.404	2.237	0.029

For the sake of establishing a general measure of how NZE-like each NZer's LOT realisations were, irrespective of gender, a model was fit to the F2 data with just gender as a main effect, and random intercepts for speaker and word (LOT Model 3, shown in 2.16). F2 was chosen since it was shown in LOT Model 2 to significantly represent the distinction between US and NZ artists. The speaker intercepts from this model will be used in the following section, when I make a final investigation of how the variables cluster together for different artists.

Table 2.16: LOT Model 3: F2, with main effect for gender, based on all data. The speaker intercepts from this model are used to determine the performers with the most NZE-like realisations of LOT.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1657.211	29.748	114.992	55.709	<0.001
GenderMale	-209.263	33.491	87.485	-6.248	<0.001

Looking at the speaker intercepts from LOT Model 3, it seemed that the significant difference between countries in terms of F2 might be driven by a few artists. To test the hypotheses laid out at the beginning of this section, we can see whether the significant effect of country on F2 goes away after identifying and removing those who appear to be consciously adopting NZE through the avoidance of salient features of SPMSS. Of course, it would be circular to remove people based on their intercepts for LOT, but by removing people on the grounds of their realisations of BATH and rhoticity, we have an objective measure of each artist's 'identity goal'. To define this subgroup of own-accent singers/rappers, I found all individuals who had both zero /r/ realisation in non-NURSE environments and zero instances of BATH as TRAP. Six individuals fit this pattern: one pop artist, Bic Runga (Māori/Chinese) and five hip hop artists: David Dallas (Samoan/European), Manu Walters of Eno X Dirty (Māori/Welsh), Edgar Mahon from the 9-5ers (Kiwi), Jody Lloyd (NZ European) and Maitreya (Pākeha)[29].

Only four of these six artists have tokens in the male LOT dataset that was shown to have a significant effect of country in LOT Model 2. By removing the 13 tokens contributed by these four artists and re-running the model on F2, the country difference becomes non-significant (p=.07 with own-accent singers excluded, n=254; p=.029 with them included, n=267). Given the tenuous nature of these models, it is necessary to show that this difference is not merely to do with decreasing the statistical power of the dataset. To ensure this difference was not due to reducing the sample size, 15 models were run with 254 observations randomly sampled from the full dataset of 267. Fourteen of these models had a p-value of less than 0.05, with a mean p-value of 0.031. Thus, the difference between NZ and US artists was robustly significant when the own-accent artists were included, and non-significant when they were removed.

2.7.3 LOT: Discussion

NZ artists had significantly lower F2 than US artists (amongst males), suggesting a more retracted and rounded variant, consistent with NZE. However, once objectively defined own-accent singers, such as David Dallas, are removed, the two groups become statistically indistinguishable by place of origin. This finding is not surprising upon listening to David Dallas. He uses a particularly NZE style LOT, with a very strong lip-rounding gesture on every occurrence, reflected in his low average F2. This is consistent with his total avoidance of non-prevocalic /r/ and his consistent realisation of BATH words with PALM. He adopts a NZE phonetic style across a range of variants, and does so intentionally, as reflected in one of his lyrics: 'all we do is do us, ain't from the States' (in the song 'Don't Rate That').

Before looking at the hypotheses again with respect to these results, I present a couple of other qualitative observations, beginning with Nakita's unexpectedly NZ-like LOT vowels, and then continuing the story of how contextual salience seems to play a role in the use of NZE features by Sons of Zion.

There was no statistical evidence of a distinction between NZ and US artists in the models with females included, but I briefly consider the raw data here. If we relax the criteria for own-accentedness used above to allow up to 10% rhoticity in non-NURSE tokens, there are two female artists singled out as own-accent performers: Bic Runga and Anika Moa. The raw data for female pop singers are shown in Figure 2.13. Bic Runga is on the NZ end of an otherwise homogeneous cloud encompassing

[29]Ethnicity labels based on either personal communication or quotes from the media, see Appendix A for details.

data from both US and NZ singers, the two tokens of LOT in Bic Runga's song are both rounded, with an F2 of about 1200Hz. Anika Moa shows clearer separation from the SPMSS norm.

The female singer with the lowest F1 and F2 values, however, is Nakita. She used SPMSS variants for both BATH and non-prevocalic /r/, so this token goes against my predictions. A qualitative analysis of the rest of the song confirms her use of a wide range of other SPMSS variants including monophthongal PRICE and retracting GOAT. Both THOUGHT and LOT however are strongly rounded in many, though not all tokens. She also has a notably centralised onset to FLEECE in the word 'deep' which reflects NZE pronunciations of this vowel. Nakita's performance provides plenty of evidence for the dominance of SPMSS, but makes me question the salience hierarchy proposed in Table 2.1. The next section will present a systematic synthesis of the results with respect to the proposed salience hierarchy and the Salience Hypothesis more generally.

- Dominance Hypothesis — The variant of LOT used almost categorically in NZ pop music is more open, unrounded and fronted than would be found in NZ speech, supporting the Dominance Hypothesis.

- Accuracy Hypothesis — The NZ artists are statistically indistinguishable from the US artists once a few objectively chosen individuals are removed from the dataset. This supports the hypothesis that SPMSS will be performed with a high degree of accuracy to the model. There are no obvious cases of overshoot in the direction of the US variant.

- Genre Hypothesis — No differences were found in the NZ data according to demographic characteristics, in support of the first aspect of the Genre Hypothesis, that pop will be largely homogeneous. While several of the most NZ-like tokens came from hip hop artists, there were also several such realisations from pop artists. This variable shows a less obvious distinction between pop and hip hop in terms of the construction of a NZ style.

- Salience Hypothesis — The point just mentioned provides some support for this hypothesis. If LOT is somewhat less of a marker of NZ distinctiveness than BATH and rhoticity, then we would expect a drop-off in the number of artists using the NZ variant. This appears to be the case, supporting the Salience Hypothesis. However, the use of NZE LOT by Nakita calls into question the proposed salience hierarchy. It is entirely possible that female pop artists are sensitive to different variables than male rappers. The danger of proposing a hierarchy is that it makes it appear fixed, when the relative salience of variables is no doubt dynamic, and as with all things sociolinguistic, will exhibit individual variation.

  Regarding contextual salience, as opposed to sociolinguistic salience, I present three examples providing anecdotal support for the Salience Hypothesis, that the goal to use NZE is better carried out by singers where there is more contextual salience — this in turn provides further evidence for the Accuracy Hypothesis, that the use of NZE involves intentional stylisation.

  The first example comes from Sons of Zion, who were discussed above with respect to BATH. The word *shocked* in the song 'Now' is raised and rounded

(F1=523Hz, F2=1410Hz) unlike the three tokens of *got*, which are open and unrounded (F1mean=687Hz, F2mean=1585Hz), and one token of *drop* in a prominent position at the end of a musical phrase (F1=713, F2=1472). The word occurs in a key hook that is repeated in the song, it is a low frequency, high informativy word, (though not at the end of a musical phrase). These factors mean that it may be subject to a higher level of conscious processing than, for example the high-frequency word *got*. As discussed earlier, Sons of Zion do seem to be in a position of wanting to occasionally perform a NZE phonetic style.

Another example occurs in Summer Thieves' song 'Coast Roads' where an instance of high-frequency *got* is unrounded, fronted and open (F1=1027; F2=1488), but a prominent, though not long, instance of the low frequency word *rock* in utterance final position is rounded, backed and raised (F1=755Hz; F2=1157Hz). A similar pattern is found in Jamie McDell's song 'Fly Honeys'. Two tokens of *got* are fronted and unrounded (F1 mean=940Hz, F2 mean=1635Hz) while the very infrequent proper noun 'John', at the end of a musical phrase, is rounded, backed and raised (F1=854Hz, F2=1351Hz). It should be emphasised that these are just anecdotes. It is hard to distinguish between the argument I am making here and one of more general processes of reduction (in the form of centralisation, and thus raising of F2) on high frequency words such as *got*. In order to make solid conclusions, contextual salience would need to be carefully defined and quantified, as would each individual's apparent desire to use NZE. Only then could we model whether their identity goals are more consistently met in positions that stand out from their surroundings. Anecdotally at least, these examples support the Salience Hypothesis. These singers all have a desire to use their 'own voice', but are only able to do so at sites of greater prominence. Without special attention and effort, their selection of a target for production reverts to the mean of their experience with that word in song, and SPMSS is produced as a result.

2.8 Testing the Salience Hypothesis

The analysis of rhoticity and LOT above both involved simple ways of using an artist's production in one variable to predict behaviour in another. In this section, I will do this more systematically, across all NZ artists, to test the Salience Hypothesis, which claims that artists will be able to better enact their identity goals on variables with greater salience. This hypothesis can be tested in part by looking at variation in contextual salience, though I have only addressed this with respect to a few anecdotes. To test the relative sociolinguistic salience of the variables, we can assess how systematic performers are in their adoption of NZE styles.

2.8.1 Addition of data for GOAT

BATH, non-prevocalic /r/ and LOT are all members of the USA-5 group studied by Trudgill (1983) and many others, specifically because they were deemed to be salient markers of the distinction between British and AmE dialects.[30] Since I am trying to establish *relative* salience, I include data from one further variable: GOAT. This may be acquiring marker status in NZ popular music, but my impression is that NZE variants are still very rare. The variable was highly salient for one of the three singers in Gibson (2010b), and the distinction between NZE and SPMSS is phonetically large (see below), making it a good candidate for marker status.

2.8.1.1 GOAT: Method and Results

A quick and simple auditory analysis was conducted for NZ artists only. The PoPS corpus was searched for instances of interconsonantal GOAT, excluding any tokens with a preceding or following glide, and any tokens preceding a lateral. A total of 533 tokens were found, and the csv file and audio were exported. Using a Praat script, I listened to all tokens whilst reading along in the spreadsheet for the context of each token. I was specifically listening for instances of a fronting trajectory, the trajectory used in NZE speech, and which occurs occasionally in NZ songs.

This fronting variable is prevalent in Australian hip hop (86% of the 548 potential tokens in hip hop analysed by O'Hanlon, 2006, compared with 9% of the 88 tokens in Australian pop in her sample of songs). Coddington (2004) also studied GOAT, and found the NZE variant to be nearly absent from eight of the twelve albums analysed. She did find that the one alternative independent band in the dataset, The Coolies, used a fronting variant in just over half of the tokens of GOAT analysed. A handful of tokens of fronting GOAT were also found for Dylan Storey in Gibson (2010b).

For the purposes of this analysis, I simplify my results to the broad distinction between those who used at least one token of NZE GOAT and those who did not. Out of 67 artists, just seven used the NZE variant. This binary distinction is in the spirit of those used above, defining a small subset of vocalists who defy the normative form in their use of NZE phonetics.

Returning to the idea that there is a strong phonetic distinction between the NZE and SPMSS variants of GOAT, I present a small acoustic analysis of two NZ rappers' realisations of the vowel in Figure 2.14. These measurements were taken at roughly points 0.2 and 0.8 of each token, once again making some allowances for cases where the formant tracking was affected by the instrumentation. Even though the NZE variant and the SPMSS variant occupy a similar F1/F2 space, they have trajectories heading in opposite directions with respect to F2. This distinction appears to have taken on sociolinguistic salience for at least some singers.

Before moving on to my methods for testing the salience hierarchy, I would like to point out here a second reason for adding this brief introduction to the GOAT vowel. It was the one remaining variable to be discussed in order to provide background information for the lexical decision task presented in Chapter 4. The stimuli for that experiment were designed to include variables where there is a large phonetic distinction between NZE and SPMSS. These are: BATH, rhoticity, LOT and GOAT. The analyses presented in this chapter, then, not only stand on their own, but also provide empirical ground truth about the kinds of phonetic realisations the general

[30] This distinction also applies to the comparison of NZE and AmE (with the exception of intervocalic /t/, not addressed here. NZE is 'clearly British-derived rather than American-derived' (Clyne, 1997, p. 297).

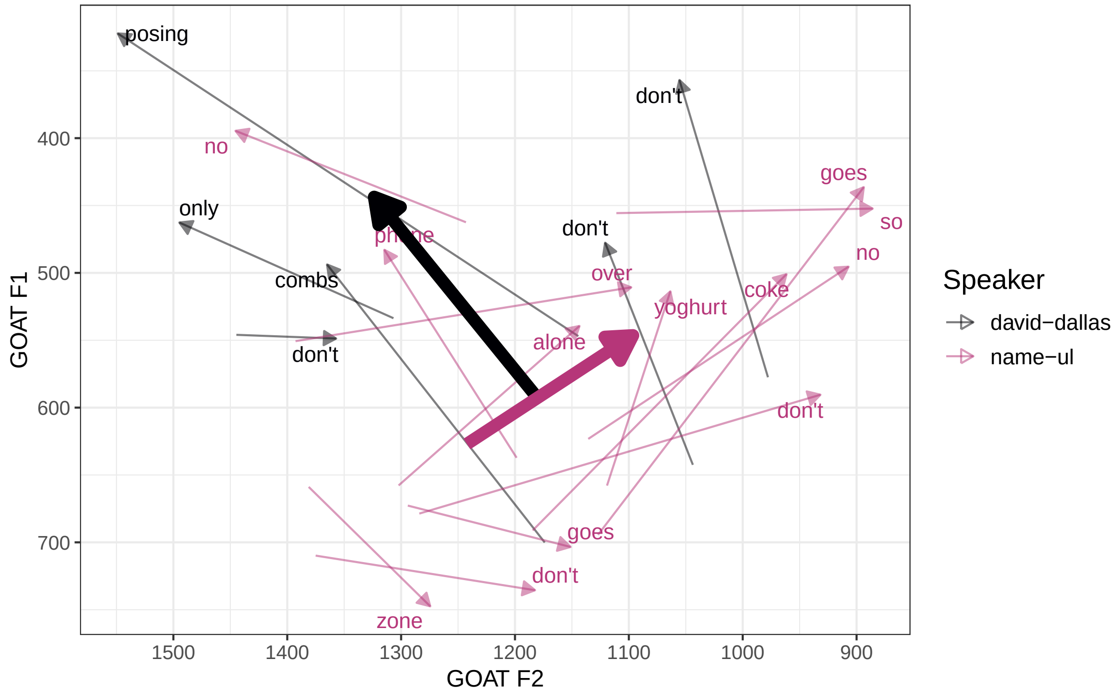


Figure 2.14: Examples of sung GOAT: David Dallas uses NZE (front-rising), and Name UL uses SPMSS/HHNL (back-rising).

NZ music listening public are exposed to when they encounter commercial popular music.

2.8.2 Defining 'strong use of NZE'

A boundary was decided upon for each variable, to distinguish NZ artists who did or did not adopt NZE in their singing style for that variable. These boundaries were based on my intuitions about what counted as 'strongly-NZE' for each variable, but could have been decided upon in a more systematic way, for example including exactly 10 artists (or some similar number) in the 'strong-NZE' category for each variable. Since each distinction is applied evenly to all artists, however, they remain a reasonably objective measure wherever the boundaries are set. The decisions made were as follows, defining the minimum requirement to count as 'strong use of NZE variant':

- NZE-BATH: all tokens of BATH realised with PALM
- Non-rhotic: less than 10% /r/ in non-NURSE environments[31]
- NZE-LOT: speaker random intercept of below -100, based on LOT Model 3, presented in Section 2.7
- NZE-GOAT: at least one token of fronting GOAT

[31]Since there was a larger dataset for rhoticity, and since there is a large degree of variability amongst almost everybody, a minimum of five tokens was required for an artist to be considered for this variable.

Thus, each of the NZ artists had a total of between zero and four cases of 'strong use of NZE'. There were 41 artists with zero, who are not further considered, though it should be kept in mind that the majority of NZ artists did not show strong use of NZE on any of the variables studied. The remaining 35 artists are shown in Table 2.17, sorted from least to most use of NZE. To determine the sorting order, '+' was assigned the value 1.1 and '—' was assigned -1. Missing data was assigned 0. The mean across the four variables was taken for each speaker. Tied values for those with three or four NZE variables were sorted by mean level of rhoticity in non-NURSE environments, with most rhoticity (i.e. most SPMSS-like) at the top.

2.8.3 Assessing the salience hierarchy

Recall the premise for this table: the variables to the right are less salient. Use of such variables would thus require a greater degree of awareness and control to go against the SPMSS. Such awareness and control is likely to come about for artists who have stronger identity goals around their projection of an 'authentic identity'. Thus, wherever there is a '+', all cells to the left of it should also have a '+', except where data is missing (as for example with Maitreya, for whom the criteria of the implicational scale are met).

Scanning the contents of Table 2.17, we can see very quickly that there is no strict implicational scale with respect to the salience of these four variables. Even amongst those with just one strongly-NZE variable, all four variables are represented.

In terms of raw numbers, there are six artists who only use a strong form of NZE BATH, that support the predictions. They project a NZ identity in the one place they know how to, in a variable which is highlighted even in the process of writing a song, through its contrastive phonemic status across the two 'native-like' dialects (spoken and sung) of the artist. There are three artists with just NZE BATH and rhoticity, supporting the predictions, along with a further four artists who are non-rhotic but did not have any instances of BATH. They also support the predictions.

Amongst the most NZ-accented singers, there are two artists with strongly-NZE use of BATH, rhoticity and LOT, and four artists who use fronting GOAT in a way consistent with the proposed salience hierarchy, three of whom are rappers whose NZ accents span many other variables. However, David Dallas and Sid Diamond also frequently use HHNL features. In sum, 19 out of the 35 artists shown here follow the predicted order of the salience hierarchy.

As for the exceptions, there are plenty of artists who are non-rhotic or produce rounded LOT whilst still using the American variant of the supposed shibboleth — BATH. The cut-off for these variables was perhaps too lenient. One way to further test this scale would be to make the cut-off for each variable increasingly stringent and see whether it is the exceptions that fall away first. A particularly interesting exception comes from Machete Clan. These rappers are most definitely engaged in stylisation, with multiple targets for their accent performances. They embrace salient US features, using some rhoticity as well as consistently realising BATH as TRAP in a rhyming sequence: 'smoking grass, smoking glass, hope you're paying attention that's the end of the class'. What is striking about their phonetic style, particularly given the overt use of US forms in these salient positions, is their performance of highly detailed stretches of NZE, including raised DRESS and TRAP, a particularly open and retracted nucleus to FACE vowels, and several tokens

Table 2.17: Salience Hierarchy: Results. NZE represented by +, non-NZ by —. Blank cells denote missing data. Artists sorted from least to most NZ-accented. Artists with no NZE variables (n=41) are not shown.

Speaker	Ethn.-Gender-Genre	NZE-BATH	Non-rhotic	NZE-LOT	NZE-GOAT
NZE in one variable					
ladi-6	mp-f-pop	+	—	—	—
pnc	mp-m-hh	+	—	—	—
diaz-grimm	mp-m-hh	—	+	—	—
lukas	pak-m-hh	—	+	—	—
matthew-young	pak-m-pop	—	+	—	—
name-ul	pak-m-hh	—	+	—	—
stan-walker	mp-m-pop	—	+	—	—
swidt	mp-m-hh	—	+	—	—
jamie-mcdell	pak-f-pop	—	—	+	—
lorde	pak-f-pop	—	—	+	—
nakita	pak-f-pop	—	—	+	—
summer-thieves	pak-m-pop	—	—	+	—
youmi-zouma	pak-f-pop	—	—	—	+
NZE in one variable (some data missing)					
dennis-marsh	mp-m-pop	+	—	—	
rickey-okay	pak-m-hh	+	—	—	
savage	mp-m-hh	+	—		—
9-5-ers-sabe	mp-m-hh	+	—	—	
balu-brigada	pak-m-pop	—	+		—
benny-tipene	mp-m-pop	—	+		—
chelsea-jade	pak-f-pop		+	—	—
deach	mp-m-hh		+	—	—
beau-monga	mp-m-pop		+		—
ginny-blackmore	pak-f-pop		+		—
NZE in two variables					
9-5-ers-edgar	pak-m-hh	+	+	—	—
eno-x-dirty	mp-m-hh	+	+	—	—
tom-scott	pak-m-hh	+	+	—	—
machete-clan	pak-m-hh	—	—	+	+
NZE in two variables (some data missing)					
lili-bayliss	pak-f-pop	+	—	+	
the-slacks	pak-m-pop		+		+
NZE in three variables					
anika-moa	mp-f-pop	+	+	+	—
bic-runga	mp-f-pop	+	+	+	—
NZE in three variables (some data missing)					
maitreya	pak-m-hh	+	+		+
NZE in four variables					
sid-diamond	mp-m-hh	+	+	+	+
jody-lloyd	pak-m-hh	+	+	+	+
david-dallas	mp-m-hh	+	+	+	+

+ strong use of NZE variant

— non-use of NZE variant

blank cells: missing data

'pak'=Pākehā 'mp'=Māori/Pasifika

of rounded raised and backed LOT. Their performance, about a multi-day drug and sex fuelled bender, is made to be confrontational and shocking, but also light-hearted and funny. Their mixing of NZ and US styles is also done in a totally non-conventional way, highlighting features of NZE not heard in any other rap songs in this corpus. The use of US variants alongside a 'fresh' NZ rap accent dispels any potential criticism that the accent is 'contrived', since the performers themselves put the contrived-ness centre stage. This is similar in many ways to the process of deauthentication through hyperbole described for the Flight of the Conchords in Gibson (2011). This could also be a response to the pervasive discourse that adoption of a US accent in song or rap is 'fake'. During discussions with singer Shaan Singh from Drax Project, he expressed pride in his adoption of SPMSS. The HHN has been put forward as a translocal speech community, people can be proud to signal their belonging to hip hop culture irrespective of their place of origin. There is no reason why this discourse could not also extend to pop music. Pop singers are also part of a translocal cultural practice, albeit one with more deeply engrained commercial aims.

The analyses here are exploratory. While I have used the term hypothesis, the predictions made were meant as both a way to structure my ideas, and as illustrative tools. The proposed hierarchy of relative sociolinguistic salience, along with my methods for testing it, have hopefully demonstrated a concept which could be tested with rigour in future, as well as providing multiple case studies which support that concept. The idea that some variables are more salient than others is not new. But the idea that it is difficult to access one's own speech variety in the context of singing or rapping is still counter-intuitive to many. I have demonstrated in this chapter that NZ artists perform SPMSS with great accuracy, and shown how the apparent intent to sing in a New Zealand accent tends to surface in places where there is greater attention to speech, in a peculiar inversion of Labov's original 1972 conception of style.[32]

2.8.4 Continuous representation of covariation patterns

The group of pop singers that used a rounded LOT alongside SPMSS variants of BATH and rhoticity led me to further investigate this pattern. Figures 2.15 and 2.16 show all NZ artists, and encapsulate the results of all four variables. What is striking is the group of artists who have high rates of rhoticity and use TRAP for BATH but have low F2 for LOT as represented by their intercepts from LOT Model 3.

Upon visual inspection of these graphs, one very salient pattern emerged. I have not tested this statistically, but I looked up the ages of the cluster of five Pākehā pop artists in the top right quadrant of Figure 2.15 who also realise BATH as TRAP. They range from 19–26 at the time of writing, with a mean of 22.2. I compared these ages to those of the five artists who fit my predicted salience hierarchy, falling in the bottom right quadrant of their respective panels of Figures 2.15 (Anika Moa and Bic Runga) and 2.16 (Sid Diamond, David Dallas and Jody Lloyd) whilst also consistently realising BATH as PALM (i.e. 0 for MeanTRAP). They

[32] I echo here the very insightful comment made in Bell (1984, p. 195): 'a shift back to local dialect - which Trudgill (1983) shows occurring in British punk music - [is] a peculiarly inverted initiative design'.

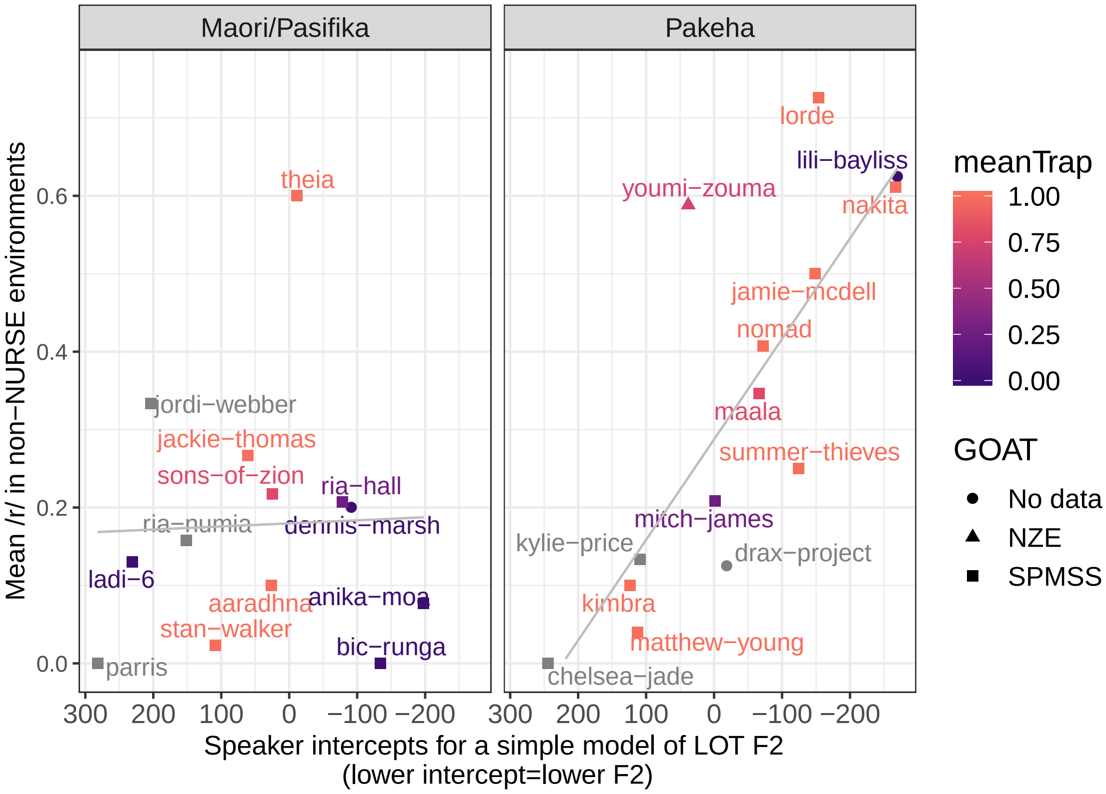


Figure 2.15: Covariation of BATH, rhoticity, LOT and GOAT in NZ pop, by ethnicity, with linear smooth lines. Greater values of 'meanTRAP' indicate greater use of the SPMSS variant of BATH. Points in grey signal that there was no data for BATH.

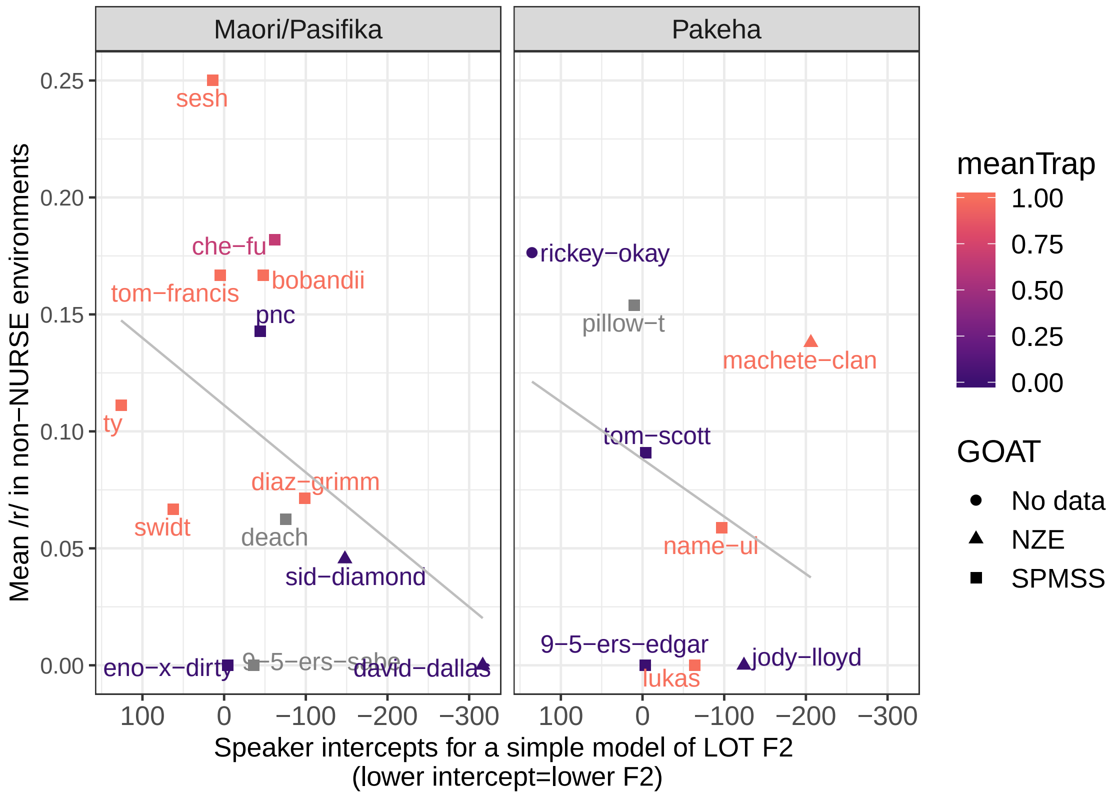


Figure 2.16: Covariation of BATH, rhoticity, LOT and GOAT in NZ hip hop, by ethnicity, with linear smooth lines. Greater values of 'meanTRAP' indicate greater use of the SPMSS variant of BATH. Points in grey signal that there was no data for BATH.

range from 33–43 with a mean of 38.8 years old. My salience hierarchy might date me. But it also shows that the intuitions I had about what constitutes an 'authentic NZ accent' in popular music have some validity for those in my cohort, whilst a younger generation has apparently come up with something new. I do not have an intuitive grasp on the phonetic style of that younger group, which means I also lack intuitions about what this constellation of variants might mean for these singers. It may well be a new way of signalling authenticity, though there is perhaps a more parsimonious explanation. These singers are all post-Ed Sheeran, whose super-star status and influence should not be unterestimated. They are contempories with a wave of British singer-songwriters that achieved prominence riding in his wake, using a variety of English drawing strongly on British as well as American phonology. While I do not know of any studies of Ed Sheeran's singing accent, and I have not analysed this in any detail, he certainly uses (rounded) British English variants of LOT and THOUGHT at least some of the time.

This presents a potential development in SPMSS, and a clear justification of why it is prudent to use a term like SPMSS over AmE to describe the sung variety of English under investigation here. A sociophonetics of popular music needs to detach itself from physical geography and follow the path of transcultural flows (Pennycook, 2007), wherever they may lead. What is particularly exciting about this is that it presents us with something that looks a bit like 'sound change'. What are we to make of singing accents if they display the kind of rich phonological processes seen in speech communities? In the absence of any significant face-to-face singing interactions, can processes of diffusion and transfer operate, and if so, how? I will return to the contentious issue of media influence on language change in Section 5.4.

2.9 General Discussion

From a corpus of 190 songs, a selection of variables were analysed with the objective of establishing the presence of SPMSS features in NZ singing and rap, the relative importance of genre and speaker characteristics in structuring variation, and the role of 'conscious awareness' in examples of own-accent singing or rap.

With the potential exception of linking /r/, the variables were presented in descending order from (according to my impressions) most to least salient. BATH is notable because the US artists present a unified model (in their lack of a TRAP–BATH split), which is easy for NZ listeners to grasp. It's obviously different from NZ speech, BATH words rhyme with a whole different set of words in singing than they do in speech for NZ artists. Rhyme is especially important in song and rap lyrics — it does not take subtle sociolinguistic awareness to grasp this dialectal difference. For these reasons, it is the variable in which we see New Zealanders behaving relatively consistently, and I argue that this is a result of being forced by the salience of the variable to make an identity/style based decision about pronunciation. Those who wish to present minimal role distances, that is, they wish to project an 'authentic self', are presented with an easy opportunity to do this by choosing to rhyme *can't* with *start* and not with *hand*.

As the most salient variable, realisations of BATH by the NZ artists were then used as a gauge of an artist's identity goals, and then also used to form hypotheses about other variables. If an artist chooses to realise BATH as PALM, then it was expected that they would also use NZ variants in other places where they have conscious access to a NZE vs. AmE phonetic distinction. There was some support for that hypothesis, though individual variation was prevalent.

Overall, this quantitative analysis of phonetic styles in popular music has forced me to soften my views about the extent to which SPMSS is performed homogenously and without effort in pop music. There are several examples of ethnic identity taking precedence over conformity to genre norms, and there is one example, albeit based on a small subset of the data (Pākehā female pop artists), of overshoot, signalling stylisation of SPMSS by NZ artists. However, the large majority of evidence does point to a native-like command of SPMSS by NZ singers, and adherence to HHNL (which bears many resemblances to other American Englishes) amongst rappers, with signs of own-accent rap more prevalent in the high-salience variables (BATH and rhoticity) than the low salience ones (LOT and GOAT). I consider now how these findings fit into the evidence for and against each of the hypotheses more specifically.

Singers juggle a range of conflicting motivations when deciding (consciously or not) between linguistic variables. There is a tension between convention and innovation, responsive and initiative, automatic and intentional. The results as a whole support the foundational assumption of this thesis, that an American-derived phonology takes the former position in each of these dichotomies. Let us step through each hypothesis in turn.

2.9.1 Evidence for and against the Dominance Hypothesis: Presence of SPMSS features

In the simplest terms, the hypothesis driving all data collection in this chapter, and the initial motivation for building the PoPS corpus was as follows:

- Dominance Hypothesis — SPMSS will be prevalent in the NZ performances, especially in pop music.

While the various nuances involved in the other hypotheses proved to be complex, this first hypothesis was conclusively supported: The Standard Popular Music Singing Style, derived from American Englishes, is not just present in NZ singing and rap, but highly dominant. The hypothesis was supported across all of the variables studied. The vast majority of artists, in both pop and hip hop, realise BATH as TRAP, have rhoticity where NZE does not, have an open unrounded LOT vowel, and a retracting, rounding GOAT vowel. The broad prediction that SPMSS variants would be more prevalent in pop than hip hop was also supported, though the difference between pop and hip hop is much less extreme than it is in Australia (cf. O'Hanlon, 2006).

To be fair, there was little doubt that the Dominance Hypothesis would be supported. By collecting a balanced sample of data from both NZ and US artists this chapter has been able to make more specific and interesting hypotheses.

2.9.2 Evidence for and against the Accuracy Hypothesis: Native-like production of SPMSS, stylisation of NZE

While still an indirect measure, the Accuracy Hypothesis in particular gets to the heart of the debate between the two main approaches to understanding non-US singers' adoption of SPMSS. An approach (e.g. Trudgill, 1983) that describes such adoption as effortful, predicts inaccuracy. For example, through the hallmarks of stylisation outlined by Bell and Gibson (2011a, p. 568): selectivity; mis-realisation; overshoot; and undershoot. The approach taken here, by contrast, describes SPMSS as a responsive style (Gibson and Bell, 2012), involving some degree of automaticity, and driven by the prominent role of context in language cognition.

The Accuracy Hypothesis predicted that NZ artists would have a native-like command of SPMSS, and thus be highly accurate in their performance of it: This is much more difficult to prove given the presence of some NZ-accent artists. There is generally solid support for a native-like grasp of the patterns of SPMSS amongst NZ pop singers. That is, there are exceptional uses of NZE, but once these are taken into consideration New Zealanders exhibit a great deal of overlap with American artists.

A possible exception is in the use of linking /r/, which showed particularly high rates of /r/ use by Pākehā female pop singers. This provides evidence against the hypothesis — my interpretation is that these singers value the usage of post-vocalic /r/. They realise it at accurate rates in non-prevocalic contexts, but in linking contexts, it seems that they have not absorbed the detailed patterns of the SPMSS model. This could be accounted for in one of two ways: the high rates of /r/ here may be the result of an additive process. Since linking /r/ occurs at high rates in the US model, there is a baseline level of /r/ usage in this environment represented in song. In non-prevocalic conditions, however, /r/ is highly represented in memories of song, but not of speech, and so the singers may have abstracted a rule that singing is more /r/-ful than speech. If such a rule were then applied to linking environments, there would be addition of the baseline /r/ levels and this song-related boost. Alternatively, it could be the addition of their own speech style and such a boost that leads to these high rates. However the result is construed, it is hard to explain without the inclusion of an intent to sound American that goes above and beyond a responsive imitation of context-relevant memories.

There are, on the other hand, instances where the use of NZE appears to be stylised, notably in the use of [ɑː] rather than [aː] for the PALM vowel used in NZ-like tokens of BATH.

In outlining the hypotheses at the start of this chapter, I described the kind of counter-evidence that could falsify each one. In the case of the Accuracy Hypothesis, it was stated that hyper-correction or overshoot of SPMSS would provide evidence of an intention to perform the US style. NZ female pop artists do not appear to overshoot the rates of /r/-fulness used by their American counterparts, and this extends to the phonological conditioning of previous environment. There is no significant difference between US and NZ female pop in this respect. This suggests a native-like internalisation of the phonology heard in pop music, rather than a 'putting on' of an accent of which they have limited knowledge. However, the linking /r/ and LOT data provide a counterpoint, and some potential counter-evidence to the foundational assumptions of this thesis. While linking /r/ rates in NZ speech are high (65% in Hay et al., 2018), the rates of linking /r/ used by the NZ female pop singers were higher than both those rates, and the rates produced by their American counterparts. This could thus be construed as a case of overshoot, and a wilful stylisation of SPMSS by pop singers.

The absence of any notable tokens of hyper-correct /r/ lend some support to this hypothesis, in line with the findings of O'Hanlon (2006), who specifically noted the absence of any hyper-correct /r/ in her analysis of 60 Australian songs.[33]

2.9.3 Evidence for and against the Genre Hypothesis: Homogeneity in pop and some diversity in hip hop

This hypothesis received the least support. There were findings in several variables where African American and also Māori/Pasifika artists used the style of their speech community in pop music. One possible reason for this relates to the difficulty in finding African American and Māori/Pasifika pop singers described in Section 2.3.1.3. It may be that my efforts to enforce independence of genre, gender and ethnicity failed. The songs may actually be on a spectrum from pop, through r&b, to hip hop, with the female Māori/Pasifika and African American pop artists falling in the middle of that spectrum. This might explain to some extent a stronger use of own-accent styles amongst these groups. Or it could be that projecting an 'authentic identity' is not just a value within hip hop, but actually a value within oppressed communities more generally. Throughout this discussion, I have been minimising the distinction between the commercial mainstream and the oppositional underground due to the fact that I have not structured that dimension into the PoPS corpus as it currently stands. Even if we were to capture that missing dichotomy, it might *still* be the case that ethnic identity is more powerful than the norms that govern pop music singing styles. This is something I did not foresee, and it is certainly worthy of further exploration.

Regarding diversity in hip hop, this hypothesis was strongly supported, even within the US artists, where regional rhoticity patterns surfaced in hip hop but not pop. Additionally, there was an interesting parallel regarding ethnicity between the BATH and rhoticity data: Pākehā rappers used more NZE than Māori/Pasifika rappers did.

In hip hop, HHNL is the dominant standard, and it is derived from AAE. The use of HHNL more by Māori/Pasifika than by Pākehā rappers provides insights into what it means to be an authentic rapper, and how this differs with ethnicity. Through 'parallels of oppression', including the differences in socio-economic status that have become entrenched through institutionalised racism, Māori/Pasifika artists are closer to the core of the HHN on multiple dimensions of hip hop authenticity (McLeod, 1999), and thus have greater licence to HHNL. Pākehā rappers can authenticate themselves through an honest representation of their background, which includes a local accent.

2.9.4 Evidence for and against the Salience Hypothesis: Identity goals are enacted where there is salience

It was hypothesised that, since NZE is an initiative style in popular music, it will be successfully enacted only when two criteria are met: firstly, a desire to use NZE (in the simplistic terms I have been using, to project an 'authentic identity'), and

[33]Though potential environments for hyper-correct /r/ insertion were not examined systematically, I am quite confident that I would have noticed such an occurrence when time-aligning the transcripts to the audio. That process involved a dedicated auditory scan of the entire corpus, during which I made notes about any interesting or unusual features.

secondly, awareness of the difference between NZE and SPMSS, along with the ability to manipulate the given variable (that is, to control their speech production, cf. Le Page's 1985 riders). This awareness, I argued, can be at a consistently greater or lesser level for different variables, though the analysis of the proposed salience hierarchy suggested that there is also plenty of individual variation in these broad levels of awareness of different variables. Awareness of any variable, whether or not it has higher order sociolinguistic salience, can also attract attention by jutting out from its surroundings in a large number of ways which revolve around contrast against the local context and surprisal. Cases where this *contextual salience* appeared to encourage the use of NZE were presented, providing some anecdotal support for the hypothesis.

Recall that the variables analysed here were chosen for exactly the opposite reason than they were initially selected by Trudgill (1983). He believed they would provide a good environment for catching instances where British singers use and mis-use AmE in song. In the present project, I chose these same variables because I believed they would be more likely to reveal instances where NZ singers could break away from the American accent that is so deeply engrained in the popular song context. The synthesis of results presented in the analysis of the salience hierarchy did support this fundamental premise. There is more use of NZE in the variables that saliently differentiate it from SPMSS than in those variables that are less salient. In the larger scheme of this thesis' research questions, the specific ordering of salience levels is a side-note. The key finding is this: *singing or rapping in a NZ accent takes effort and awareness.*

2.9.5 Improvements on previous work

Almost all previous studies of singing accents have left out US voices, relying, at best, on descriptions in the literature of dialects of American English speech. Too often, however, such studies have relied on essentialised notions of American English that come from a UK-centric perspective, paying attention only to features that mark AmE styles as distinct from Southern British English, or regional varieties from the North of England, New Zealand, Australia etc. This study is a step forward in this regard. By including US singing voices, we now have a reasonable estimate of what 'rhotic' actually means in the context of song, for example. With quantitative knowledge about American singing accents, we can now make much more well-founded judgements about the mechanisms involved in the singing styles of non-Americans.

O'Hanlon (2006) follows Trudgill (1983) in the use of Le Page's 1985 riders to linguistic modification to explain variation in the data. In her discussion, O'Hanlon (2006, p. 200) states that an Australian singer with 28% rhoticity was 'unable to fully rhoticize her singing' because of a lack of control over production of the variable (rider 2: ability to modify linguistic behaviour) and a lack of access to the reference group (rider 4). However, these conclusions were made without any knowledge of the actual rates of rhoticity in US pop music. In the PoPS corpus, 20 out of the 51 US pop artists' mean rhoticity rate is less than this. It would now seem unwarranted to conclude that the Australian singer's (apparently) low rate of rhoticity was due to inability to accurately emulate the US model. This is just one example of many in the literature that make claims about intentionality and (in)ability of non-US artists to accurately adopt an American accent in song, based on guesswork about the US model. Such claims can be reconsidered in light of the results for US artists presented in this chapter. Another question arises here: why is it that the US singers themselves have such low rates of rhoticity? It may relate to the importance of AAE in the formation of singing accent norms. This is something I will not be able to address, but it is a reminder that a rigorous comparison of the singing and speech of American artists would also strengthen our understanding of the sociophonetics of popular music.

2.9.6 From the singer to the listener

By looking at the results, not as isolated variables, but as connected semiotic resources, we see signs of both structure and idiosyncrasy. The fact that different singers can latch onto different variables to do identity work could relate to idiosyncratic cognitive histories — perhaps Nakita was exposed to *contextually* salient instances of NZE LOT in some favourite song. Perhaps repetition of that song led to strong acoustic memorisation (see Section 5.5) or perhaps there were just a few occurrences that caused surprisal, and were then encoded in memory with extra weight (Sumner et al., 2014; Hay et al., 2018). A few examples of such NZE tokens could set up feedback dynamics that lead to divergences of salience levels across individuals. Or perhaps, this pronunciation is nothing to do with dialect at all, but something to do with singing technique and hyper-articulation. This would be an interesting counter-example to the general trend towards sonority argued for by Andres Morrissey (2008).

Whatever the mechanisms were that led to the occurrence of the innovative NZE variants presented in this chapter, each one of them has been laid down into a recording that has the potential to affect a generation of music listeners (with a 'generation' of pop music listeners perhaps lasting just the length of one's teens and early twenties). I have argued that awareness is crucial for the performance of NZE in song — but not sufficient. There must also be a reason to do any linguistic styling that goes against the grain. What better reason is there than referencing a role model, a musical idol? That is most likely how this whole phenomenon got started, and in the period of music studied by Trudgill, that process of Nativisation (Schneider, 2007) was not yet complete: music listeners did not have a complete grasp of what a SPMSS should sound like, or how to make all of those sounds. But as the American music industry doubled-down on its control of the mainstream, the glimmers of Diversification documented in this chapter began to emerge, fueled by oppositional subcultures like punk and hip hop (though we may have also seen here signs of a move toward SPMSS in commercialised US hip hop). The global commercial pop music speech community may never really reach Phase 5 of its post-colonial development, since it shows no signs of cutting ties from its 'motherland' and forming a (global) national identity. The splintering off of subcultures, however, is constantly driving the development of more local and specific systems of musical, visual and sociophonetic differentiation (Irvine, 2001).[34]

[34]Meanwhile, non-Anglo music communities face similar processes at different scales. Singers in Montreal seem to be more adept at breaking with Standard French than New Zealanders are with SPMSS. It would be revealing to analyse the burgeoning field of music production in te reo Māori where there is a general rejection of code-switching, and a singing style that is phonetically similar

We are agents, but we also have limits on our working memory and other executive functions. The Standard Popular Music Singing Style is the lingua franca for a global community of popular music consumers and performers. They have a wide range of other native languages and native dialects, but in the context of song, they are all native singers/listeners of the American-influenced style that has shaped the majority of their musical experiences.

2.9.7 Future work

Understanding what a NZE style actually entails can come from a second pass of the data. Once we identify a singer like David Dallas, who is shown to carefully use NZE variants in LOT and BATH, we can use his performances to get a perspective on what he sees as authentic. Take, for example, his use of rhoticity in NURSE environments, and his use of vowel hiatus rather than glottal stops or linking /r/, even in weak–strong environments that strongly disfavour hiatus. With a qualitative approach, we can bootstrap our way towards understanding what a tough, conscious, South Auckland Samoan/European New Zealand *speech* style consists of, through David Dallas' stylisation of the characterological figure of himself in rap.

The perspective taken in this thesis, that a musical context activates memories and phonological knowledge associated with singing, might predict one rather striking phenomenon: the unmerging, in singing, of two phonemes that are merged in speech. The obvious example would be the near-merger of NEAR and SQUARE in NZE. A cursory examination of the NZ singers'/rappers' realisation of SQUARE shows that there are very few examples of the raised nucleus that would be found in speech. This 'unmerging' is of interest, as it shows knowledge of these categories for these singers. However, the merger is not uniformly complete across the NZ population, and we have no spoken data for the performers. An examination of the distance between NEAR and SQUARE in speech vs. singing would be more revealing than this examination of just the singing. Since NEAR and SQUARE are also always examples of potential post-vocalic /r/ (be it non-prevocalic or linking), we would expect a relationship between rhoticity and an open nucleus, and the concomitant increase in raised nuclei in non-rhotic tokens. Another approach would be to look at the realisation of SQUARE between NZ and US vocalists. An absence of significant difference is predicted. A different approach would consider general listeners, as will be done in the chapters which follow. Future work could test whether New Zealanders are better able to assign NEAR and SQUARE words to their lexical set if the words are presented in musical vs. non-musical contexts (cf. Hay et al., 2006b).

New Zealand English has a lexical merger of the words *woman* and *women*, where the plural form takes FOOT, not KIT, in the first syllable (Warren et al., 2017). Mergers could be a useful phenomenon for the study of how restricted the knowledge of sung phonology is to sung contexts. If, in speech, a person is unaware of the distinction between *woman* and *women*, it could be hypothesised that they will be able to make the distinction in singing. A qualitative analysis of the five instances of *women* produced by NZ vocalists (who all happen to be Māori/Pasifika male rappers) in the corpus was conducted. Diaz Grimm produces two instances of the plural with a clearly raised and fronted KIT vowel in both syllables. PNC

to the spoken variety. While these observations are merely impressions, I believe this could be a fruitful area for study.

uses KIT as well, though not as fronted and raised as those by Diaz Grimm. Sesh has one instance of plural *women*, and it is produced with KIT . The last token is by Savage, who has a lyric with FOOT in the first syllable of what is most likely the word *women*, though there is some contextual ambiguity around the plurality of the referent. Given the fact that Savage's accent is largely American-accented, that is, there are few signs of intent to produce NZE in his rap, this may be treated as evidence against the Accuracy Hypothesis. Given the conclusion by Warren et al. (2017) that the lexical merger may now be nearing completion for younger New Zealanders, we can be relatively confident that these artists would pronounce the plural *women* with a FOOT vowel in their speech. These few instances of 'unmerging' are thus of some interest, despite being anecdotal. The data presented in Warren et al. (2017) are largely based on Pākehā speakers, so an alternative interpretation of these results could be that Māori/Pasifika NZ Englishes have not adopted this merger.

While the inclusion of a wide range of US singers in this study was notable, this study still does not complete all desirable aspects of comparison, which would include spoken style. An ideal study would have a well balanced set of music recordings covering both US and non-US artists, matched with speech samples of those singers. This was the approach that Duncan (2017) took in his analysis of Australian country singer Keith Urban. If it could be scaled up to include a range of artists varying by genre, gender, ethnicity and commerciality, many of the questions than remain after the present analysis could be addressed.

2.9.8 Limitations

There are some limitations to this study which should be acknowledged here, including a detailed consideration of the structure of the corpus itself. Amongst smaller methodological issues was the reliance on auditory analysis for several variables. Recent findings by Hay et al. (2018) show just how subjective monitoring for the realisation of /r/-sandhi is, in non-linguist participants at least. Listeners 'reconstructed' an /r/ according to the likelihood of it occurring in that position. Another issue relates to the statistical analyses presented. While I made attempts in the sections above at disciplined and systematic model fitting procedures, the process was more exploratory and guided than is ideal. In the coming chapters I will go to great lengths to provide transparent statistical methods, that are pre-defined and preregistered. The experiments will have specific and testable hypotheses, unlike the high-level ones in this chapter.

Another limitation worthy of more careful attention is the imposition of researcher-defined macro-social structure onto the individuals studied. The most glaring limitation of the analyses presented in this chapter is the absence of female rappers, which came about as a result of my stringent song selection methods. The lack of female rappers means that I have presented results for female pop as if they have something specifically to do with gender, but this conclusion cannot be drawn. We can make conclusions about gender associations in the context of pop, vis-a-vis the male pop singers, but we cannot make conclusions about gender in popular music more generally. In the same vein, by not including female rappers, this analysis has not brought us much closer to understanding the sociophonetic portrayal of masculinity in hip hop, since all we can say conclusively is how male rappers differ from male pop singers.

The collection of songs controlled for country of origin, ethnicity, gender and genre. All artists came from music charts which were intended to provide parity across levels of commercial success. However, due to the differing sized music markets, this approach inadvertently hard-wired some other relevant discrepancies into the dataset. Primarily, the USA music industry is absolutely massive compared to the NZ music industry. Total retail revenue in 2016 for the NZ music industry was NZ$112m (US$70.6, PWC, 2018), and in the US was US$7.6b (IFPI, 2018). Even if we take the national populations into account, the revenue per capita is still higher for the US market, with about US$20.40 per capita in the USA and US$14.70 per capita revenue in NZ.

In the context of this ongoing American dominance, however, there is greater opportunity than ever before for non-USA artists to break through globally, exemplified by the fact that the South Korean commercial hip hop group BTS were the second highest selling artist globally in 2018, and the first ever non-Anglo artist to enter the IFPI global chart. This is referred to as the 'local becoming global phenomenon', with Martin Jessurun, of Warner Music quoted as saying 'globalisation leads to localisation and vice versa' (IFPI, 2018, p. 23).

2.9.8.1 A not-so-balanced corpus: Non-independence of ethnicity, gender and genre

By designing a 'balanced corpus', I actually had to fight very hard against the realities of the recorded music industry.

What was striking in the collection of the NZ chart data was the extent to which genre is predicted by ethnicity and gender. If you are a Māori/Pasifika female performer, you are most likely classified as r&b/soul, not pop. If you are a Māori/Pasifika male, you are likely classified as hip hop, though reggae/dub is also possible. Pākehā males are hard to find in hip hop but dominate in rock, while Pakeha females are stereotypical pop singers. This may reflect genuine sub-cultural preferences, or it may reflect the essentialising forces of the music industry on artists. Aaradhna strongly argued for the latter when she rejected the New Zealand Music Award for 'Best Urban/Hip Hop Album', stating in her non-acceptance speech: 'It feels like I've been placed in a category for brown people, that's what it feels like' and clarifying her genre of music: 'I'm a singer, I'm not a rapper. I'm not a hip-hop artist.' (NZHerald, 2016)

Related to this issue is the fact that the 'balanced corpus' created for this study is not actually a good reflection of listeners' experiences. Those experiences carry the very collinearities that the sampling method sought to overcome: more African American males in hip hop, African American females singing r&b and European American females singing pop, and so on. This has an impact on the conclusions of analyses. When looking, for example, at a variant such as [eɪ] in *can't*, the artificially balanced corpus is ideal for showing us that African American performers use this more than European American performers, irrespective of genre. However, if this were a random sample of songs, representative of an average listener's experience, then there would be more tokens of 'caint' in hip hop than pop, merely by virtue of there being a higher proportion of African American performers in that genre. A link between 'caint' and hip hop in the minds of listeners is thus still very likely, even though the approach taken here has hidden it. I do not raise this as a limitation. I still believe that within the quantitative empirical method, the most effective way to disentangle collinear semiotic dimensions is to fill all cells in a cross-tabulation of social categories. It is just important to remember that listeners operate in an environment where all of these dimensions are weighted by their rate of occurrence in-situ. Ultimately, what we are doing as sociophoneticians is exploring not just the links between phonetic forms and the social categories in a speaker-listener's mind, but also the ways in which such links overlap with one another and shift dynamically, being called forth with different likelihoods according to their relevance to the current cognitive scene.

Chapter 3

Phonetic Categorisation Task

In the context of an exemplar theory of language production, the findings presented in Chapter 2 reflect the representations held in memory by singers. The distinction between NZE and SPMSS for these singers can be viewed as a manifestation of non-overlapping clusters of memories for conversational language and language in song. The difference between these contexts is supported by differences in dialect, a range of other differences in the voice including steady state pitches in song as compared to pitch contours in speech, along with a range of other acoustic differences, notably musical instrumentation, which is expected to be bound together with phonetic detail in memory. Additionally, there are differences across other senses, such as the absence of a visible face in a substantial proportion of music listening experiences (seeing a singer's face actually leads to greater attention to and intelligibility of lyrics, Jesse and Massaro, 2010), or the presence of such a face in two dimensions on a screen, that can neither see nor respond to the viewer. The latter point is an example of the functional differences between language in general usage and language in song. Finally, recorded vocals are heard in the exact same form multiple times — for favourite songs, hundreds or even thousands of times — whereas memory for spoken words is dominated by a lack of invariance (Johnson, 1997) between occurrences.

Exemplar theory predicts that context activates sub-spaces of a large number of dimensions of memory. In the crudest sense, one such dimension would range from speech to singing. Through statistical learning, a musical context will activate for a listener a sub-cloud of speech sounds that are distinct from those activated in a conversational setting such as chatting with a friend or ordering a coffee. Crucially, such statistical learning is expected to happen not just for singers and musicians, but for all language users who encounter recorded music. These perceptual processes can be considered to be cognitively prior to the production results described in the preceding chapter. Just as in Pierrehumbert's 2001 original description of speech production, singing involves sampling from a relevant portion of phonetic memories. In the case of song, that subspace happens to involve a different dialect than the one activated for the purposes of speech. In sum, the distinction seen above for singers should not exist solely in the minds of music performers, but also in the minds of the general music-listening public.

Having now considered the phonetics of popular song, and having foregrounded the relationship between production and perception, this chapter begins to explore the representation of SPMSS in the minds of listeners. A phonetic categorisation task (PCT) will examine whether expectation of singing primes the speech perception system for SPMSS. The experiment uses a two-alternative forced choice task to establish the position of a perceptual phoneme boundary in musical and non-musical environments.

The next section provides the background for the phonetic categorisation task, introducing relevant literature and outlining the predictions of the experiment.

3.1 Background for Phonetic Categorisation Task

In this experiment, respondents hear resynthesised words that fall between *bed* and *bad*, and choose which of the two words they heard. While the literature reviewed in Chapter 1 presented a range of research exploring differences and similarities between language and music cognition, this section explores in more detail the studies which laid the foundation for the present experiment. I will first discuss early studies that used synthesised continua between speech sounds, and key studies in the sociophonetics literature using this methodology. The two studies by Drager (2006, 2011) upon which the present experiment directly builds will then be reviewed, followed by a more recent study utilising the same methodology (Hay et al., 2017). The section concludes with a summary of the methods and results of a previous version of the present task that I ran as part of my Masters research (Gibson, 2010b). Several problems with that earlier version of the experiment will be outlined, motivating the present replication.

3.1.1 Using synthesised continua to explore perception

Many studies have shown that speech perception is affected by listeners' expectations about the identity of the speaker. In a pioneering study, Ladefoged and Broadbent (1957) found that by altering the formant frequencies of an introductory sentence, the perception of the following word could also be altered. Further work at this time involved the development of continua of sounds synthetically manipulated to range between different phonemes. Liberman et al. (1957) synthesised a continuum spanning the stop consonants /b, d, g/, by manipulating the transition of the first two formants from the stop to the vowel. In a phonetic categorisation task, they showed that participants have a sharp perceptual boundary between consonants, hearing either one phoneme or another, with strong agreement between participants about where the boundaries are. That is, there is categorical phoneme perception. For vowel continua, however, Fry et al. (1962) showed that listeners have much more fluidity and uncertainty about where the boundary between categories lies. They synthesised a continuum spanning KIT, DRESS and TRAP, and found much greater variability in responses. They drew insightful conclusions about the different functions of consonants and vowels, with the latter being especially suited to conveying dialectal information and idiosyncracies of individuals' voice quality (p. 173). They suggest that in the kind of gradient perception used for vowels, context will be crucial to assigning sounds to their phoneme, whereas the difference between places of articulation in stops will be more robust across various contexts: 'listeners rapidly form an appropriate reference frame against which they judge the quality of and identify the sounds which occur ... the particular phonemic category selected is dependent on context, that is more specifically on the vowel reference frame which is operative for the listener at the time of reception' (Fry et al., 1962, p. 174–175). They also note that the *local* context effect is one of contrast: 'a given vowel sound will appear more open when preceded by a sound closer than itself' (Fry et al., 1962, p. 180). These points are highly relevant to the current study, and have not always been taken into account in modelling the variation found in more recent experiments using (re)synthesised continua. I will return to this issue in the analysis of results.

The insights from these earliest studies continue to be relevant, and the methodology in which participants categorise stimuli that vary along a synthesised continuum has had ongoing utility in psycholinguistics over the intervening decades. It has become a particularly useful tool in research adopting exemplar theory approaches to the study of speech perception. A series of studies in the late 1990s using two-alternative forced choice tasks for stimuli ranging between /s/ and /ʃ/ (Strand, 1999), and between FOOT and STRUT (Johnson et al., 1999), found that expectations about speaker identity (specifically, gender) affect the position of the boundary between the phonemes. Interestingly, this difference could be evoked by asking the participants to merely imagine they were listening to a female or a male. These effects were strongest in the parts of the experiment when the participants were most likely to be following that instruction, at the start of the experiment and at the end, prior to a test. Similarly, Rakerd and Plichta (2003) (cited in Clopper et al., 2010) found participants to shift their phoneme boundary between /ɑ/ and /ae/ according to the dialect of a preceding carrier phrase. This phenomenon has recently been demonstrated in neuronal populations in auditory cortex, through intracranial electrocorticography (Sjerps et al., 2019). Shifting the F1 of a carrier phrase creates a local contrast effect on the placement of listeners' phoneme boundary in a continuum of high back vowels, both in behavioural and neuronal responses.

D'Onofrio (2018) has shown that such phoneme boundary shifts can be elicited not just by static macro-social categories but by more nuanced social persona labels. Those primed with 'Valley Girl' are more likely to perceive a backed vowel on the TRAP–LOT continuum as TRAP than those primed with 'Chicago Bears Fan'. These perceptions reflect the speech patterns of the associated characterological figures (Agha, 2005).

Another series of studies using synthesised continua and binary forced-choice tasks was spawned by Ganong's 1980 seminal experiments. In these experiments, participants wrote or typed whether they heard /b/ or /p/, /d/ or /t/, and /g/ or /k/ as the initial sound in words which had VOT manipulated such that the voiced-voiceless series of stimuli formed a word–nonword pair (e.g. *dash–tash*) or a nonword–word pair (e.g. *dask–task*). The critical question in this paper was not *whether* a bias towards perceiving words would exist, but whether such bias would be stronger at the phoneme boundary than at the endpoints of the continuum. The results showed a strong bias towards hearing the initial phonemes so as to make words rather than nonwords and that this bias occurred on ambiguous stimuli more than unambiguous stimuli. One interesting aspect of the analysis was that participants were carefully grouped such that they each were exposed to a different step of the continuum on their first encounter with a given word pair. In this way, Ganong tested whether the bias existed even prior to the establishment of expectations around the word pairs and the extent of the continua. This condition assured that the lexical bias is 'truly perceptual' (p. 116) and not related to learning about the set of stimuli. The result was significant even for that small subset of the data. I follow this technique in my design and analysis of the PCT.

In one follow up study of particular relevance to the present experiment, Connine et al. (1993) found that this effect is also sensitive to word frequency for pairs of words — responses to ambiguous stimuli will be biased towards higher frequency words. For example, by incrementally adjusting the voice onset time of the initial plosive, a continuum of stimuli ranging from *best* to *pest* was created. Whilst the two ends of the continuum attract uniform responses, participants are more likely to hear the ambiguous intermediate tokens as the higher frequency word, *best*. Borsky et al. (1998) showed how semantic contextual congruence affects phonetic categorisation, with a bias towards hearing *goat* in the phrase 'milk a coat' (using a continuum from *coat* to *goat*). Regarding the question of whether such effects come from a warping of perceptual space or decision processes, they conclude that 'potentially ambiguous categorizations may be subject to additional evaluation in which a context-congruent response is both preferred and available earlier' (p. 2670).

A phoneme identification task by Pinnow and Connine (2014) examined the effect of recent experience on categorisation of phonemes, with one group of participants being exposed to high rates of schwa deletion. In an experiment building on the seminal study by Ganong (1980), word to nonword (e.g. *p(o)lice–b(o)lice*) and nonword to nonword (e.g. *plove–blove*) pairs of stimuli were created, by synthesising continua between voiced and voiceless initial stops. In deciding the initial phoneme of each word, participants were biased towards real words (with schwa deleted), the classic Ganong effect, but this effect was especially strong for words they had recently heard in the schwa-deleted form. Importantly, this knowledge also transferred to novel items, showing not only that the listeners updated their expectations based on exposure within the experiment but also abstracted a rule which could be extended by analogy to other forms. This balance between the impact of phonetic detail in memory and the formation of abstractions is important to our understanding of speech perception and representation, and will be further discussed in Chapter 5.

While these studies all focus specifically on social or linguistic effects on phonetic categorisation, Holt (2006, p. 4016) found specific evidence that 'speech and non-speech context stimuli jointly influence speech processing', by providing an acoustic context to a target stimulus that included either speech only, or speech and a series of non-speech tones, phonetic categorisation was affected in a way that reflected a combined influence of the two aspects of acoustic context. The role of non-speech context has also been shown in intracranial recordings of neuronal populations (Holt, 2006).

An adaptive resonance theory (ART) (Grossberg, 1980; Grossberg and Kazerounian, 2016; Goldinger and Azuma, 2003) approach to determining the fundamental unit of speech perception employs the concept of 'masking'. This is the process whereby higher order levels of mental representation are more accessible to conscious experience, even though lower level, more localised, resonances of activity may still be occurring and playing a functional role in speech perception processes. So, phonemes may always be operating, but once words are also activated, the conscious perception of the phonemes that are nested within that word will be masked. The exception to this is the transient higher level experience of the larger contextual scene, which does not mask words, but may still act in resonance with them. In this way, ART 'naturally resolves ambiguity: Just as phonemes and syllables are masked by word-level dynamics, ambiguous words will evoke multiple local resonances that

are masked by global, contextually coherent states' (Goldinger and Azuma, 2003, p. 309).
This account of the scaling nature of cognitive abstractions provides a sophisticated theoretical backdrop to the present experiment.

3.1.2 Perceiving the NZE short front vowel shift

The remainder of this section focuses on studies involving the DRESS and TRAP vowels. These vowels are part of the New Zealand short front vowel chain shift, and are raised and fronted compared to other dialects of English (Hay et al., 2008). Drager (2006, 2011) conducted two phonetic categorisation tasks which provide the foundation for the present experiment. The stimuli were a set of words which ranged on a continuum from sounding like the word *bed* to sounding like the word *bad*, and participants decided which word they heard. Drager (2006) examined whether perceived speaker age affects categorisation of these phonemes. Because the short front vowel chain shift is ongoing, younger New Zealanders tend to have more raised vowels. It was hypothesised that participants would also expect a younger speaker to have raised vowels, and thus be more likely to categorise an ambiguous token as *bad* if heard in the younger sounding voice, than a word with the same vowel formants spoken by the older sounding voice.

Nineteen listeners categorised words from the *bed–bad* continuum which were based on instances of the word *bad* uttered in a wordlist reading style by two males in the Canterbury Corpus. The two voices were manipulated to have identical vowel formants at each step of the continuum. The age of the two speakers was estimated by participants after completion of the experiment. Participants responded that they heard *bad* more often to the younger sounding speaker, as hypothesised, with perception thus reflecting patterns in production. One issue with the design of this experiment, acknowledged by Drager (2006, p. 65) was that the younger sounding voice's token was longer (234ms) than the older speaker's base token (174ms). This is problematic because DRESS is generally shorter than TRAP (though the difference was not significant in one study of early NZE, Langstrof 2004, cited in Drager, 2006). Hearing an ambiguous word, length could be used as a cue, with the shorter word being categorised as *bed* and the longer one as *bad* – a pattern which aligns with the results found in this paper, potentially undermining the conclusion that the differing response to the two voices was driven by expectations related to perceived speaker age. The present experiment will directly examine whether vowel length does indeed affect phonetic categorisation in this way. Another finding of note in this study was that female participants tended to respond *bad* more than males, reflecting the fact that females lead the short front vowel raising in NZE. This gender effect will also be tested in the present study.

In a second study, Drager (2011) looked again at the role of age on the perception of the boundary between these two vowels using the same stimuli (including two female voices which were included in the procedure of the experiment presented in Drager (2006) but excluded from its analysis). The second paper ran a similar experiment with new participants, adapted vowel continua and an adapted design. This time, words were paired with a photo, which visually suggested a speaker age to the participants. The pairing of photos with voices was crossed across participants, to mitigate the issue of vowel duration discussed above. Results were complicated by multiple interactions, but the overall finding regarding age supported the hypotheses, at least for the older participants: photos of younger-looking speakers increased the likelihood of responding *bad*. This is in line with the hypothesis that perception should mirror patterns of production in the speech community. This pattern did not hold for younger participants, for whom photo-age did not affect perception. There was also a significant interaction between participant sex and the sex of the speaker and their paired photo, with participants responding that they heard *bad* more when listening to a speaker of their own sex. This unexpected interaction is interpreted as possibly reflecting a pattern whereby the participants have more exposure to young same-sex voices (who would be more advanced in the short front vowel shift), but a more even distribution of voice-sex for older speakers. While this is an interesting suggestion, there are various problems with the design of the experiment which could also play into the unexpected findings in ways which were not controlled for nor examined. Three of these issues are briefly discussed below.

Firstly, while the effect of duration differences between voices was mitigated by the crossed design used in Drager (2011), we cannot be sure that the interactions in the model presented are not actually by-products of the potentially uneven assignment of the different voice-photo pairings to participants of different ages and genders. The duration difference between the two female voices (68ms) was similar to the difference between the male voices (60ms), and thus equally problematic.[1] It could be that older participants happened to be assigned the pairing of the shorter tokens with the older-looking speakers more often than were younger participants. This could lead to a length-based effect showing up as a subject-age by photo-age interaction, as reported.

Another issue with the design of both studies is that it is the participants themselves who estimated the age of the voices (Drager, 2006) or photos (Drager, 2011), after having already made judgements about the boundary between their DRESS and TRAP phonemes. There is a potentially problematic circularity here, which could have been avoided by having a different set of respondents estimating the voice and photo ages. A final critique of these studies is the absence of any control for the quality of the previously heard stimulus in the modelling of results. As suggested by the quote from Fry et al. (1962) above, hearing a very DRESS-like token will make a respondent more likely to hear TRAP on the next trial of the experiment. This local contrast effect will be explored in the present study, along with a direct examination of the role of vowel length on perception of these phonemes. While I have raised several methodological critiques here, it should be emphasised that Drager's studies introduced an original paradigm for exploring the role of social information in phonemic categorisation, and the studies do provide evidence that the position of a perceptual phoneme boundary can be shifted by subtle variations in the stimuli relating to patterns in the speech community.

More recently, and following the paradigm established by Drager's studies, Hay et al. (2017) marks one of the first attempts to explicitly test the effects of a non-speaker-related context on speech perception and production. This context is a location, specifically, the interior of a car. The authors argue that this context likely exhibits structured co-occurrence with speech variation. When we experience talk in a car, it will tend to be when the engine is running, creating a noisy environment.

[1] The decimal points shown in the duration information printed in Table 1 of Drager (2011) are most likely typos.

Talk in noisy environments (Lombard speech) is characterised by greater intensity, increased vowel durations and higher f0, and has indeed been documented for speech in cars (Jung, 2012). Importantly for the study by Hay et al. (2017) and also for the present experiment, Lombard speech also involves higher F1 (van Summers et al., 1988). The study asked whether this experience with Lombard speech in cars also affects production and perception of speech in a car even when it is totally silent.

Strong evidence for a location-based context effect was found for speech production, with higher amplitude, higher f0, and higher F1 in a silent car than in the lab. But the results of a perception experiment using synthesised continua from *bed* to *bad*, following the experiments by Drager outlined above, were not conclusive. There was some evidence of expectation of higher F1 for participants listening in the car versus those in the lab, the former being less likely to hear *bad* than the latter. This result was problematic, however, due to the lack of any significant effect of whether or not participants were exposed to actual noise in headphones. The authors argued that this was because participants did not believe they were really listening to a voice speaking against noise, but rather a voice recorded in a quiet environment, being played back against noise. Even though these perception results were fragile, they do present some important insights which need to be taken into account in designing the present experiment, namely the possible effect of noise on phonetic categorisation. These studies all provide a backdrop against which the present study is conducted. Let us, then, turn to the experiment I previously conducted as part of my Masters research, which the present experiment will replicate and extend. It is necessary to review that original version of the experiment in some detail in order to motivate the present replication.

3.1.3 Original version of experiment and reasons for replication

While Drager's studies hypothesised that perceived speaker age would affect vowel perception in the context of a chain-shift, the present study focuses on the combination of dialectal and stylistic information that has been conventionalised into accent differences between singing and speech. The premise is that all native speakers of NZE will have been exposed to systematically different vowel realisations for the phonemes DRESS and TRAP in musical and non-musical contexts, with raised variants in speech and open variants in song. As will be shown below, using data from the PoPS corpus, there is considerable acoustic overlap between spoken exemplars of NZE TRAP and sung exemplars of DRESS irrespective of the place of origin of a singer. The corpus data presented below will strengthen the empirical basis for this experiment.

Gibson (2010b) tested whether listeners perceive the boundary between DRESS and TRAP differently in speech and singing by setting the stimuli, which ranged on a continuum from *bed* to *bad*, to a musical background in one condition (with an instruction that they would hear singing), and no background audio accompaniment in the other. The results supported this hypothesis, with participants less likely to report hearing *bad* in the 'singing' condition. While the result was significant, and in the expected direction, there were several methodological issues with the design that warrant a careful replication of the study. The main issues were the pitch-shifting of half of the stimuli in only the music condition, the absence of a control condition with non-musical noise, and the inconsistent spacing of the vowels in the resynthesised continuum, with some stimuli very similar to their neighbours and others having very large gaps in F1/F2 space. These problems will be addressed in turn below, along with the way in which they will be rectified in the replication.

3.1.3.1 Pitch-shifted stimuli

In the original study, in order to give the impression that the voice was actually singing, half of the stimuli in the music condition were pitch-shifted up one semitone. This difference affected spectral structure subtly, but problematically. While the direction of the formant shift caused by the pitch shifting was in the opposite direction to the hypothesised result (half of the stimuli in the music condition had higher F1, thus sounding slightly more BAD-like, while the hypothesis was for participants to respond *bad* less often in the music condition), it may have had a range of other knock-on effects. It was a flawed experimental design to have unmatched stimuli across conditions, and the present replication remedies this issue by having identical stimuli in all three conditions. This replication simply uses the same pitch in both music and speech conditions, in hopes that the background music and the instructions that participants will hear singing will be enough to create the desired effect of priming song related expectations.

3.1.3.2 Noise control condition

There were only two experimental conditions in Gibson (2010b): music, and absence of music. Given the findings from Hay et al. (2017) discussed above, it seems plausible that any type of non-silent condition could invoke expectations of Lombard speech. There is thus no way of firmly concluding that it is the music itself causing the effect found in Gibson (2010b) — it could simply be evidence that listeners expect opener vowels when listening to words in noise than in silence. The present experiment thus introduces a third condition, in which the stimuli are heard in the context of a non-musical background noise. If the results of Gibson (2010b) were caused by expectations about singing as opposed to speech, then the music condition of the present experiment should have fewer *bad* responses than either the Noise or Silence conditions. If both of these phenomena are operating — expectation of more open vowels in singing, and also in noise — then they could be additive, with responses in the noise condition being intermediate to those in silence and in music. The results for F1 of DRESS and TRAP presented by Hay et al. (2017) find only subtle differences of around 50Hz in speech production between the car and the lab. Though the car is actually quiet, these results related specifically to these vowels in NZE and are therefore relevant. They suggest that the difference between noise is likely to be very much closer to silent conditions than it is to music, where F1 differences are in the hundreds of Hertz.

3.1.3.3 Continuum step sizes

The continuum of stimuli used in Gibson (2010b) were not manually measured prior to running the experiment. After data collection, it was discovered that the tokens used in the experiment actually exhibited wide variability in the difference between neighbouring tokens on the continuum (see Figure 6.2 in Gibson, 2010b).

A basic premise of the experiment is that the continuum should represent small and even variants of the vowel which range between the two endpoints. This issue was also the case for the stimuli used in Hay et al. (2017). In order to create more carefully controlled stimuli in the present experiment, the resynthesis process involved a pragmatic and iterative approach, to get the formants of the resynthesised soundfiles to be as close as possible to the intended values. This process will be discussed in 3.4.2.

3.1.3.4 Vowel length

One innovation to this experimental paradigm will be added in the present study, namely vowel length. Despite the findings of Langstrof, 2004, (cited in Drager, 2006), many varieties of English have a vowel length distinction between DRESS and TRAP, with the latter being longer. To test this in modern NZE, a basic analysis of wordlist data in the Canterbury Corpus was undertaken, which includes the words *bed* and *bad*. To determine the length ratio between the two vowels, the average duration of the vowel segment in all wordlist instances of these words was extracted, using the force-aligned boundaries existing in LaBB-CAT. For *bad* the average vowel length was 294ms, and for *bed* it was 235ms. The length of *bed* is thus 80% the length of *bad*. Since the experiment involves citation-style, not conversational, instances of the words, it was the ratio of lengths in the wordlist data that was used to guide the length ratios used in the present experiment. More detail is given below in Section 3.4.2.1.

Length will be included in this experiment to see whether people shift the boundary between the vowels to a more open position when a longer stimulus is heard. Length may be of particular importance in the present study, given the focus on singing, in which the length distinction between DRESS and TRAP may be neutralised, due to the rhythmic demands of a song. If this is the case, then we should see more *bad* responses for long stimuli in the silence and noise conditions, but no length difference in the singing condition. This will be tested and discussed in the analysis of results.

3.2 Using PoPS to Motivate the Phonetic Categorisation Task

In this section, I briefly return to the PoPS corpus, using it now not just as an object of study, but also as a methodological tool — to provide reference values for the design of the vowel continuum. The results themselves extend the findings of Chapter 2, but the main aim here is to establish the kind of acoustic experiences the participants in the PCT might have had with the DRESS and TRAP vowels in the context of song. For this reason, the analysis is reported in this chapter rather than having been reported in Chapter 2. Additionally, the analysis here is restricted to F1, since vowel height is the key acoustic variable of interest in the PCT. F2 was not measured for these vowels.

3.2.1 DRESS and TRAP: Method

A search was conducted in LaBB-CAT for DRESS and TRAP vowels produced by males before non-nasal coronals so that tokens would be comparable with *bed* and *bad*, the words to be used in the PCT. The first viable token in each song was analysed for F1 in Praat, with number of formants set to 6 and maximum formant set to 5500Hz (the use of these formant tracker settings was an accident, they should have been set to 5000Hz and 5 formants maximum for the analysis of these male voices).

For each token, the first step was to check that the syllable was stressed, and then assess whether the vocal was isolated enough to allow Praat to sensibly track the formants. The most problematic cases are loud noise bursts co-occurring with the vowel, and ongoing synth pads, which are especially problematic since they look rather similar to vowel formants. This was determined auditorily and visually. If the token looked reasonable, I took as close to the mid-point of the vowel as I could, to mitigate the influence of coarticulation, whilst taking care to avoid any obvious deviations in formant tracking caused by instrumental sounds. Falsetto tokens were excluded.

In addition to the sung data I added measurements from the Canterbury Corpus to allow for comparison between SPMSS and NZE.[2] This dataset included 20,116 tokens of DRESS and 9,210 tokens of TRAP.

3.2.2 DRESS and TRAP: Results

The raw results are shown in Figures 3.1 and 3.2, along with the F1 values from the Canterbury Corpus and a normal distribution around the mean values for Western male speakers reported in Clopper et al. (2005) (DRESS mean F1=550Hz, sd=60Hz; TRAP mean F1=700Hz, sd=100Hz). The first thing to notice in Figure 3.1 are the areas in US pronunciation (right hand panel) where black and black overlap, that is, where sung and spoken instances of DRESS have the same realisation, and the areas where orange and orange overlap (where sung and spoken TRAP are similar). These are areas of F1 space where DRESS and TRAP are unlikely to be confused, irrespective of whether a US person is singing or speaking. Both singing and speech also have large areas of overlap between DRESS and TRAP. This reflects speaker variability — the core issue of lack of acoustic invariance that makes indexicality so important to speech perception (Johnson, 2006). On the NZ side (the left panel of Figure 3.1), we see that DRESS and TRAP are both raised in speech compared to song, they have a lower F1. The important finding here is that spoken TRAP and sung DRESS are centred around roughly the same F1 value. When a listener hears a male voice producing an F1 of 550–600Hz, then the best way to determine the vowel category is through context: if it is song, it will be DRESS, and if it is speech, it will be TRAP.

In Figure 3.2, the same data is presented, but faceted this time according to context. This shows that NZ male pop singers realise DRESS with an identical vowel height to US singers, while their sung TRAP is somewhat raised compared to the US singers. For DRESS, the mean F1 for US singers was 588Hz and for NZ singers it was 578Hz. The difference was not significant (2-tailed t test p=0.66). For TRAP,

[2]These values had already undergone data cleaning and outlier removal for a different project.

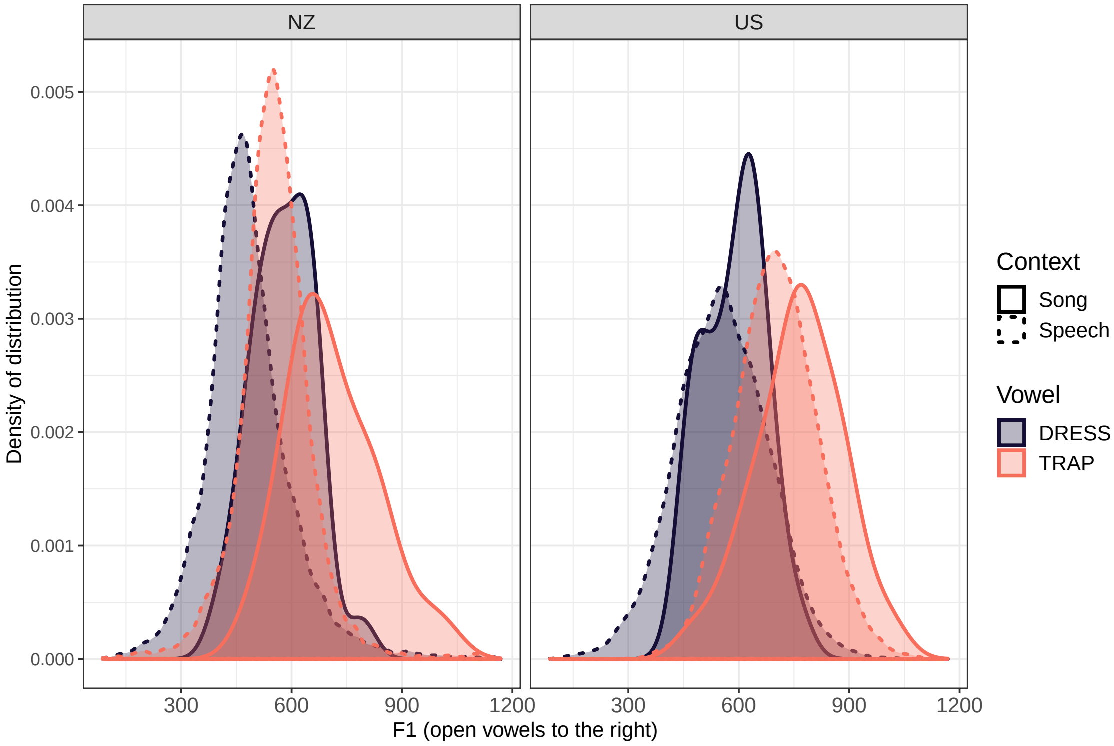


Figure 3.1: Density distribution of F1 for US and NZ males' DRESS and TRAP vowels, grouped by context. Sung data from 120 tokens of PoPS. NZ speech from 13756 tokens of males in Canterbury Corpus (CC) born after 1965, US speech is a normal distribution around the mean values from Clopper et al. (2005) for Western male speakers, with the same standard deviation as the CC data.

the mean F1 for US singers was 763Hz and for NZ singers it was 709Hz. This difference was not significant, but there was a trend (2-tailed t-test p=0.09) for New Zealand male pop singers to use a somewhat NZ-like TRAP vowel in singing when compared to US singers. However, comparing New Zealanders' spoken and sung TRAP (Figure 3.1) shows that NZ sung TRAP is still very much more open than it is in speech. While this will not be investigated here, it is likely that NZ-accented rappers identified in Chapter 2 such as David Dallas, Jody Lloyd, Sid Diamond and Machete Clan are driving this difference between NZ and US singing for TRAP. The lack of difference between registers for DRESS suggests that TRAP may have greater sociolinguistic salience. The important point is that the distributions for spoken NZE TRAP and SPMSS DRESS are almost completely overlapping, creating a region of ambiguity in which the acoustic value of F1 is not enough to determine vowel categorisation — context must be employed. It is this region of ambiguity that is exploited in the present experiment.

3.2.3 Formant values for the DRESS–TRAP continuum

With several studies having now looked at these particular variables using this same task, it seems prudent to review the formant values used in the various continua described in previous studies. It is important to keep in mind that in the context of the task, the position of the boundary between DRESS and TRAP for a listener will be related to the outer limits and step sizes of the continuum itself, which help

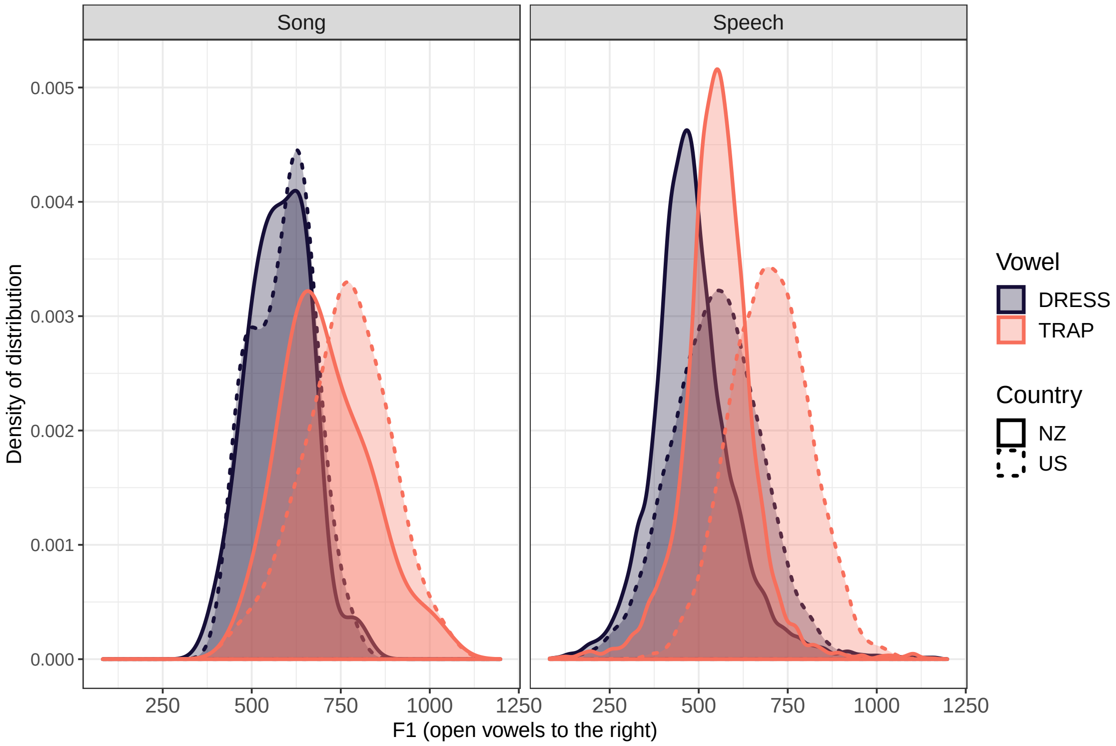


Figure 3.2: Density distribution of F1 for males' DRESS and TRAP vowels in singing and speech, grouped by place of origin. Sung data from 120 tokens of PoPS. NZ speech from 13756 tokens of males in Canterbury Corpus (CC) born after 1965, US speech is a normal distribution around the mean values from Clopper et al. (2005) for Western male speakers, with the same standard deviation as the CC data.

to form the listeners' frame of reference for vowel perception. Participants will tune in to the stimuli that are presented to them, and may attempt to provide a roughly even number of responses. Table 3.1 shows the minimum and maximum values for F1 and F2 in each of the four studies reviewed above, along with the number of steps that this continuum was divided into. The table also introduces the values to be used in the present study.

Table 3.1: Summary of formant values (male voices only, rounded to 5Hz) at outer ends of DRESS–TRAP continua in previous studies and present experiment.

Study	F1 range	F2 range	Continuum steps	Manipulation
Drager (2006)	410–590	2235–1970	10 steps	Perceived speaker age
Drager (2011)	420–615	2140–1950	9 steps	Speaker age and gender
Hay et al. (2017)	500–680	2040–1810	7 steps	Location and background noise
Gibson (2010)	450–660	2110–1800	8 steps	Background music vs. silence
Present Study	385–730	2130–1810	6 steps x 2 lengths	Music vs. noise vs. silence

Given the measurements for singing and speech presented above, and taking into consideration the vowel continua presented in Table 3.1, the following values were decided upon as the desired end points for the continuum: 390–730Hz for F1 and 2030–1810Hz for F2. These end points were then broken up into six equal steps. The reason for having only six steps in the continuum, where the original version of the experiment had eight, was to allow for the addition of the length distinction such that each vowel quality had a short and a long token.

The average F1 values for DRESS and TRAP for male pop singers in the PoPS corpus were 583Hz and 736Hz respectively. Thus, the most *bad*-like token, step 6, is similar to an average sung TRAP (which for NZ singers was 709Hz). The first step of the continuum, however, is much more raised than anything the listeners are likely to have encountered as DRESS in a song, but does represent spoken NZE DRESS. The most raised token of DRESS in the PoPS corpus was 404Hz, so while not inconceivable, Step 1 would be an outlier DRESS vowel in a sung context.

3.2.4 Lexical frequency of *bed* and *bad*

To conclude this background section, I will briefly consider the relative lexical frequency of the words *bed* and *bad* in singing and speech and discuss how that might impact the experiment. I will then summarise the predictions, before moving on to the experiment's methods.

It is ideal that *bed* and *bad* are both high frequency words, and of a similar frequency. The frequencies of *bed* and *bad* were calculated in Celex (based on the EFW file, which is for frequencies of English wordforms, as opposed to lemmas) and in the lexical frequencies obtained from the LyricsPlanet website (see Section 2.3.3 for an introduction to these lexical frequency measures). The sum of frequency counts in Celex is 18.8 million occurrences, and for Lyrics Planet is 14.9 million. The number of occurrences per million words for each word in each of the two corpora is as follows: *bed* occurs 244 times per million words in Celex (raw n = 4376) and 229 times per million in Lyrics Planet (raw n = 3378) — *bed* thus occurs at a similar frequency in Celex and in song lyrics; *bad* occurs 209 times per million words in Celex (raw n = 3755) and 584 times per million words in Lyrics Planet (raw n = 8624) — *bad* thus occurs more frequently in songs than it does in speech/writing. In a spoken context, *bed* and *bad* are of a similar frequency, while in a musical context *bad* is more frequent than *bed* . This pattern is supported by the frequencies of these words in the (much smaller, 36,109 word) PoPS corpus: *bed* occurs 12 times (for comparison's sake, this would scale up to 332 occurrences per million words) and *bad* 21 times (582 per million). In the ONZE corpora (totalling 3.4 million words), *bed* occurs 1547 times (455 per million) and *bad* 1584 times (466 per million), this provides further evidence that the frequencies of *bed* and *bad* are similar to one another in speech contexts.

3.2.5 Summary of predictions

> Reference Frame Hypothesis: New Zealanders expect to hear SPMSS in popular music, and will therefore shift the boundary between DRESS and TRAP according to the context. Raised TRAP will be expected in non-musical conditions, resulting in a larger proportion of *bad* responses. In the musical context, however, open short-front vowels will be expected, resulting in greater likelihood of perceiving *bed*.

Various predictions have been established from the literature reviewed above. The primary prediction of this experiment is that the findings of Gibson (2010b) will be replicated: participants should expect the boundary between DRESS and TRAP to be in a more open position in the Music condition than in Silence (that is, there will be fewer *bad* responses in Music). Additionally, there should be fewer *bad* responses in Music than Noise, ruling out the possibility that the results of Gibson (2010b) were based on masking rather than the music itself. There should be no significant difference between Silence and Noise, though if there is a trend, it should be for participants to hear less *bad* in Noise than Silence (that is, an expectation of opener vowels in the Noise condition, since Lombard speech has opener vowels).

Regarding length, shorter stimuli should sound more *bed*-like than longer stimuli in general, though this length effect may be reduced or absent in the music condition. Regarding participant variables, younger participants will be further along in the short front vowel shift (Hay et al., 2008), having more raised vowels in speech. This should show up in the perception experiments as having a raised perceptual boundary, and responding that they hear *bad* more often than older participants. A similar prediction can be made for gender, such that females are leading the sound change and thus have a more raised boundary between DRESS and TRAP. It is therefore predicted that females will hear *bad* more often than males. Finally, while all of these pre-existing factors will effect a participant's perceptual reference frame going into the experiment, it is expected that there will be a local contrast effect, such that participants will be more likely to hear *bad* after a very *bed*-like stimulus, and vice versa. They will also likely adapt to the continuum as the experiment progresses, with a bias towards distributing their responses evenly between the two choices.

3.3 Description of Participants

This section begins with information about ethics and recruitment, and then presents a summary of the participants' survey responses. Recruitment was done through University of Canterbury social media pages, along with physical placement of posters around campus. These recruitment materials (which can be found in Appendix B) asked for native speakers of NZE, defined as people 'who have lived in New Zealand for most of their life' (as defined below), to 'be involved in a linguistics experiment which examines speech perception'. My own network through Facebook also yielded several participants. The cohort is thus largely, but not entirely, made up of students of the University of Canterbury, with 28 of the 36 participants identifying themselves as a student on the questionnaire (sometimes alongside another occupation). One participant had a linguistics background, but did not have any knowledge about the topic of the experiment. The recruitment procedures, along with all the text used, were approved by the University of Canterbury Human Ethics Committee (under application number HEC 2016/20/LR-PS Amendment 2). Relevant files are included in Appendix B. All participants (the 36 analysed and also those involved in the piloting phase) received a $10 voucher for their participation.

The experimental design phase involved running several pilot participants, particularly to get the timing of the stimuli to work correctly with the music without any latency issues. Once all main aspects of experiment design such as timing and instruction wording were settled, and the preregistration had been filed, the data from three further participants was used to calibrate the signal to noise ratios on the basis of error rates to the Music and Noise conditions of the lexical decision task (see Section 4.2.2.4 for further details about this process). The data from these three participants was discarded on the basis of these error rates and not analysed further. After the amplitude of all the soundfiles had been settled upon, one further participant's data had to be excluded from the analysis because they did not meet the criteria of having lived outside of NZ for less than six years of their life. This person's data was never analysed. Their non-eligibility only became clear when they were filling out the survey after having completed the experiment. After this, I asked potential participants much more directly about whether they met the inclusion criteria — having spent less than six years outside of New Zealand, and having no diagnosed hearing impairment — prior to confirming their participation.

Regarding the time spent overseas criterion, a discrepancy between the preregistration and the survey wording was discovered midway through data collection. The questionnaire had the following categories for total years spent living outside NZ: 'less than 2'; '2–4'; '4–6'; '6 or more', while the preregistration says: 'Participants must have grown up in NZ, defined as having spent less than 5 years total outside of NZ in their life'. It was decided to change the cut off to six years once the oversight in the survey design had been noticed, since it was not possible to make a five year cut-off with the available information.

As well as collecting demographic information and asking questions about the kind and amount of music the participants listen to, the survey also asked participants how much they felt like the stimuli in music conditions sounded like singing. It was deemed important that some listeners might be convinced that they are hearing a singing voice in the music conditions, while for others, the words may not be as strongly associated with the musical background. The questionnaire itself can be found in Appendix B, and the survey responses are summarised below. Some of these responses are of interest in their own right, while others provide background information for the statistical models which follow.

3.3.1 Questionnaire responses: Demographics

The distribution of the 36 participants across the various categories for each survey question is shown below:

- Gender: female, n=25; male, n=11; other, n=0.
- Age: under 20, n=7; 20–24, n=14; 25–29, n=9; 35–39, n=1; 40–49, n=2; 50–59, n=3.
- Ethnicity: New Zealand European/Pākehā/New Zealander n=33; Māori/Pasifika, n=3.[3]
- Years outside NZ: less than 2, n=23; 2–4, n=10; 4–6, n=3.
- Handedness: right-handed, n=33; left-handed, n=3.
- Languages spoken: monolingual, n=31; bilingual/multilingual, n=5.
- Socio-economic status: This was calculated using the New Zealand socio-economic index published by Statistics New Zealand (Fahy et al., 2013), by

[3]Further detail about ethnicity is concealed to ensure the anonymity of the participants, since some of the combinations of ethnicities may make individual participants identifiable.

finding the relevant score for participants' parents' occupations, and also the participant's own occupation if they included anything other than 'student'. The index provides a score between 0 and 100 for a wide range of occupations, with higher numbers representing higher socio-economic status, with categories grouped hierarchically from general areas of employment to highly specific job titles. The questionnaire responses for all occupations were assigned a value as specifically as could be warranted by the wording given by participants on the survey. The mean of the available scores was taken. This mean was based on a minimum of one score, in the case of a single parent's occupation being given, to three scores, in the case where occupations were given for two parents and for the participant themselves. The mean of these mean scores was 58.7, with a standard deviation of 13.6. The maximum mean score was 90 and the minimum was 25.

3.3.2 Questionnaire responses: Music consumption patterns

The remainder of the questionnaire pertains to music. A total of 8 of the 36 participants self-identified as musicians. A range of genres were listed on the questionnaire, and participants circled the genre(s) they 'like listening to the most'. The results of this section of the survey are shown in the left-most columns of Table 3.2.

Table 3.2: Description of the two participant clusters' genre preferences, sorted from Cluster 1 (C1) favoured genres to Cluster 2 (C2) favoured genres.

Genre	Total 'likes'	% of C1 liking genre	% of C2 liking genre	C2–C1
R&B/Soul	9	43.75	10	-33.75
Electronic	9	37.5	15	-22.5
Hip Hop/Rap	12	43.75	25	-18.75
Classical	8	25	20	-5
Folk (responses volunteered in 'Other')	2	6.25	5	-1.25
Blues	8	18.75	25	6.25
Pop	26	68.75	75	6.25
Jazz	4	6.25	15	8.75
Country	5	6.25	20	13.75
Rock	15	18.75	60	41.25
Alternative	13	6.25	60	53.75
Singer-Songwriter	13	6.25	60	53.75
Other genre (open-ended)	6			

In order to sensibly reduce the dimensionality of this information down to something which could be included in statistical models, a clustering method was used (this was also in line with the wording of the preregistration). Participants were assigned to two groups based on the top level split of a cluster analysis of their binary responses to the twelve genres. This analysis was performed by creating a Jaccard distance similarity matrix (using R command *dist* with method = binary), and then running a hierarchical clustering algorithm on that matrix (using R command *hclust* with method = ward.D2). A transposed version of the data, i.e. clustering the genres based on their patterns of responses from the 36 participants, gives a more interpretable view of this cluster analysis. Figure 3.3 shows how the genres were grouped. This is an intuitively logical grouping, with related genres ('art-music' genres jazz and classical, or 'urban' genres R&B and hip hop) appearing under the same branches of the tree.

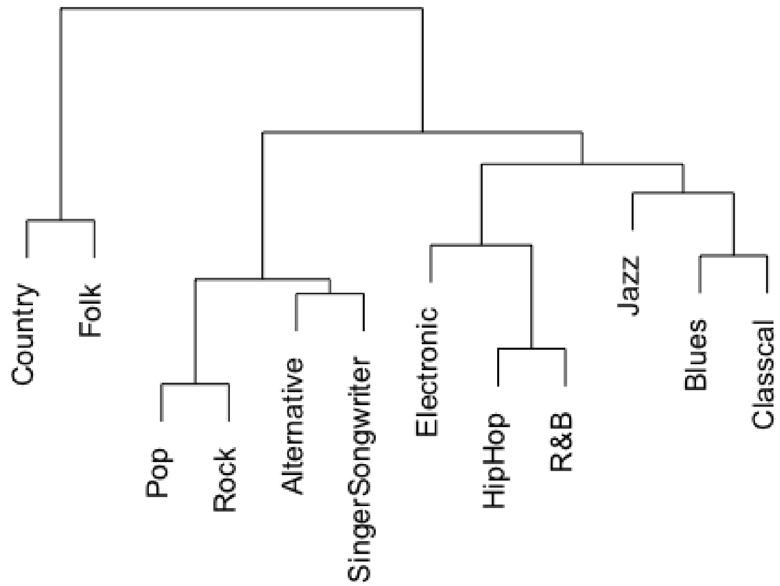


Figure 3.3: Clustering of genres based on binary responses by 36 participants.

For modelling purposes, it is the grouping of *participants* that was performed, splitting participants into clusters based on genre-liking. The top level split of this cluster analysis yielded 16 people in Cluster 1 and 20 in Cluster 2. By looking at the percentage of people in each cluster that like each genre, it is possible to see the logic behind this grouping of participants. While the majority of people in both clusters like pop, Cluster 1 participants also tend to like hip hop, r&b, and electronic music. Those in Cluster 2 tend to like the alternative, rock and singer-songwriter genres. These details are shown in Table 3.2, with the genres sorted according to the 'C2–C1' column, which represents the percentage of people in Cluster 2 who like the given genre, minus the percentage of people in Cluster 1 who like it. These clusters will be operationalised as an independent variable in the modelling of the results.

Table 3.3: Summary of responses (counts, means and standard deviations) to music-related questions on Likert scales.

Survey question	Response frequencies					Mean	S.D.
	1	2	3	4	5		
How much time do you spend listening to music?	3	6	14	5	8	3.25	1.23
How much of the music you listen to is by American artists?	0	5	12	19	0	3.39	0.73
How much of the music you listen to is by New Zealand artists?	2	26	8	0	0	2.17	0.51
Do you think it is surprising to hear New Zealand accents in songs?	5	9	6	10	6	3.08	1.34
In the parts with music, did it sound like the voices were singing?	16	12	6	2	0	1.83	0.91

The other music related items in the survey, all of which use Likert scales from 1–5, are summarised in Table 3.3, showing frequency distributions across the response scale along with the mean response and its standard deviation. There is a wide range of variation with respect to the amount of time people spend listening to music (mean=3.25, s.d.=1.23), and in general people say that a high proportion of the music they listen to is by American artists (mean=3.39), and a smaller proportion is by NZ artists (mean=2.17). In general, people say they would be surprised to hear a NZ accent in a song, though there is wide variation in the responses to this question (mean=3.08, s.d.=1.34). Finally, most people said that the stimuli did not sound at all like singing in the music conditions (mean=1.83). Note that since the survey was presented just after the lexical decision task (LDT), which was much longer than the PCT, the responses to this final question are likely to be more relevant to the latter experiment.

3.4 Method

With the participant population already described above, this method section presents a description of the stimuli used in the experiment, with some detail about the process of designing those stimuli. The procedure of the experiment and how it was set up in E-Prime will then be explained.

3.4.1 Participants

Thirty-six native speakers of NZE participated in the experiment. They had all spent less than six years outside of NZ in their lives, and had no diagnosed hearing impairments. See Section 3.3 above for a full description. To ensure their anonymity, each participant was assigned a number prior to their arrival. This was marked on the questionnaire sheet, but not the consent form which would have their signature on it. At the end of the session, these two forms would be separated so that the identity of participants could not be tracked from that point on.

3.4.2 Experimental stimuli

The stimuli used in the phonetic categorisation task include the twelve tokens that comprise the resynthesised continuum from *bed* to *bad*, the musical accompaniment, and the background noise. This section provides detail about the creation of these stimuli. The noise file is shared with the lexical decision task to be presented in the next chapter, but is described in this section, along with some details about two other types of noise that were piloted and ultimately rejected.

3.4.2.1 Stimuli for resynthesised vowel continuum

The original sound recording used for resynthesis was recorded by the author with the intention of falling roughly in the middle of the desired continuum, by producing a 'normal' NZE *bad*. It had an F1 of 505Hz and an F2 of 1910Hz. The formant values of the original recording are however essentially irrelevant since all stimuli used in the experiment had F1 and F2 resynthesised, but the aim was for neither end of the continuum to sound more or less manipulated than the other. The intended pitch of the recording was the musical note E3, which equates to a fundamental frequency of 164.8Hz, the recorded token was near to this, with an f0 of 165.7Hz. This fits with the key of the musical accompaniment, producing the 5th note of the A major scale. The stimulus plays on the first beat of every second bar, over the A major chord and the D major chord. While not melodic as such, the 'sung' word fits in to the tonality of the music coherently.

The length of the original recording was 525ms, from onset of the pre-voicing of [b] to the tail of the burst from the [d], with the vocalic portion lasting 380ms. The word was performed with the tempo of the music in mind (80 beats per minute), such that it would sound like it was 'in time' with the music, by being an eighth-note length (an eighth note at this tempo is 375ms, roughly matching the vocalic portion of the recorded word). A short version of the soundfile was made prior to vowel resynthesis. The vowel in this shorter stimulus was made to be 80% of the length of the vowel in the long stimuli, based on the analysis of Canterbury Corpus wordlist data discussed above (in Section 3.1.3.4). To create the short stimulus, 75ms was removed from the middle of the vocalic portion of soundfile, taking care to make cuts at zero point crossings. To ensure imperceptible edits, these cuts were made by closely inspecting the waveform and making the cut at a point which appeared to be at an analogous part of the complex waveform. The shorter soundfile was thus 450ms long with a vocalic portion of 305ms. The vowel continuum was created in Praat using a script originally written by Paul Warren and subsequently adapted by Jen Hay and then by myself for this project. The continuum has six different vowel qualities including the two most extreme tokens. Note that this script is largely the same as that used in Gibson (2010b), but has a different sample rate setting (9500Hz instead of original 11000Hz) in the resampling options for Praat's 'To LPC (burg)' command. Through trial and error, it was found that these new settings made for a more naturalistic and less muffled sounding resynthesis of the vowel. Another aspect of the resynthesis that involved some trial and error was the length of the transition allowed by the script to move from the original formants to the target values. This was eventually set to 50ms at each end of the vowel.

Initially, two methods were trialled for choosing the values of the intermediate steps of the continuum: even steps based on Hz values, and even steps based on ratios of Hz values. The latter was expected to produce more psycho-acoustically similar gaps between neighbouring stimuli, but after listening to the two resulting continua, it was decided that linear steps in Hz produced more even sounding intervals between the vowels.

Once the desired values for the vowel stimuli had been determined, the F1 and F2 of the resulting vowels was measured, and discrepancies between the intended and resulting vowel formants were found. The script does not produce results exactly matching the intended values, which was the reason for the uneven step sizes in the continuum used in (Gibson, 2010b). To get around this issue, an iterative approach was taken, whereby the formants of resulting audio files from the first resynthesis were measured. The difference between the target value and the achieved value was then used to offset the values fed into the script a second, and third time. This way, the resulting audio files were close to the intended targets.[4] It should be emphasised that the final stimuli were the result of only a single resynthesis. The tuning process described here was used solely to improve the target values selected for that resynthesis. This process was undertaken for both F1 and F2 of each of the 6 continuum steps. Table 3.4 shows the desired formant values, the formant

[4]For example, the target F1 for continuum step 1 was 390Hz. When 390Hz was used as input to the script, the resynthesised token had an F1 of 361Hz, much too low. So the second attempt used an input value of 390Hz + 29Hz (the difference between intended and actual) = 419Hz. This resulted in a file with an F1 of 400Hz, much closer to the target. A third attempt then used an input value of 419Hz - 10Hz = 409Hz. The resulting file had an F1 of 385Hz, which was deemed acceptably close to the target.

measurements from a set of stimuli created by using the desired targets as the input to the resynthesis script, the values actually used to input to the resynthesis script once tuning of values had been done, and measurements of the final stimuli. It can be seen that the problem with uneven steps was overcome. Once this process had been undertaken with the short version of the original recording, the tuned input values were applied to the long version. The final step in creating the stimuli was to reduce the amplitude of all twelve stimuli to 0.105 Pascal (using the 'Get root-mean square...' and 'Multiply' functions in Praat). This equates to 74.4dB SPL.

Table 3.4: Summary of formant values (Hz) relevant to creation of the vowel continuum from *bed* to *bad*.

	Desired formant values		Resulting formant values if using desired values as input to script		Iteratively tuned final input values		Resulting formant values of final stimuli	
Continuum Step	F1	F2	F1	F2	F1	F2	F1	F2
1	390	2130	361	2093	409	2157	384	2132
2	458	2066	459	2021	457	2103	458	2067
3	526	2002	521	1952	529	2050	526	2000
4	594	1938	618	1905	576	1980	599	1935
5	662	1874	647	1848	690	1902	652	1873
6	730	1810	670	1792	782	1835	731	1810

3.4.2.2 Music stimuli

The music condition used the same background musical accompaniment as that used in Gibson (2010b). The accompaniment is in a soft pop style consisting of guitar, keyboard, bass and drums cycling around the chord progression |A |F#m |D |E7 | in a 4/4 time signature with a tempo of 80 beats per minute. Once the length of the experiment had been finalised, the soundfile was edited to fit that length. The music condition had a four bar long introduction to enhance expectations of singing, and the ending of the music was edited to occur a few seconds after the last stimulus (the length of the music file was 2:41).

3.4.2.3 Noise stimuli

Several versions of the noise condition were attempted during piloting, and for the purposes of completeness, I include some detail on this process here. The ideal for the noise condition would be a noise source which is spectrally identical to the music passages for the experiment, but without priming 'music' for the participant. In my first attempt to do this, the actual music track was chopped into short segments of 1–5 milliseconds and played in a random order. Max 6.1.10 (`https://cycling74.com`) was used to automate this process. Max is an open source tool for working with audio, graphics and interactivity. I made a Max 'patch' (interactive programme) to create this version of the background noise. Upon opening the Max patch, a dialog appears so that the user can select an audio file. Arguments are entered for three customisable properties of the noise output: a) the length of the shortest sample, b) the difference in length between shortest and longest samples, and c) how often to change and randomly select the size of the sample within that specified range. The reason for this third parameter was to avoid a sense of rhythm. When the resulting noise soundfile was presented to colleagues in a workshop, I received general feedback that the noise was still too musical sounding, and so it was abandoned.

My second attempt involved making a music-spectrum shaped noise in Matlab. This was created based on the two background music files used in the experiments along with four other pop songs. The resulting file was used with a pilot participant. The participant reported hearing 'call-centre' music 'behind' the noise. Overlapping tonalities across the various songs may have resulted in chords being easily imagined in this noise. If the noise condition sounds like music to the participant, then it is not serving its function as a control environment, and so this version of noise was also abandoned.

In the end, pink noise with a wide-band cut at 3kHz was used. Pink noise (or 1/f noise) has equal energy across each octave, that is, the intensity decreases as frequency increases. White noise, by comparison, has equal intensity at all frequencies, so the high frequencies are psycho-acoustically dominant. Due to the upward spread of masking (see, e.g., Oxenham and Plack, 1998), pink noise is appropriate for masking speech. The 3kHz cut was made since there is additional sensitivity to frequencies in the range of 2–5kHz associated with the resonance of the auditory canal, the mean of which is about 2.8kHz in adults (Pierson et al., 1994).

The noise file was cut to the appropriate length for the experiment (2m31s), and given a fade-in at the start and a fade-out at the end, ensuring this fade began after the final response would be given. To summarise, the noise condition is not spectrally equivalent to the music, but the use of this type of noise does ensure that any result for music cannot be attributed to its non-silence. This is the main function of including the noise condition, as a control.

Because the noise is spectrally and temporally dense, it interferes more with speech perception than does the music. It thus needs to be at a lower average amplitude in order to cause a similar degree of masking. The signal to noise ratios were carefully calibrated by conducting pilot runs of the lexical decision task and looking at the error rates in each condition. The aim was to have similar error rates in the Music and Noise conditions. It was decided through this process that the music should be 4 dB SPL louder than the noise. It was also decided that the PCT could afford to have a lower overall signal-to-noise ratio since the task is much easier than the LDT. The assumption behind this decision was that louder music will increase participants' likelihood of perceiving the voice as singing. The resulting amplitudes, then, for the PCT, as measured by Praat's 'Get root mean square' function, were .105 Pascal (74.4 dB SPL) for the test stimuli, .065 (70.2 dB SPL) Pascal for the music and .041 Pascal (66.2 dB SPL) for the noise. Since calibration of the amplitude of the music and noise soundfiles was tested using the LDT, it will be described in more detail in Section 4.2.2.4.

3.4.3 Procedure

Participants were welcomed and asked to read the information sheet and sign the consent form (see Appendix B.3). As can be seen from the information sheet, there is no mention of 'music' or 'singing' in these materials, nor in the recruitment phase, so those who had the silence and noise blocks first did not know that the experiment involved music until encountering the instructions for the Music condition. Participants listened to the stimuli through headphones, while seated in a sound-treated booth at the University of Canterbury. The laptop running E-Prime was situated in an adjacent room, and a monitor presenting instructions was placed in this room, facing the participant, through a window. The volume on the laptop was set to the same level for all participants which was deemed through piloting to be a comfortable listening level, and while participants were told they could ask for the volume to be adjusted if they wished, none of them did so.

Once they were seated, I said 'you will be listening on headphones — all the instructions you need will be on the screen, and you will be deciding whether you hear this word or this word (pointing at the *bed* and *bad* labels) and responding by pressing these buttons'. In this way I avoided using the words *bed* and *bad* verbally. The visual instructions read: 'you will hear a New Zealander singing/saying either the word *bed* or *bad*' ('saying' is used in the Silence and Noise conditions and 'singing' in the Music condition). Participants signalled which word they had heard by pressing a button on an SRBox response box with labels next to the buttons. Condition order was counter-balanced across participants, as was the position of the words on the response box (*bed* was on the left for half of participants and on the right for the other half). In between blocks, participants were given encouragement and told to take a break for as long as they like and then to click any button on the response box to continue. The on-screen instructions prior to subsequent blocks stated 'in this next block, you will hear singing/speaking'.

At the beginning of each block, the background noise file was triggered (in Silence a placeholder silent file was used) after which a wait time elapsed which was longer for Music (12,120ms), to allow for the four bar instrumental introduction, and shorter for Silence and Noise, where the interval between pressing the button (which triggered the start of the noise file) and the first stimulus was 3000ms. The inter-stimulus interval was fixed, at 6000ms, or two musical bars (eight beats), rather than being determined by the response, so that the stimuli in the Music condition would always fall on the first beat of the bar. The first stimulus in every block is the short version of continuum step 4 (4S). This is followed by the other eleven stimuli, in random order. Then all twelve stimuli are played again, in random order. This way, identical stimuli are separated from each other to an acceptable degree. People will sometimes have encountered the short and long versions of stimuli in neighbouring positions, and very occasionally they may have encountered the same stimulus in Trial numbers 12 and 13.

3.4.3.1 Experiment design in E-Prime

E-Prime version 3.0.3.43 was used to run the experiment. The Order parameter was set to 'permutation'. This means that all possible condition orders are cycled numerically by participant number. The participant number which I began with happened to be in the middle of this cycle, not at the start of it, meaning that the condition orders were not distributed perfectly evenly, with one extra participant in one of the condition order permutations. The 'Session' number was used to keep track of whether the *bed* label was on the left- or right-hand side of the response box. This allowed for the data to be easily recoded so that the responses would be aligned despite having different raw values. These responses of *bed* and *bad* were exported from E-Prime along with a wide range of other information, and were then prepared in RStudio for the analyses presented below.

3.5 Results

A clear distinction between deductive and inductive reasoning processes is an important value of the scientific method. To reduce the potential for experimenter bias, this experiment was preregistered on aspredicted.org prior to the commencement of data collection.[5] These documents provide a roadmap for data analysis, and increase transparency.

Analysis of results begins with a summary of the raw data, including a between-participants analysis of the first token of the experiment, complemented by the preregistered analysis of responses to an individual stimulus (4S). This is followed by a statistical analysis of the full dataset which strictly follows the preregistered model fitting procedure. Finally, a refined model-fitting procedure is explored and the resulting model is presented.

3.5.1 Data processing, raw results, and analysis of data subsets

Data was collected for a total of 2592 responses (72 trials for each of 36 participants). This included 1207 *bad* responses (46.6%) and 1385 *bed* responses (53.4%).

3.5.1.1 Analysis of first-trial of experiment

Part of the preregistered analysis of results was a simple between-participants comparison of the very first token of the whole experiment, which was always stimulus 4S (continuum step 4, short). The preregistration called for a Fisher's Exact test of this first trial. Across the 36 participants, 12 had the Music condition first, 13 had Noise, and 11 had Silence. The fact that there were not 12 participants starting in each condition was an accidental quirk caused by starting the permutation procedure (to counter-balance condition orders) in E-Prime on a number not at the start of the permutation sequence (see 3.4.3.1). Of the twelve who had Music first, three responded that they heard *bad* (25%). Nine out of the thirteen who had Noise first heard *bad* (69%), and ten of eleven who had Silence first heard *bad* (91%). Fisher's exact tests were performed on each 2x2 table of Condition by Response. The differences between the Music condition and each of the non-music conditions were significant (p=0.003 for Music vs. Silence; p=0.047 for Music vs. Noise), while the difference between Noise and Silence was not (p=0.33)[6]. Table 3.5 summarises this simple but striking illustration of how differing context-driven expectations can result in differing perception of the same auditory stimulus.

3.5.1.2 Outlier removal

Before moving on to an examination of the raw results of the dataset as a whole, the preregistered outlier removal procedure will be briefly summarised. Details of this

[5] The preregistration is available at `http://aspredicted.org/blind.php?x=qu8ze7` and is also included in Appendix B.

[6] Amongst these 36 trials, there are two which will be removed as outliers for having long reaction times. If those trials are removed from the analysis at this point, the difference between the Music and Silence conditions is still nearly significant (p=0.051), but the difference between Music and Noise is not (p=0.181)

Table 3.5: First token of entire experiment across three conditions for 36 participants.

Condition	n *bad*	n *bed*	Total Participants	Percent *bad*
Music	3	9	12	25%
Noise	9	4	13	69%
Silence	10	1	11	91%

process can be found in the preregistration document (included in Appendix B). No participants were excluded for having unusually high or low rates of perceiving *bad*, defined as having a mean response rate more than 3 standard deviations above or below the mean of participant means. The minimum for an individual was 33% *bad* responses and the maximum was 61%. No participants were removed for having particularly fast or slow mean reaction times — all participant means were within 3 sds of the mean of participant means. The fastest participant had an average RT of 548ms and the slowest average was 1187ms. Thus, all 36 participants remained in the analysis.

As for removal of individual trials, none were removed for having an onset delay[7] of more than 75ms (the maximum delay was 15ms). Individual trials where the reaction time was <3SDs below or >3SDs above the given participant mean were excluded. This resulted in the removal of 44 tokens which had long RTs and 1 token with a short RT. The number of trials included in all data analyses which follow is thus 2547.

3.5.1.3 Summary of raw results

Once outliers were removed, 46.3% of remaining responses were *bad*. This rate varied by condition, and supported the hypothesis, with 43.8% *bad* in Music, 47.0% in Noise, and 48.1% in Silence (see Table 3.6, below). Figure 3.4 shows the percentage of *bad* responses in each condition, for each of the six vowel qualities, in the raw data. It shows that steps 1, 2 and 6 were completely unambiguous, with greater than 99% agreement between responses to stimuli with these more extreme vowel qualities. There was also near-unanimous agreement that Step 5 was *bad* (97.0%). Step 3 was somewhat ambiguous, being heard as *bad* 17.7% of the time, while Step 4 was the most ambiguous (62.5% *bad*).

It is on these ambiguous stimuli, Steps 3 and 4, where the difference between conditions thus plays out. The Music condition (solid red line in Figure 3.4) attracted the lowest percentage of *bad* responses (13.7% for Step 3, and 51.4% for Step 4), the Noise condition (dotted green line) somewhat more (15.6% and 66.2%), while the Silence condition (dashed blue line) had the most *bad* responses (23.7% and 70.4%).

While it is interesting and important to get a feel for the raw data, a full analysis requires that we ascertain other factors that explain variation in the dataset, and hold those constant. In a case where the raw data appear to support the hypothesis, as they do here, it is especially important to ensure the result stands up to statistical

[7] E-Prime's reported latency between triggering a soundfile and it playing.

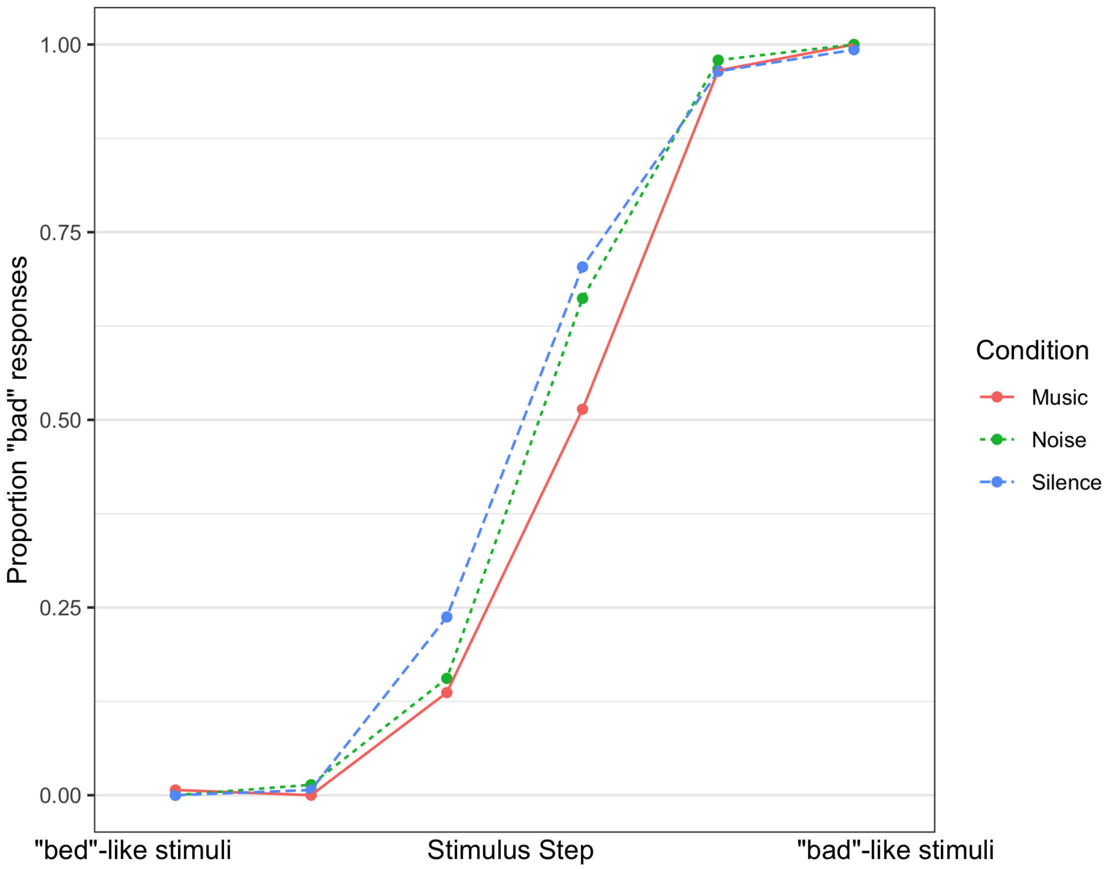


Figure 3.4: Proportion of *bad* responses to each of the six vowel qualities in the continuum, across the three conditions.

modelling. The statistical analysis begins in the next section with a very simple model of responses to just the most variable stimulus.

3.5.1.4 Simple GLMER model of responses to stimulus 4S

In the preregistration, there was a decision to run models on individual stimuli that showed a 'high degree of variability in responses', which I subsequently defined as being a response rate of between 40–60% *bad*. The only stimulus that was sufficiently variable to conduct such an analysis was the short version of continuum step 4 (4S), for which 51% of responses were *bad*. The next most variable stimuli were 4L (step 4, long) which had 74% *bad* and 3L, which had 23% *bad*. Note that half of the 4S stimuli occurred at the starts of blocks. The higher variation for this stimulus may thus be partly due to that privileged positioning — it occurred more often at points in the experiment when participants would be less certain about the range of the continuum. The present analysis includes the tokens discussed in Section 3.5.1.1 above, but extends to all instances of 4S, half of which were at the starts of blocks, and half of which were at a random position in the second half of each block, giving a total of six occurrences per participant. The raw difference between conditions for the 204 responses to stimulus 4S are as follows: 39% *bad* in Music; 55% in Noise, and 59% in Silence. Table 3.6 shows these values alongside the equivalent values for whole dataset, and the values for the first trial of the experiment discussed above.

Following a simple model fitting procedure which took into consideration the

Table 3.6: Percent *bad* responses in each condition: for the whole dataset; for all stimulus 4S trials; and for the subset of 4S trials that were the first token of the whole experiment (after outlier removal).

Dataset	Music	Noise	Silence	n Trials
All data	43.8%	47.0%	48.1%	n=2547
All 4S trials	39.4%	55.2%	59.1%	n=204
First 4S trial only	27.3%	69.2%	90.0%	n=34

small dataset, and thus did not test for interactions, a final generalised linear mixed-effects regression model (GLMER), using the *glmer* function in R's *lme4* package (Bates et al., 2015), was reached which included significant main effects for Condition and Gender, with a random intercept for Subject. Table 3.7 shows the output of the model. As compared to the Music condition (the reference level for the Condition variable), participants were significantly more likely to hear 4S as *bad* in the Silence condition (p=.012) and near significantly more likely to hear *bad* in Noise (p=.054). Additionally, males were significantly less likely to hear *bad* than females. Participants' age and the amount of time they had spent overseas were also tested, but were not significant.

Table 3.7: Output of simple model for stimulus 4S in phonetic categorisation task.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.152	0.34	-0.45	0.655
ConditionNoise	0.747	0.39	1.93	0.054
ConditionSilence	0.989	0.40	2.50	0.012
GenderM	-1.189	0.49	-2.42	0.016

3.5.1.5 Discussion of preliminary analyses

The results presented thus far show strong support for the Reference Frame Hypothesis, that participants will hear *bad* more in non-musical than musical contexts, when they are expecting to hear speech rather than singing. This can be seen in the raw data, and is particularly evident in the between-participants comparison of the very first token of the experiment. While this first-trial analysis is extremely simple, it is also very well controlled. The 4S stimulus token is ambiguous — it could be a good example of either a raised spoken NZE TRAP vowel, or an open sung DRESS vowel — and context is used by participants to resolve that ambiguity. In the absence of any knowledge about, or experience with, the speaker, at the beginning of the experiment, the contextual information recruited is at a more global level than once the participant becomes familiar with the speaker, and the range of variability involved in the stimuli becomes apparent. Recall that for those who had Noise or Silence first, they had no idea that the experiment would involve music at the point of this first response.

The analysis of responses to stimulus 4S, despite including tokens from throughout the experiment, also provided support for the hypothesis, though the difference between Music and Noise failed to reach significance. The significance of gender supports the prior finding by Drager (2006, 2011) that males have a slightly more open perceptual phoneme boundary between DRESS and TRAP, reflecting the reality in NZE, where females lead the short front vowel shift.

The fact that the result is stronger in the subsets of data discussed in this first part of the analysis, and more subtle when looking at the aggregated results (Table 3.6) is interesting. While it is in part a simple reflection of the ceiling/floor effects for the outer-most continuum steps, it also suggests that participants quickly update their frame of reference based on the incoming stimuli. As they progress through the experiment, they 'tune in' to the local context of the continuum itself and the broader musical context has less of an effect. As will be revealed in the final statistical model, below, this pattern is, in itself, statistically significant.

Two models of the full dataset will be presented in the next two sections. The first strictly follows the modelling procedure outlined in the preregistration, while the second follows a refined modelling procedure, correcting for some oversights in the preregistration.

3.5.2 PCT Model 1: Preregistered model fitting procedure

Fitting of statistical models is a process open to analyst bias at multiple steps. It is possible to convince oneself that a model with more 'desirable' results is worthy of presentation, while another is not (Simmons et al., 2011). Preregistering a specific model fitting procedure reduces the scope for this approach to statistics. Hypotheses made prior to data collection can be tested with deductive, rather than inductive reasoning — they are predicted, not post-dicted. The preregistration for the present experiment provided a clear and detailed roadmap for the process of statistical analysis. Without describing all the details of this in full here (the preregistration document appears in Appendix B), the four main steps of the backwards modelling procedure were as follows:

1. Initial model: modelling began with a predefined list of main effects, an intercept for participant, and no slopes.

2. Pruning: independent variables (IVs) were removed based on least significance, with each removal ratified by log-likelihood model comparison. Variables listed with * in the preregistration were tested in all 2-way interactions prior to removing, and kept in the model until after pruning if those interactions approached significance ($p<.1$).

3. Variance Inflation Factor (VIF) tests: VIF tests were carried out whenever new terms were added to the model (that is, at the start of modelling, and when adding interactions). If the highest VIF was greater than 15, variables were removed to reduce the multi-collinearity. Note that the preregistration did not set a final threshold for maximum VIF, and I ensured all VIFs were under 10 in final models. I report the highest VIF in each final model throughout the thesis.

4. Adding slopes: the pruning phase ended once a model was achieved where all IVs were either significant, or significantly worsened the model fit when removed. After this, a maximal slope structure was added, and slopes were

removed based on the amount of variance they explained, until the model converged.

Having worked through this process, exactly as preregistered, the following final model was fit: Response ∼ Condition + StimStep + StimLength + PrevStimStep + Gender + ListenMusic + Block * Trial24 + Musician * Trial24 + (1|Subject). Each of these significant predictors will be defined below, and the output of the model is shown in Table 3.8. In the model output, positive coefficients mean higher log-odds of responding *bad*. Each of the significant predictors is discussed in turn:

- Condition — Most importantly, the two 'speech' conditions (Noise, Silence) have significantly higher log-odds of a *bad* response than does the 'singing' condition (Music, the reference level), when all other variables are held constant. The difference between Noise and Silence was tested by running the same model with different level ordering for Condition, and this difference was not significant (p=.101), but the trend was for there to be more *bad* responses in Silence than Noise, as was seen in the raw data (see Figure 3.4).
- Continuum step of the stimulus (StimStep) — Unsurprisingly, more *bad*-like vowel qualities are more likely to be heard as *bad*.
- Length of stimulus (StimLength) — Short stimuli are less likely to be heard as *bad* than long stimuli, as predicted.
- Continuum step of the preceding stimulus (PrevStimStep) — The more *bad*-like the previous stimulus was, the less likely the present stimulus is to be heard as *bad*, as predicted. That is, there is a local contrast effect.[8]
- Gender — Males are less likely to respond *bad* than females. This is in the expected direction.
- Amount of music a participant listens to (ListenMusic) — People who listen to more music are less likely to respond *bad*. This variable was treated as continuous despite being an ordinal Likert scale response, see Table 3.3.

In addition to these main effects, two interactions were found:

- Block number in interaction with Trial number (Block*Trial24) — As the experiment goes along, participants are less likely to respond *bad*. This occurs as trials unfold within a given block, and across each of the three blocks, with the trial effect diminishing in the second and third blocks. Block is treated as a three-level factor, while trial is a continuous variable from 1–24, representing the serial position of a stimulus within a given block. This is an unexpected behaviour since the majority response across the whole dataset was *bed*. It was predicted that participants would try to even out their responses as the experiment went on, but they actually gravitate more strongly to *bed* over time.

[8]Note that for the first trial in each block, where there *was* no preceding stimulus, NA cells were avoided by entering a 'neutral' continuum step equivalent to the present stimulus. Since all blocks started with stimulus 4S, PrevStimStep was thus set to 4 for the first trial of each block.

- Whether participant is a musician, in interaction with Trial number (Musician*Trial24) — Musicians did not show the gradual trend towards responding *bed* as they progressed through each block in the way that non-musicians did.

Table 3.8: Output of model of all phonetic categorisation task data, using preregistered model fitting procedure (PCT Model 1).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.775	0.77	-10.08	<0.001
ConditionNoise	0.627	0.22	2.88	0.004
ConditionSilence	0.985	0.22	4.48	<0.001
StimStep	3.139	0.16	20.20	<0.001
StimLengthS	-1.012	0.18	-5.54	<0.001
PrevStimStep	-0.434	0.06	-7.43	<0.001
Block2	-0.393	0.40	-0.99	0.324
Block3	-1.123	0.41	-2.75	0.006
Trial24	-0.077	0.02	-3.52	<0.001
GenderM	-0.981	0.41	-2.40	0.016
Musiciany	-0.699	0.58	-1.22	0.224
ListenMusic	-0.334	0.16	-2.08	0.038
Block2:Trial24	0.025	0.03	0.85	0.393
Block3:Trial24	0.086	0.03	2.98	0.003
Trial24:Musiciany	0.056	0.03	2.01	0.044

3.5.2.1 Discussion of PCT Model 1

Expectation of singing, and the presence of background music, are related to expectation for more open pronunciations of DRESS and TRAP. This holds when comparing the Music condition to both of the control conditions. The other significant fixed effects are generally in the expected direction, with males having a more open phoneme boundary, and the previous stimulus demonstrating a significant local contrast effect. The interactions are more difficult to interpret, but their main function in the model is to hold constant as much variability in the data as possible, so that the hypothesis about the effect of Condition could be tested.

As argued at the start of this section, the benefit of conducting a fully preregistered analysis is a reduction in the role of analyst bias. However, many decisions made by the analyst in the course of model fitting are well-founded (that is, not all are driven by biases). The rigid approach taken to reach the model presented here has thus missed out on the ability to learn about the data and respond to that new knowledge in the process of working with it. For example, in the preregistration I did not consider the possibility that the effect of Condition would change as the experiment unfolded, a pattern which became clear upon examination of the raw data. Since an interaction between Condition and Trial was not preregistered, it could not be included in the model presented in this section. Additionally, the main effect regarding the amount of music a person listens to would be an interesting finding if it was carried by the Music condition, but since the interaction of music-related questionnaire items with Condition was not preregistered, it could not be tested in the present model. The interaction between Musician and Trial is dubious, given that there are only eight musicians amongst the 36 participants. With the ability to make active decisions during modelling, it is possible to avoid tests which may have data sparsity issues.

These oversights in the preregistration will be remedied in the next section, whilst risking the introduction of my own bias to find a model that is to my liking. I use the information gleaned from the first model fitting procedure to more carefully decide which interactions should be tested, and in what order items should be pruned from the model.

3.5.3 PCT Model 2: Refined model fitting procedure

Prior to commencing the model fitting for this second model, I considered what I had learned about the data in the process of fitting the preregistered model. I then decided which interactions and main effects I had clear predictions for, and restricted my modelling to those terms. Additionally, having learnt that modelling of this dataset was subject to convergence issues, I decided not to add all potential interactions for a given variable at once, but rather to test each one on a base model and note the log-likelihood comparison of the model with and without the given interaction. Only those that approached significance would then be added and pruned after the removal of other non-significant main effects.

The details of the model fitting procedure for PCT Model 2 are similar to the preregistered procedure outlined above. I outline below the variables and interactions which were tested, along with the prediction for each. The main effects tested, but not expected to be involved in interactions were Age (younger should be more likely to respond *bad* due to having more raised short front vowels); Time Overseas (people with more overseas experience should be less likely to respond *bad* having had more experience with unraised short front vowels); and Gender (males should be less likely to respond *bad*, having less raised vowels). The various interactions tested, and the prediction for each, is as follows:

- Trial number * Condition — We know from the preregistered model that there is a drift towards *bed* as the experiment goes on. What would be of more interest, however, is if there is a drift towards being less affected by Condition as the experiment goes on, with the prediction that the difference between the Music and Noise/Silence conditions will be greater at the start, and diminish as the experiment goes on. I attempted modelling this with two different representations of Trial number: 'Trial72' represents how far through the entire experiment the trial occurred, and 'Trial24' represents how far through the given block the trial occurred. It turns out that both of these interactions are significant, with the latter providing a slightly better model fit. This will be discussed below.

- Stimulus Length * Condition — Long stimuli may be more likely to elicit *bad* responses in the Noise and Silence conditions, but not in the Music condition (see Section 3.1.3.4).

- Musician * Condition — Musicians might be more shifted towards *bed* responses in the music condition than non-musicians.

- Music Listening * Condition — People who listen to more music might shift more towards responding *bed* in the music condition than those who listen to less music.

- Proportion of music listening devoted to USA versus NZ artists * Condition — Those who listen to more singers from the USA, or who have the greatest difference between their USA versus NZ music listening, might be more likely to respond *bed* in the Music condition.

- Surprising to hear a NZ accent in a song * Condition — Those who said they find it surprising to hear a NZ accent in a song may be more likely to hear *bed* in the music condition than those who say they would not be surprised.

- Experimental stimuli sound like singing * Condition — Those who were more convinced they were listening to singing may be more likely to hear *bed* in the Music condition.

- Genre Cluster * Condition — Those who listen to hip hop (Cluster 1 participants) might have been exposed to more NZ accents in music (see Chapter 2) and thus shift less towards *bed* in the Music condition.

After following the model fitting procedure to its conclusion, the following final model was chosen: Response ~ Condition * Trial24 + StimStep + StimLength + PrevStimStep + Gender + (1 |Subject). The highest variance inflation factor value for this model was 5.6. The model output is shown in Table 3.9.

Table 3.9: Output of model of all phonetic categorisation task data, using refined model fitting procedure (PCT Model 2).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.174	0.65	-15.66	<0.001
ConditionNoise	1.390	0.40	3.45	0.001
ConditionSilence	2.270	0.42	5.45	<0.001
Trial24	0.030	0.02	1.50	0.133
StimStep	3.123	0.15	20.32	<0.001
StimLengthS	-1.005	0.18	-5.50	<0.001
PrevStimStep	-0.425	0.06	-7.32	<0.001
GenderM	-0.932	0.43	-2.19	0.029
ConditionNoise:Trial24	-0.065	0.03	-2.25	0.024
ConditionSilence:Trial24	-0.107	0.03	-3.73	<0.001

Each significant term in this model is discussed below in turn, but it is worth noting first the *non-significance* of the wide range of music-related interactions tested. Amount and type of music listening did not affect how strongly participants supported the hypothesis. Models with the interaction of Genre Cluster with Condition faced convergence issues, but this interaction did appear to be significant, with Cluster 2 participants less likely to support the hypothesis (that is, less likely to shift towards responding *bed* in the Music condition)[9]. This lack of significant results will be discussed further below, but first, the significant effects:

- Condition * Trial24 — The first thing to take note of when looking at the model output is that the central hypothesis of the experiment was supported: participants respond *bad* more often in the non-music conditions, where they are told to expect speech, than the Music condition, where they are told to expect singing. This difference is highly significant for the distinction between both Music and Silence, and Music and Noise. The role of Condition in the model output cannot be understood, however, without also considering its significant interaction with trial number. Variables involved in interactions cannot be treated as if they stand alone as main effects. To summarise this interaction, then, the hypothesis (more *bad* responses in Silence/Noise than Music) is strongly supported early on in each block (Trial 1), but the difference collapses by the end of the block (Trial 24). This result is examined further below. To test for any difference between the Silence and Noise conditions, the same model was re-fit with a relevelled Condition variable that had Silence as the reference level. This revealed that the interaction between Condition and trial number was not significant when comparing Noise and Silence (Noise vs. Silence main effect: Estimate = -0.880, p=.033; Noise vs. Silence * Trial24: Estimate = 0.042, p=.147).

- Continuum step of the stimulus (StimStep) — More *bad*-like stimuli are more likely to be heard as *bad*.

- Length of the stimulus (StimLength) — Short stimuli are less likely to be heard as *bad*.

- Continuum step of the preceding stimulus (PrevStimStep) — The contrast effect is again significant.

- Gender — Males are less likely to respond *bad* than females.

Figure 3.5 shows the interaction of Condition by Trial described above. The Reference Frame Hypothesis was supported early on in each block, but after hearing *bed* or *bad* more than a dozen times in a block, the effects of expectations primed by the condition disappear. Recall that a similar model using the trial number throughout the whole experiment (Trial72) also interacts significantly with Condition, showing a trend for the effect of music on responses to diminish as the experiment goes along. Log-likelihood comparison of each of these models with a minimally smaller model, and comparison of AIC values, revealed the model with Trial24 (trial number within a given block) to be slightly stronger. Either way, the important trend is the same — the difference between conditions is strongest at the start of the experiment and diminishes both as each block goes on and as the experiment overall goes on.

In Figure 3.6, the related raw data is shown. In order to show raw data, the mean proportion of *bad* responses was calculated for each participant in each sub-block. Since all twelve stimuli were presented in the first and second half of each block, this is the smallest time-scale across which means of the raw data can be sensibly

[9]Cluster 2 participants include those who tend to like the Alternative, Rock and Singer-Songwriter genres. Note that this tendency was in the opposite direction than I had predicted.

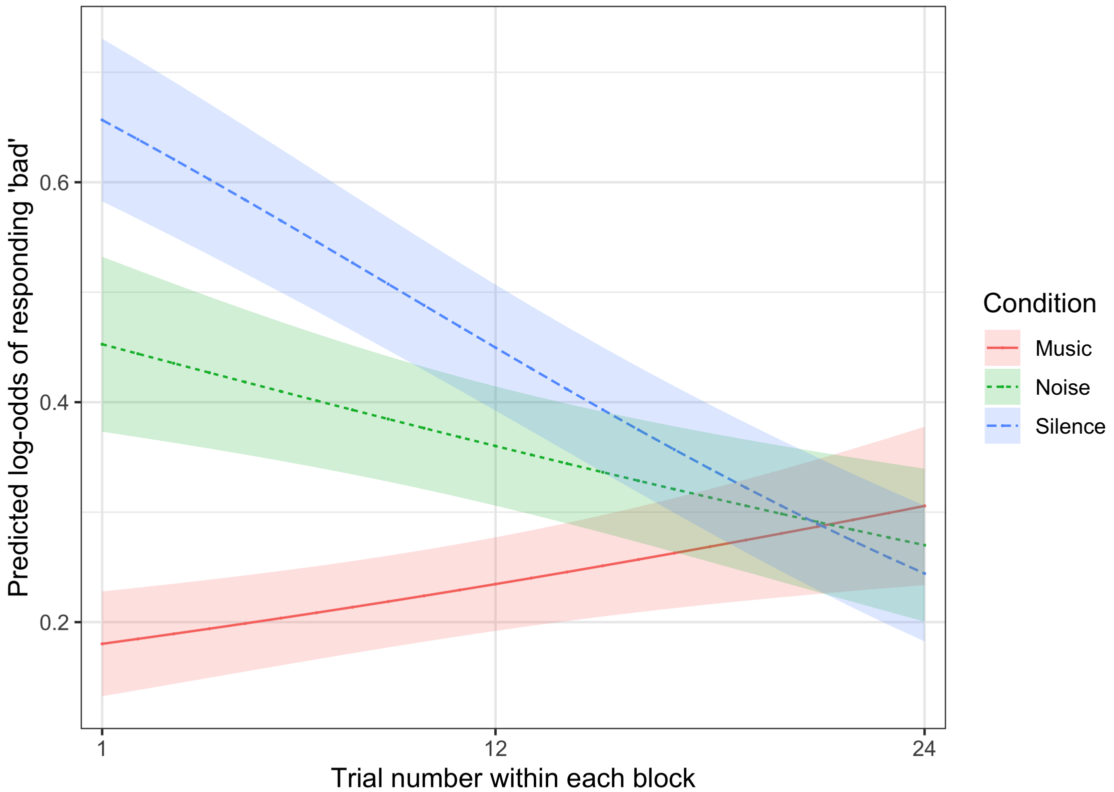


Figure 3.5: Predicted log-odds of responding *bad* in each of the conditions by trial number.

calculated (since stimulus quality is the main predictor of response, these need to be balanced before taking any mean of responses). Each data point represents the mean of participant means for the sub-block (for 12 trials across each of the 11–13 participants), and the error bars show the 95% confidence interval around each mean of participant means. While we can see the pattern captured in the model by the interaction, we can also see that this diminishing of the hypothesised effect is particularly pronounced in the first block, where the raw data shows a strong differentiation between the three conditions in these first 12 trials of the experiment. This shows how quickly participants 'tune in' to the continuum they are hearing, but it also shows that they return to relying more on the context at the start of each new block, with its slightly different intstructions, and different background audio.

3.6 Discussion

3.6.1 Summary of results

Participants heard *bad* more often in the 'speech' conditions than the 'singing' condition, providing strong support for the Reference Frame Hypothesis. This result is borne out across the range of data analyses presented above. It is clear in the raw results, and is particularly well demonstrated by a between-participants comparison of the first trial of the entire experiment, strengthening the findings of the original version of the experiment (Gibson, 2010b). Most importantly, the preregistered generalised linear mixed effects model on the entire dataset found a significant

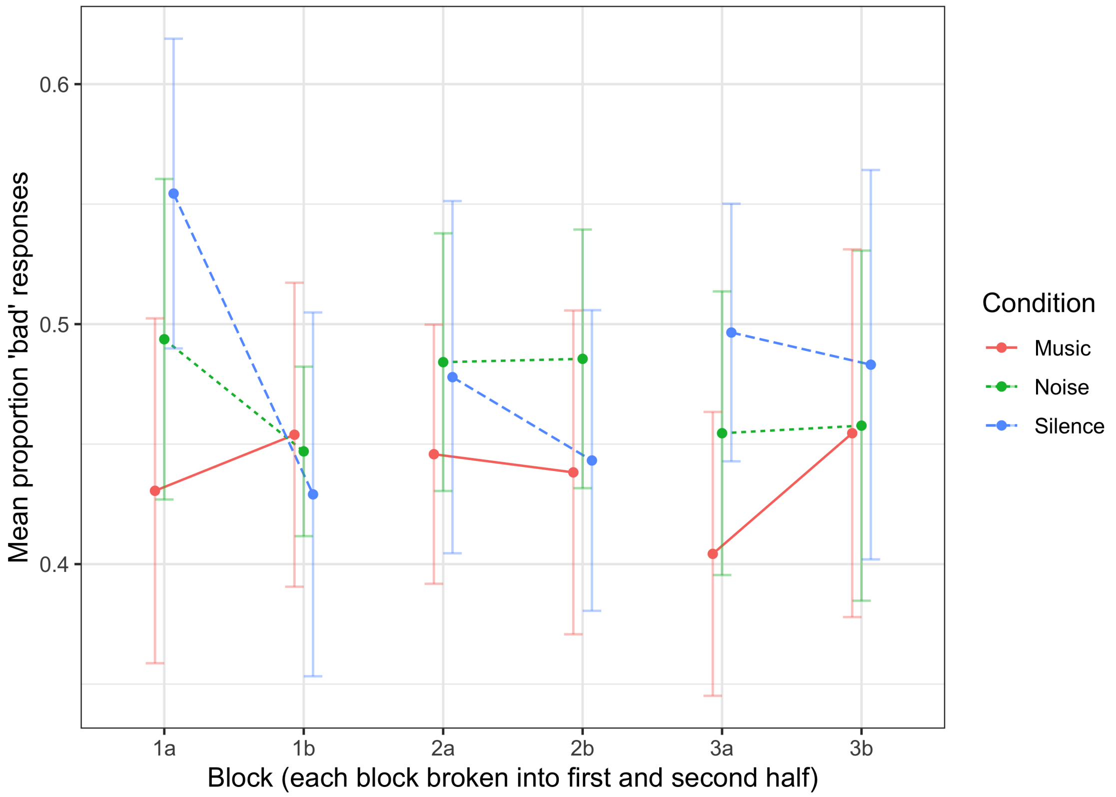


Figure 3.6: Proportion of *bad* responses (mean of subject means) in each of the six sub-blocks for each of the three conditions. Each sub-block contains one occurrence of each of the twelve stimuli. Error bars show 95% confidence intervals around the mean of participant means.

effect of Condition, with both of the 'speaking' conditions having higher rates of *bad* than the 'singing' condition, when holding constant the effects of other factors. This significant effect is also evident in the final model, which uncovered a nuance in the data: the hypothesised effect is strong at the start of the experiment and of each block, but diminishes as the participant becomes more familiar with each new combination of voice and context. This kind of habituation effect has been attested in other studies that involve surprisal (van den Brink et al., 2012; Nieuwland and Van Berkum, 2006).

The results were mixed with respect to the effect of background noise on phonetic categorisation. In the preregistered model, there was no significant difference between the two control conditions (Noise and Silence), strengthening the special status of music, and providing some evidence against the prediction that listeners might expect Lombard speech in the Noise condition. In PCT Model 2, however, once the interaction of Condition * Trial was introduced, a distinction between the Noise and Silence conditions did emerge as significant. Participants were more likely to respond *bad* in Silence than Noise, in the early stages of each block (see Blocks 1a and 3a in Figure 3.6, where Noise is roughly halfway between Music and Silence).

This provides some support for the finding in Hay et al. (2017), that there may be an expectation of more open vowels in a typically noisy environment. It does not provide clear evidence for or against their explanation, however, of the lack of a significant perceptual effect between conditions with and without actual auditory background noise. That explanation rested on the idea that listeners can tell the difference between words genuinely produced in noise, and those which were produced in silence but are being played back against noise. Further experimentation is needed to clarify these issues, but the results of this experiment do at least provide some evidence that NZ listeners may shift their perceptual phoneme boundary between DRESS and TRAP when listening to words in noise, in expectation of Lombard speech. While interesting, that trend does not in any way take away from the main finding, that music-related expectations exist, and their impact goes above and beyond any effect of noise. When expecting singing, NZ listeners expect more open DRESS and TRAP vowels.

A range of other factors were significant in the statistical models. Foremost amongst these is the obvious role of the vowel quality of the stimulus itself (more *bad* responses to more *bad*-like stimuli), but also the quality of the preceding stimulus. This is a contrast effect, whereby if the previous stimulus was more *bed*-like, the present response is more likely to be *bad* . This effect has not always been tested in the statistical models of previous studies of this kind, and should be considered if this methodology is employed again in future.

An important discovery in this experiment was the significant effect of manipulating the length of a stimulus. Despite identical formant frequencies, participants are more likely to categorise short stimuli as having DRESS vowels. This reflects their experience in the speech community, and is a result which calls into question the findings of Drager (2006). Drager (2011) is less problematic with regards to this issue, since photo-age and voice were crossed in the design. A reanalysis of the results of that experiment could potentially uncover what role length had to play, however, and maybe clarify some of the unexpected interactions involving participant age and gender.

Regarding the participants' social characteristics, there is a clear gender effect which shows up throughout the analyses: males have fewer *bad* responses, supporting prior studies, and suggesting that they themselves as a cohort have less raised DRESS and TRAP vowels than the female participants. Age was also expected to affect phonetic categorisation, but did not reach significance in any model. This may simply be due to the large majority of participants being in their twenties (see Table 3.3). Participants were not selected with the aim of testing such an effect, though it is likely that an age effect would emerge if a balanced sample of participants were recruited.

In the refined modelling procedure, a range of interactions between Condition and the participants' music exposure was tested, and none of these reached significance (though an interaction of musical genre preference with Condition held some promise). Perhaps counter-intuitively, this is actually in line with the overall argument of this thesis, that there is strong homogeneity in singing accents. Whether a person listens to more or less NZ or American music, for example, is a moot point if the realisation of DRESS and TRAP is similar in music from any location. The use of raised DRESS (and to a lesser extent TRAP) does not appear to be a marker of NZ identity even for 'own-accent' singers in NZ (see Section 3.2), so more exposure to NZ music should not necessarily break down the expectation for open vowels in song. In fact, if anything, those who have strong exposure to New Zealand singers using open variants in song may actually have stronger *backgrounding of place indexicalities* (a proposal that will be discussed in some detail in Chapter 5) than those who do not.

3.6.2 Methodological issues

There are obvious limitations to a task involving only two lexical items in a laboratory setting. A number of task-related strategies may come into play, and the effects of waning attention and 'zoning out' are unpredictable. As has been stated several times now, however, one of the strengths of this task is that it does not take a lot of trials to see the hypothesised effect — one trial is sufficient. In fact, the more trials there are, the more muddy the results become. This provides important lessons about experiment design, especially if we are to apply methodologies such as this to the collection of electrophysiological data (as is suggested we might, in Section 4.5). The importance of considering how participants learn over the course of an experiment is important. For example, Hay et al. (2018) found that American participants learnt about and adjusted to a novel form, intrusive /r/, over the course of an experiment. With respect to EEG, methods for single-trial analysis are rapidly improving (De Vos et al., 2012).

Beyond methodological issues, the change in responses over the course of the experiment tells us about the role of context in speech perception. When we know nothing about the voice we are about to hear, environmental cues will strongly shape expectation in the absence of more specific information. Within seconds of listening to a voice, cues within the voice itself, including phonetic patterns, will become evident and the general context will become less relevant to the core phonetic frame of reference. The context and the voice will interact at that point to determine the linguistic and social meaning of variants. Upon hearing the first Step 1 stimulus in the music condition of this experiment (a very raised *bed*), listeners may have a range of reactions. The stimulus may break any illusion of listening to a singing voice, being too out of place to fit a sung context. Or perhaps the supposed singer will be considered unusual or alternative for using a raised DRESS vowel in the context of a song. The first of these reactions deserves further attention.

As mentioned in the methods section above, Step 1 of the continuum is much more raised than any sung DRESS in the measurements of the PoPS corpus. It is possible that the impression of singing in the music condition would be destroyed upon encountering one of these tokens from the *bed* end of the continuum. Since they are not plausible tokens of sung DRESS, and the experiment insists to participants that the word is either *bed* or *bad*, their vowel perception may be switched back to a speech-like frame of reference upon encountering these tokens. This, along with the strong contrast effect which was found to be significant, would weaken the hypothesised context-induced perceptual vowel shift. The striking between-participants difference in responses to the first token of the experiment, along with the decreasing effect of condition as the experiment progresses, provides clear evidence that the effect of music is greatly weakened as the experiment goes on, and this provides another possible reason for that. To mitigate this issue, future experiments could restrict stimuli to the ambiguous space, rather than providing listeners with strong reference points at each end of the continuum in the way that this, and previous studies, have tended to do. Alternatively, the vowel continuum could span more than two phonemes, as in Fry et al. (1962).

A more technical potential limitation needs to be addressed, particularly given the intermediate results found for the Noise condition: could the effect of music be related to it having a greater amplitude than the background noise? Recall that the average amplitude of the background music is 4 dBSPL greater than that of the background noise. We see a slight trend for fewer *bad* responses to Noise than to Silence, and less again for Music. Could this result, then, be explained by Music being just 'more noisy'? I would argue that this cannot explain the results, since the same Noise–Music intensity ratio leads to more errors in Noise in the lexical decision task to be presented in the coming chapter (see Section 4.3.1.2 for details). The higher error rates in the LDT suggest that the noise was actually psycho-acoustically louder than the music. Additionally, the steep drop-off in the effect of Condition strongly suggests that the result is to do with participants' expectations about what they are likely to hear, rather than any effect of masking.

I conclude this chapter by presenting one final limitation, not of the methodology of the experiment, but about how much this experiment can tell us about the listeners' expectations. The distinction between DRESS and TRAP rests mainly on an F1 difference, and there is a general trend for singing to involve greater jaw opening (and thus a greater F1) than speech, in general. The next chapter includes the same group of participants, and uses a lexical decision task to explore their expectations specifically about dialectal norms in popular music.

Chapter 4

Lexical Decision Task

> If lexical identification involves matching an incoming signal to a distribution of stored exemplars of words in context, then identification should be facilitated when the incoming signal is well matched to past experience(s) of the word. (Hay, 2018, p. 7)

4.1 Congruence Facilitates Lexical Access

This chapter presents the results of an auditory lexical decision task that explores listeners' expectations about accent in singing versus speech. The experiment is based on Walker and Hay (2011), in which participants made lexical decisions to 'older-person words' (words said more by older than younger speakers in the Origins of New Zealand English, ONZE, corpora) and 'younger-person words' (words over-represented in younger people's speech) in an older and younger voice. They found faster and more accurate responses when the 'word age' matched the voice age, that is, when the word was congruent with the speaker. Rather than exploring expectations about speakers of different ages using different words, this experiment manipulates speaker dialect (using a NZ and a US voice), and background audio context (Music, Noise and Silence). It is expected that the NZ voice will be easier to process than the US voice overall, since the participants are New Zealanders and familiar dialects are processed more quickly than unfamiliar ones (Clopper et al., 2016; Floccia et al., 2006). The central hypothesis of this experiment, however, is that responses to the US voice will be facilitated (or responses to the NZ voice inhibited) in the music condition.

As with the PCT, this experiment was preregistered prior to the commencement of data collection, and that document is included for reference in Appendix C.[1] A summary of the preregistration and how it guides the analysis will be presented at the start of the results section below. First, a review of relevant literature is presented, beginning with the effect of congruence on lexical access, and continuing with a review of some key studies using the lexical decision task methodology, including the three experiments that provide the direct foundation for the present experiment.

Godden and Baddeley's 1975 seminal (and rather elaborate) experiment tested list learning and recall by participants on land and underwater. They found that

[1] The preregistration was submitted to aspredicted.org on 11 October 2018, and can also be viewed at `http://aspredicted.org/blind.php?x=hw4t4k`.

when the context of learning matched the context of retrieval, recall was improved. This provided evidence that memory encoding and recall are affected by 'extrinsic context'. From this situational learning effect flow a range of consequences. For example, we can detect when a stimulus occurs in an unusual context. The development of EEG methods has been particularly fruitful in the exploration of such congruence effects. In a now seminal study, Van Berkum et al. (2008) found that an N400 (an event-related brain potential signalling a contextually unlikely word) was elicited by words that were incongruent with stereotypical characteristics of the voice. For example, when a female mentions her *tractor*, stereotypical associations are revealed by the N400 response, signalling that the word *tractor* is incongruent with the female voice in the minds of the participants. The EEG literature on incongruence will be introduced in some detail in Section 4.5.

Recently, in what is to my knowledge the first study of sociolinguistic expectation in the processing of song lyrics, Squires (2018) found little support for the hypothesis that non-standard forms (particularly invariant *don't*, as in *he don't*) would cause less disruption to a self-paced reading task if they were in the context of song lyrics. The lack of clear support for a reduction in surprisal can be accounted for by considering the wider context of the experiment. By delivering lyrics one word at a time, for visual processing, with no auditory access to music, it is easy to imagine that the participants' experiences with music were not activated in a rich, multi-sensory way. As Squires (2018) argues, the attention to individual written words, if anything, would heighten the salience of standard language ideologies. The methods employed in the present chapter could be adapted to test morphosyntactic and lexical variables, with the hypothesis that response times to auditory presentation of non-standard forms would be facilitated in music contexts.

As was discussed in the previous chapter, effects of various speaker characteristics on speech perception have been attested. Most have focused on listeners' phonetic frame of reference for a speaker, rather than on speaker-indexical congruence per se (e.g. Strand 1999 on gender, Staum-Cassanto 2008 on ethnicity and Drager 2011 on age). An experiment by Walker and Hay (2011), however, brought this literature closer to the work on facilitation of lexical access through congruence. They provide a complement to the work of Van Berkum et al. (2008) by focusing on congruence of a certain type of speaker with frequency-based patterns rather than congruence of a speaker with semantics, using a lexical decision task. The Walker and Hay (2011) study will be introduced in further detail below.

In both Chapters 1 and 2, salience was discussed largely with reference to speech production. To extend upon this in the context of the perception experiment to be presented in this chapter, I briefly review here the work of Sumner and colleagues, who have shown that overt attention to social stereotypes affects the perception of fine phonetic detail. They argue that novelty and salience affect not just how we access information stored in episodic memory, but also modulate the strength with which episodes are encoded (Sumner et al., 2014). In a task where listeners transcribe Chinese-accented speech whilst being presented with a picture of either a Chinese or Caucasian face, McGowan (2015, p. 516) found that 'listeners transcribe Chinese-accented speech more accurately when the face they are shown provides social information congruent with the voice they are listening to. They transcribe speech less accurately when the social cues and acoustic signal are incongruent'. As will become clear, this study is highly relevant to the LDT presented in this chapter. McGowan's study provided counter-evidence to prior claims that would have predicted negative social stereotypes to lead to more transcription errors in the Chinese-face condition, through a decrease in attention to the voice. A critical detail was that the effect was significant even for listeners with relatively little experience with Chinese-accented English, suggesting that even relatively few encounters with the variety, bolstered by social stereotypes, can lead to strongly encoded (and detailed) memory traces.

4.1.1 Exploring lexical access with lexical decision tasks

The lexical decision task is one of the most widely used methods in psycholinguistics, the term having been initially coined by Meyer and Schvaneveldt (1971). It presents participants with words and non-words and requires them to judge which is which. Early examples of the LDT were generally conducted with visual stimuli. Goldinger (1996a) reviews the later development of the auditory LDT, and presents various methodological insights that guide both the methods and analysis of the present experiment. Of particular relevance to this study amongst early auditory LDTs are those which deal with lexical representation issues (Luce and Lyons, 1998; Luce and Pisoni, 1998) and those which address the integration of music and language (Poulin-Charronnat et al., 2005; Hoch et al., 2011). The LDT is often paired with a priming methodology. There have been several studies of cross-dialect and cross-language effects using this method (for a review, see Clopper and Walker, 2017).

The LDT has proven a useful tool in recent examinations of the role of congruence in speech processing. As discussed above, there is an accumulating body of evidence that lexical access is facilitated by various types of congruity. Walker and Hay (2011) found evidence that listeners have access to knowledge about how words have been distributed across social groups. In a lexical decision task, they found faster and more accurate responses to words where the word-age and the voice-age were congruent, e.g. the word *shilling* was responded to more quickly and accurately when heard in an older voice while *sexist* was accessed more easily in a young voice. This finding was replicated by Kim (2016).

An important distinction between Walker and Hay (2011) and Kim (2016) is their methods for selecting words. Walker and Hay's words are selected from a corpus and based on relatively slight skews in word-age toward younger or older speakers. Kim's words, on the other hand, were actively selected to represent stereotypically young and old speech, mainly using neologisms for the young words, with only 14 out of 96 of the young words appearing in a standard dictionary. While this will be highly correlated with distributions of usage, there will be an additional 'layer' of conscious awareness associated with the stereotypical forms that is not present for the items in Walker and Hay's experiment, where the distinction between old and young words is, in most cases, less intuitive than the examples given above. Despite these differences, Kim (2016) replicated the finding that response times are faster when word age is congruent with voice.

In a third experiment, Kim and Drager (2017) found an effect of word age on speed of lexical access in a primed LDT where only a single sociophonetic variable was manipulated between conditions, and where that manipulation occurred on the prime, not on the target word. When the prime included a younger-person variable, participants responded faster to 'young' target words (as rated in a separate study) than 'old' target words, and vice versa, though the (young) participants also responded to young words faster in general. An additional result is particularly relevant to this study: there was a significant interaction between test location (Seoul or Hawai'i) and word age, such that participants tested in Korea were faster to old-words than those tested in Hawai'i. The authors interpret this as follows: 'Being tested in a foreign country could lead to [a] difference in activation level or prior expectation for the probability of encountering an old-associated word' (Kim and Drager, 2017, p. 624).

This paradigm was recently extended by combining implicit association tasks with lexical decision tasks, and using not only age-skewed words as above, but also gender-skewed words. Across four experiments, Hay et al. (2019) found that lexical access to words over-represented in male or female speech is facilitated by congruent social cues, be they gendered faces, the labels 'male' and 'female', or unlabelled pictures of gendered pairs of objects such as a briefcase and a purse. There were significant reaction time effects in all tasks, while the hypotheses were not consistently supported by the accuracy data, perhaps due to ceiling effects.

In the present study, the independent variable expected to speed lexical access is the congruence of a US voice with a pop music context. Rather than a continuum of stimuli differing on just the F1/F2 dimensions, the stimuli used in this task employ the full suite of sociophonetic variants occurring in the US or NZ voice that produces the target word (there are no primes) and the presence or absence of background music.

4.1.2 Summary of predictions

Before moving on to describing the methods of the experiment in detail, I will summarise the main prediction once more. Listeners have strong associations between SPMSS and pop music contexts. If our ability to parse the speech stream relies on activation of contextually relevant subsets of language exposure, then we should be more able to process SPMSS sound structures in the context where they are highly likely, and we have heard them most frequently. In the present experiment, this ease of processing would manifest as a facilitation for processing the US voice in the music condition. In the statistical models, this would result in an interaction between Condition and Voice, such that in the music condition, the US voice is processed faster and/or more accurately. This leads us to the specific hypothesis of this experiment:

> Lexical Access Hypothesis: New Zealanders' lexical access will be faster (more native-like) for a US voice when it occurs in song than non-song contexts. Responses will be slower (less native-like) to the NZ voice when it occurs in music.

4.2 Method

In this lexical decision task, 36 participants listened to 300 stimuli across six conditions that cross two voices (NZ and US) with three contexts (Music, Noise and Silence). Half of the stimuli were real words, chosen for having a sizeable phonetic distinction between the two dialects, while the other half are non-words that differ from the real words by one or two phonemes.

The methods of two of the experiments described above, upon which this study builds (Walker and Hay, 2011; Kim, 2016), were considered carefully when designing the present experiment. There are fewer trials in the present experiment than in the previous ones: 300 per participant, compared to 840 in Walker and Hay (420 items, each heard in two voices) and 768 in Kim (two lists, counterbalanced by voice age).

Both previous experiments have 2*3 experimental conditions, consisting of a two-way speaker voice distinction for age, a younger and older voice (though in Kim, 2016, there are actually two older and two younger speakers, a male and female, though this distinction is never analysed), and three experimental manipulations: old, neutral and young words. The present experiment has two voices, a US and a NZ speaker. These voices were matched for age and gender, and will be further described below. Background audio and instructions were manipulated (instead of word-age) with the same three conditions as in the PCT: Music, Noise and Silence. The prior experiments proceeded in blocks for each voice, with the old, neutral and young words randomly mixed together (that is, blocked by voice, randomised by word-age). Similarly, the current experiment was blocked by voice — participants heard all three conditions first in one voice and then the other. In order to create the impression of singing in the Music condition, it was necessary to separate the three conditions into their own sub-blocks so that the Music condition could have some continuity, and give the impression of a song. It would be of interest to see whether tiny snippets of music could achieve the same result. This first baseline experiment is being conducted, however, to establish whether music can influence speech perception of natural voice recordings at all.

4.2.1 Participants

The thirty-six participants are the same as those who completed the phonetic categorisation task, and who were described in Section 3.3.

4.2.2 Experimental stimuli

This section covers the selection and recording of words and nonwords for the experiment, as well as analysis and manipulation of those recordings prior to running the experiment. The background music and noise are also described.

4.2.2.1 Creation of word and nonword lists

Words and nonwords for the task focused on several variables that vary greatly between typical singing accents and NZE. The words were inspired by the lists provided in Wells' 1982 description of lexical sets. The nonwords were derived from those words by changing/adding one or two phonemes.

The stimuli, as finally included in the experiment, included 150 words and 150 nonwords, with 31 word/nonword pairs for each of three non-rhotic sociolinguistic variables of interest: BATH/DANCE, GOAT and LOT; and a further 57 word/nonword pairs involving rhoticity, specifically comprising START (21 pairs), NORTH (19 pairs) and NURSE (17 pairs). A longer list of words and nonwords was initially recorded to allow for some flexibility in choosing which words would be used in the final experiment design. Indeed, several words were removed in cases where the pronunciation was unclear, or the recorded pitch was deemed anomalous or unstable. Others were removed because it became apparent that the nonword could actually be interpreted as a word if heard from the perspective of the other dialect. Whenever a token was removed, its equivalent in the other voice was also removed so that the lists for the NZ and US voices remained identical. The full list of words and nonwords used in the experiment is included in Appendix C.

4.2.2.2 Recording of stimuli

Two male speakers in their forties were recorded reading the lists of stimuli, each in their own session. One was born and raised in New Zealand and the other was originally from Illinois, USA, and had lived in New Zealand for almost three years at the time of recording, in April 2016. The speakers were both colleagues in the Department of Linguistics at the University of Canterbury. Recordings were made in a quiet room at the NZ Institute of Language, Brain and Behaviour using a head-mounted Beyer condenser microphone. Audio was recorded in 16 bit, with a sample rate of 44.1kHz, into Logic Pro X on a Mac Book Pro, through a USBPre 2 interface. The speakers' participation was approved by the University of Canterbury Human Ethics Committee under application number HEC 2016/20/LR-PS Amendment 2. The information sheet and consent forms for the experiment are shown in Appendix C.

After being given a general description of the experiment design, they listened to a sample of the music track (described below) in headphones. This was to give them a sense of the tempo and tonality of the piece so that the two syllables of each word would occur on two consecutive eighth-notes when occurring in the music condition. In collaboration with myself, the first speaker to be recorded (the American) chose a stable pitch on which to say the words. This pitch was C#3 (which should equate to about 138.6Hz), which felt comfortable for the speaker, and was musically coherent, being the root of the C#m chord, which occurred regularly throughout the music. Once this training phase was complete for each speaker, they were asked to read through the words, and were asked to speak in a monotone on the target pitch, with no intonation contour in either direction between syllables.

The New Zealander was recorded second. He was played some of the American's recordings and was asked to match the pitch and rhythm of those recordings, while maintaining his 'normal' NZE accent. Note that he did not *shadow* the American speaker's recording, but simply listened to a few words to establish the desired pitch and word length. It is possible that hearing the US speaker could have caused accommodation effects in his speech production, though this is not my impression. The recordings produced a formal, word-list style, evidenced e.g. by a lack of /t/-flapping in the NZ voice. This relatively formal style seemed acceptable in the context of an ongoing list. The key criterion was that the words and nonwords could sound spoken when in isolation, and could sound sung when in the context of music. The 'sungness' of the stimuli will be further examined below (in Section 4.3.2.6).

The speakers were given regular breaks, to discuss desired pronunciation of the nonwords, and to re-establish the target pitch. During the recording of the second speaker, the words were checked to have identical interpretation of word stress and broad phoneme classes, except where phonemic distinctions were dialectal. For example, the nonword *choathy* was pronounced with a voiced dental fricative by the first speaker, so the second speaker was asked to also use /ð/. As the speakers worked their way through the lists, I marked on a hard copy any words that needed re-recording, and then had the speakers repeat those problematic tokens. A second recording session was undertaken with the NZ speaker for some of the first lists recorded because his pitch was later discovered to be lower on those stimuli. As will be described below, some problems with the recordings required further attention.

4.2.2.3 Analysis and manipulation of recordings of words and nonwords

Individual soundfiles were automatically extracted for each stimulus, using a Praat script by Vica Papp which applies the 'Annotate To TextGrid (silences)' function to find the alternating patterns of speech and silence in a soundfile and then exports each spoken word to its own wav file. A subset of stimuli were manually checked to ensure there was no silence at the start of soundfiles, since it is the start of the soundfile that triggers the start of reaction time measurement in E-Prime. The Praat script performed well at marking the boundary at the optimal place. Another script was used to automatically rename the resulting soundfiles using hyphens to concatenate three elements: voice (nz/us); word/nonword (w/n); and the stimulus name. This yielded file names such as 'us-n-snurchen.wav' and 'nz-w-arming.wav'. A script was then used to normalize all of the resulting files in Praat. This meant that the waveforms of all files had an equivalent waveform peak of 0dBFS, but they did still have varying mean amplitudes. To resolve this I scaled the amplitude of all of the files based on their root mean square amplitude, using the Multiply function in Praat. The script which did this created a ratio for each soundfile according to its current RMS, and then multiplied it to achieve an RMS of 0.105 Pascal (74.4dB SPL).

An analysis of the duration of the files was conducted, and is shown in Table 4.1. The NZ files ranged from 573–1027ms, with an average of 767ms (762ms for words and 773ms for nonwords). The US files ranged from 514–975ms, with an average of 732ms (719ms for words and 746ms for nonwords). These differences were deemed small enough for the purposes of the experiment. Note that an overall speeded response to one of the voices would not give a false positive result since it is the *interaction* of Voice with Condition that is the focus of analysis, rather than any main effect relating to Voice on its own.

Table 4.1: Analysis of mean duration (ms) and mean pitch (Hz) of soundfiles (prior to manipulation) for lexical decision task.

	Duration (ms)		Pitch (Hz)	
Voice	Word	Nonword	Word	Nonword
NZ	762	773	134.5	134.2
US	719	746	137.6	137.3

To analyse the pitch of the resulting files, a Praat script was used to write the pitch contour of each file and then take the mean f0 and its standard deviation throughout the pitch track. Recordings where the standard deviation of the pitch measurements across the soundfile was over 15Hz were deemed to involve errors in Praat's pitch tracking. For example, sometimes Praat attempts to track the pitch of a fricative (creating spuriously high values). These words with high standard deviations were removed from this analysis of pitch. Forty-six such tokens were removed, leaving 601 files. Note that this checking process was done before stimuli were removed for other reasons, and began with 647 files. The average pitch of the remaining files was 135.9Hz. As shown in Table 4.1, the average pitch of these 601 files was similar between words and nonwords, however there was a difference by speaker, with the average pitch being 137.6Hz for the American voice and 134.5Hz for the NZ voice. This was deemed problematic since it meant that the US voice would be more 'in tune' with the music than the NZ voice. If this were the case, and the hypothesised effect was found, that effect could be due to either dialect itself or due to a more contextually congruent pitch. Some evidence for this suspicion comes from the finding by Gordon et al. (2011) that participants performed better in a LDT when musical beats were aligned with stressed syllables, even though that finding related to rhythm not pitch, which may have allowed participants to predict the onset time of stimuli.

In musical terms, the average pitch of the NZ voice is 68 cents (hundredths of a semitone) below the intended pitch of C#3, and for the US voice it is 20 cents below C#3. The NZ stimuli are thus on average a third of a semitone flatter (lower) than the US stimuli, and are actually closer to the pitch C3, which is not congruent with the tonality of the music. To minimise manipulating the pitch of the stimuli, it was decided that the music track itself should be tuned down by 41 cents so that it would be midway between the US and NZ average pitches, with them both then being nearer to the ideal relative pitch. Further impressionistic analysis of the pitch of the stimuli suggested that there was also more pitch movement between the first and second syllables of the stimuli in the NZ voice than the US voice. This could also make the voice seem less plausibly 'sung' in the music condition. Any discrepancies in 'singiness'[2] between conditions is a major problem. It is unknown whether the hypothesised effect depends on participants hearing the speech in music as if it were sung. Studies such as Hay et al. (2006a), where perception shifted towards Australian vowels despite participants 'knowing' that the speaker was a New Zealander, suggest that it may not be too much of a problem. However, since it is likely that the effect would at least be stronger if the American voice was perceived to be more 'singy' than the other, it was deemed important to minimise differences in 'singiness' between the two voices as much as possible.

To test whether the US voice did indeed sound more singy than the NZ voice, I conducted a small study with six participants (of varying language and demographic backgrounds) in the Department of Linguistics at the University of Canterbury. Each listener was played a total of 40 of the stimuli in the context of music, using E-Prime to collect responses. This small study used a version of the music which

[2] I introduce a distinction between two adjectives here. 'Singiness' refers to vocal delivery, how much does a voice sound like it is singing as opposed to speaking at a given time. 'Songiness', on the other hand, is a measure of relative lexical frequency. A 'songy' word is one which is more likely to occur in song lyrics than in conversational speech. In the LDT modelling, songiness of each of the 150 real word stimuli was calculated and used as an independent measure, whilst there was also a survey question asking 'how much did the voice sound like it was singing', resulting in an empirical measure of perceived singiness.

had been pitch-shifted down by 41 cents to be in between the average pitches of the two speakers as described above. Participants rated each stimulus on a scale from 1 to 5, labelled with the following categories: 1 = 'very much spoken'; 2 = 'more spoken than sung'; 3 = 'neither spoken nor sung'; 4 = 'more sung than spoken'; and 5 = 'very much sung'. The stimuli included half NZ and half US words and nonwords. Half of the NZ tokens were picked as having pitch movement between the syllables ('bad' recordings) and half were picked on the basis of having a stable pitch ('good' recordings). The results of this small study showed that there was no significant difference in ratings between the 'good' NZ recordings and the US recordings (which all had stable pitch), but that 'bad' NZ tokens were rated as significantly more speech-like, particularly for words (as opposed to nonwords). This confirmed my suspicion that these tokens could cause problematic differences in 'singiness' between the voices.

Another Praat script was developed in order to search through the 648 soundfiles and determine the difference between the mean pitch in the first and second half of each soundfile, enabling the identification of potentially problematic tokens. Stimuli with a particularly large pitch difference between syllables were discarded, but there were also many subtly problematic tokens where there was a small shift in pitch between the first and second syllables. The decision was made to use these stimuli, and to adjust the pitch of all the stimuli so they would be in tune with the music. To do this, a Praat script was developed to adjust the mean pitch of all stimuli to match the musical target (41 cents below C#3). Rather than using a 'robotise' type function which imposes an f0 upon the recording, the boundary between the syllables was located, and then the average pitch in each syllable was taken and the f0 was shifted to bring it to the desired pitch. Early tests of this script showed that these minor adjustments resulted in good audio quality with no artefacts. However, it was later found that some words with intervocalic sonorants had an audible artefact on the consonant. While not ideal, these artefacts were deemed to be an acceptable trade-off for reliably 'in tune' stimuli.

4.2.2.4 Music and noise stimuli

As with the PCT, the test stimuli are presented in three conditions which vary according to background audio: Silence (no background audio), Music or Noise. The background noise is the same as that used in the PCT (see Section 3.4.2.3), simply looped to last the length of the block (3 minutes and 10 seconds), with a fade in and fade out at each end. The music used in the LDT is 'Science Music' which was composed and produced by Ryan Podlubny and Rob Batke for Podlubny's 2019 PhD thesis work. The piece is in an electronic pop style with a 4/4 time signature. It was in the key of E major with a tempo of 130bpm in its original form. As discussed above, however, in order to make the mean pitch of the two voices equidistant from the intended pitch without altering the stimuli themselves to differing extents, the entire music file was pitch-shifted down by 41 cents (nearly half a semitone) in Logic Pro X, with its native 'Pitch Shifter' plugin. While it is possible to manipulate pitch and tempo independently, this results in much poorer audio quality. The quality of the audio is unaffected, however, if the pitch and duration remain coupled in the manipulation (the analogy is playing a tape at a slower speed). The pitch shifting thus resulted in a slight decrease in tempo. The new tempo was 127bpm.

As mentioned in Section 3.4.2.3, the LDT was used as a tool to determine the amplitude ratio between the background music and background noise so that they would have a similar perceptual loudness. To determine this ratio, the goal was to have the audio recordings set at levels that would produce a similar rate of errors in the Noise as in Music conditions. The idea behind this is that the noise control is only a successful control condition if it is at least as psycho-acoustically loud as the music is, but the noise has stronger masking properties than the music and thus can have a lower amplitude than the music and still achieve this.

The first pilot participants were exposed to music and noise with the same root mean square amplitude. This led to much higher error rates in the Noise condition than the Music condition. This 'piloting' was actually a false start at running the experiment, with fifteen participants, in which an error in the E-Prime scripting meant the stimuli were not presented in time with the music. This was a 'clock.scale' command in the in-line code which was inserted to allow synchronising of the E-prime paradigm with the recording of EEG data, as discussed in Section 4.5. Since the timing of stimuli is a crucial part of the design, these data were discarded. The large discrepancy in accuracy rates between conditions did, however, provide the useful insight that the relative amplitude of noise and music audio files needed to be adjusted. Three further pilot participants helped to determine a good ratio for the background audio files. It was found that there was a clear tendency for more errors in music than noise when the noise was 6dB SPL quieter than the music. The intermediate version of the experiment, which was deemed to have an appropriate ratio of amplitudes, had the noise 4dB SPL quieter than the music. This produced slightly more errors in the noise condition than the music condition, as desired. This ensures the noise is loud enough to cause at least as much difficulty with the task as the music condition, but without being perceived to be much louder. After trialling these various versions, the final average amplitude of the noise file was .0285 Pascal (63db SPL), while the music file had a root mean square amplitude of 0.045 Pascal (67dB SPL, 4dB louder than the noise). The word and nonword stimuli had the same amplitude as the *bed* and *bad* stimuli had in the PCT (0.105 Pascal, 74.4dB SPL), but in this experiment, since it is a much more difficult task, both background audio files were quieter than they were in the PCT (see Section 3.4.2.3 for further explanation).

4.2.3 Procedure

The 36 participants began the lexical decision task after having completed the phonetic categorisation task. They remained in the same sound-treated booth with the same equipment. After the PCT had been completed, I went into the booth and encouraged the participant for having completed the first task. On the response box, the *bed* and *bad* labels had been attached with Blu Tack on top of slightly smaller labels which said *Word* and *Not Word*. I said to participants, 'in the next task you'll be deciding whether you hear real words or made up words', and I removed the *bed* and *bad* labels to reveal the label for each of the options as I said it. Participants were then told that the second task would be a bit longer, with six blocks of three minutes, and reminded that instructions would be displayed on the screen. I then returned to the control room to load the second experiment in E-Prime.

All participants completed six blocks, encompassing the three conditions in each of the two voices. Order of conditions was counterbalanced so that half of the participants heard all three blocks of the NZ voice then all three blocks of the US voice, while that order was reversed for the remaining participants. The order of the three conditions was the same within each voice for a given participant, but the six possible permutations of that order were counterbalanced across participants. When referring to the blocks in the statistical analysis below, they are broken up into two variables: BlockHalf refers to the first and second voice, that is Blocks 1–3 vs. Blocks 4–6; SubBlock refers to the three blocks within each voice, that is, SubBlock 1 refers to Blocks 1 and 4, while SubBlock 2 groups Blocks 2 and 5, and so on. The models also work with the six-level version of Block, which is referred to as Block6. Many aspects of the procedure are similar to those used in the PCT. Responses and their reaction times are collected using the SRBox. The right-most button was always used for ‘word’ responses and the left-most button for ‘nonword’ responses. While most LDTs have participants use their dominant hand for the ‘word’ response, this was not done in the present experiment. This was a decision made to simplify processes and reduce the chance of any error in the administration of the experiment, particularly since I affixed different physical labels to the response box for the PCT and LDT.

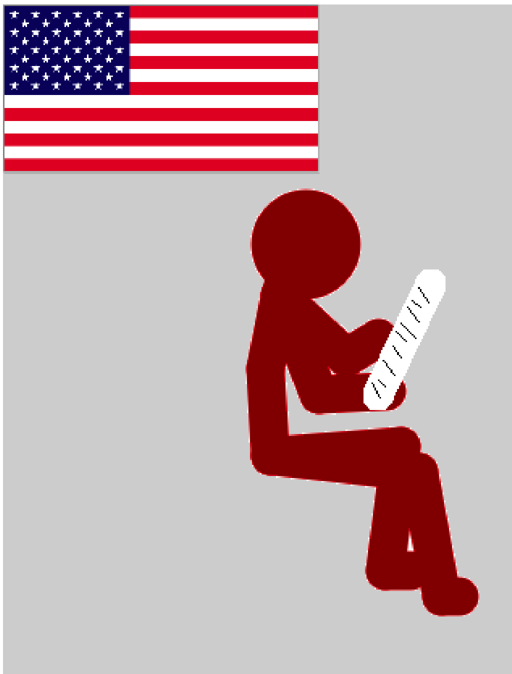
The instructions were slightly different to the PCT in that participants were told they would hear an American or a New Zealander speaking/singing, and after they pressed a button to begin the experiment, rather than going directly to the start of the experiment, one further screen was displayed, which included either the word ‘SINGING’ or ‘SPEAKING’, above one of the images shown in Figure 4.1. Multiple cues were thus given to prime the idea that the voices would be singing in the Music condition: first, the instruction that ‘you will hear singing’; then an image of a stick figure wearing headphones and holding a microphone; then the music itself; and finally, the fact that the voices are in tune and in time with the music and have a stable pitch.

After the task was completed, the participants completed a survey (see Section 3.3) that included demographic questions and questions about their music consumption practices. Finally, participants were thanked for their time, invited to ask any questions they had about the experiments, and remunerated with a $10 gift voucher.

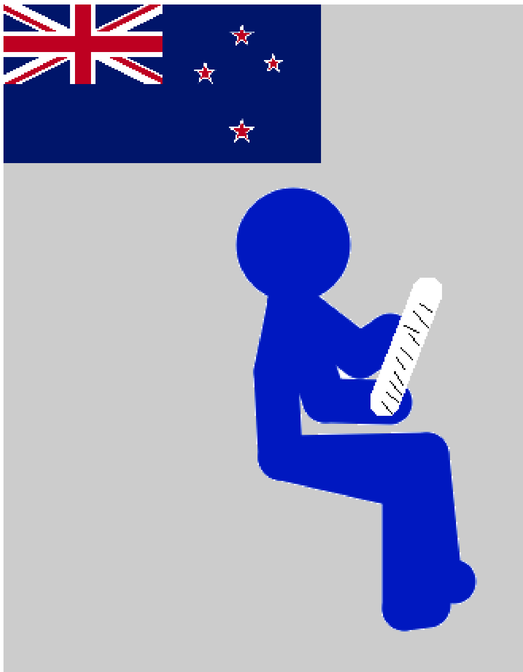
4.2.3.1 Experiment design in E-Prime

The E-Prime (version 3.0.3.60) experiment used nested tables to fully counterbalance the presentation of stimuli across conditions, randomly assigning the NZ and US voice to half of the stimuli every time the experiment is run. In this way, each participant hears half of the stimuli in each voice, and the assignment of voices to stimuli is different for everyone. In each block, half of the stimuli are words and half are nonwords. The distribution of the various linguistic variables was fully random.

As mentioned above, the music had a tempo of 127 beats per minute. At this tempo, one beat lasts 472ms, and one bar lasts 1890ms. These timings were used in the E-Prime settings to ensure that stimuli were played in time with the music using the same procedure applied to the PCT. The ‘cumulative’ timing setting was used for the stimuli to restore the rhythm on subsequent trials if an Onset Delay occurred. Crucially, however, ‘event’ timing mode was used when the background music was triggered, so that any onset delay would be passed on to the triggering of the first stimuli. That is, the stimuli would be delayed an equal amount as the background music, maintaining their relative timing. At the beginning of a given

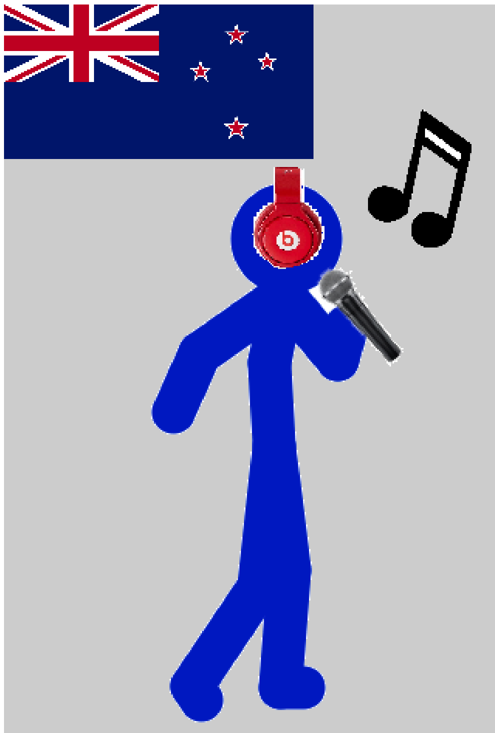


(a) 'An American speaking'



(b) 'A New Zealander speaking'

(c) 'An American singing'



(d) 'A New Zealander singing'

Figure 4.1: Pictures shown on the screen prior to starting each block, with either the word 'SPEECH' (a and b) or 'SINGING' (c and d) shown above them in large font.

trial, there is a Wait object of 1520ms, which allows E-Prime to buffer the soundfile so that it can be played on the musical beat. Following this wait, the stimulus is played and the participant has 2652ms to respond. An additional Wait object of 1ms inserted before the end of the trial was found to greatly reduce onset delays.

In the noise and silence conditions, there is a 4 second delay from the end of the screen with the image and the first stimulus. This allows for the fade in time of the noise in the noise condition, and seems a reasonable time to get ready for the first stimulus in the silence condition. In the music condition, there is a longer delay before the first stimulus, during which the four bars of instrumental music play.

4.3 Results

This results section begins with a description of the procedure for removing outliers, followed by presentation of the raw data for accuracy rates and reaction times for the interaction of Voice and Condition. This is followed by a series of statistical analyses, beginning with simplistic models, followed by a detailed description of the three preregistered models for accuracy and reaction time. Problems with the preregistration will be discussed, leading to a final model for reaction time based on a refined modelling procedure.

4.3.1 Raw results and data processing

Several sources of data were merged together: the E-Prime data; the questionnaire responses for each participant, and information about which genre cluster they belong to; the duration of each of the soundfiles; and frequency data for the 150 real word stimuli. Variables were then transformed, renamed, regrouped and relevelled, checking the distributions of continuous variables and logging and scaling several of them. As well as these various data processing tasks, outlier removal was also done at this first stage.

4.3.1.1 Outlier removal

As per the preregistration (see Appendix C), outliers were removed as follows:

- Mean accuracy rates were calculated for each participant, and the mean of these means was calculated to be 89.2%, with a standard deviation of 3.3%. The designated cut-off for removing data from whole participants on the basis of low accuracy was therefore 79.3% (mean of participant means – 3*sd of participant means). No speaker had a mean accuracy rate lower than this value (the lowest was 80.3%). The highest mean accuracy rate was 94%. No upper limit was placed on accuracy levels in the preregistration and thus all 36 participants remained in the analysis.

- All trials had an onset delay of zero thanks to the range of steps included to allow buffering, as described above. Therefore, no trials were removed for this reason

- Fifty-five trials were removed for having an RT of less than 400ms.

- The mean and standard deviation of the RT was calculated for each participant. An upper cut-off RT was then established for each participant at 3 standard deviations above their mean. A total of 174 trials were removed for being above that threshold.

- For the analysis of RT, 1058 trials with incorrect responses were removed, and finally 4757 nonword trials were removed, leaving a total of 4756 eligible correct responses to word stimuli for analysis, from a maximum possible total of 5400 trials.

- For the analysis of accuracy presented in the next section, the incorrect responses are kept in the dataset.

4.3.1.2 Raw results: Accuracy

In the preregistration, it was decided that analysis of accuracy should focus only on responses to real words. Before removing the nonwords, though, a brief inspection of the accuracy rates across the full dataset is presented here.

The raw accuracy rate across all 10,571 responses (after removing 229 outliers based on RT) was 89.9%. It was lowest in the Noise condition (87.2%), moderate in the Music condition (90.6%) and highest when stimuli were heard in Silence (92.2%). When looking at these responses to both words and nonwords, accuracy was higher for the American voice (91.0%) than for the NZ voice (89.0%). The cross-tabulation of voice by condition reveals no clear support for the Lexical Access Hypothesis — Music and Noise pattern together with similar advantage for the US voice over the NZ voice (2.4% and 2.3%, respectively). This advantage is less when the stimuli were heard in silence (US accuracy 1.2% higher). Listeners' accuracy was hurt more by masking of either type when it occurred with the NZ voice than when it occurred with the American voice.

After removing 5263 responses to nonword stimuli, the overall accuracy rate across the remaining 5308 responses was 89.6%. Once again, it was lowest in the Noise condition (86.8%), moderate in the Music condition (90.7%) and highest in Silence (91.3%). Accuracy when looking at just responses to real words was higher for the New Zealand voice (90.2%) than for the US voice (89.0%). The comparison of this data with the data presented above that included responses to nonwords shows there was a bias towards responding 'Word' for the NZ voice and a bias towards responding 'Not Word' for the US voice.

Figure 4.2 shows the mean accuracy rates for the interaction of Voice by Condition, and reveals the same pattern as that described above for the full dataset. When looking at the effect of voice, Music and Noise pattern together again, with the NZ voice having a moderate advantage over the US voice (1.0% for Music and 0.9% for Noise), and having a greater advantage in Silence (with accuracy to the NZ voice 1.8% higher than to the US voice). Once again, accuracy for the NZ voice drops more when masked, with no obvious difference between Music and Noise. The Lexical Access Hypothesis predicted that there would be an accuracy boost for the US voice in Music vs. Silence. Indeed this happened. But the purpose of the Noise control condition was to differentiate any effect of masking from any effect of expectations about accent in music. Since Music and Noise pattern together, there is no support for a facilitation to accuracy when processing the US voice in music.

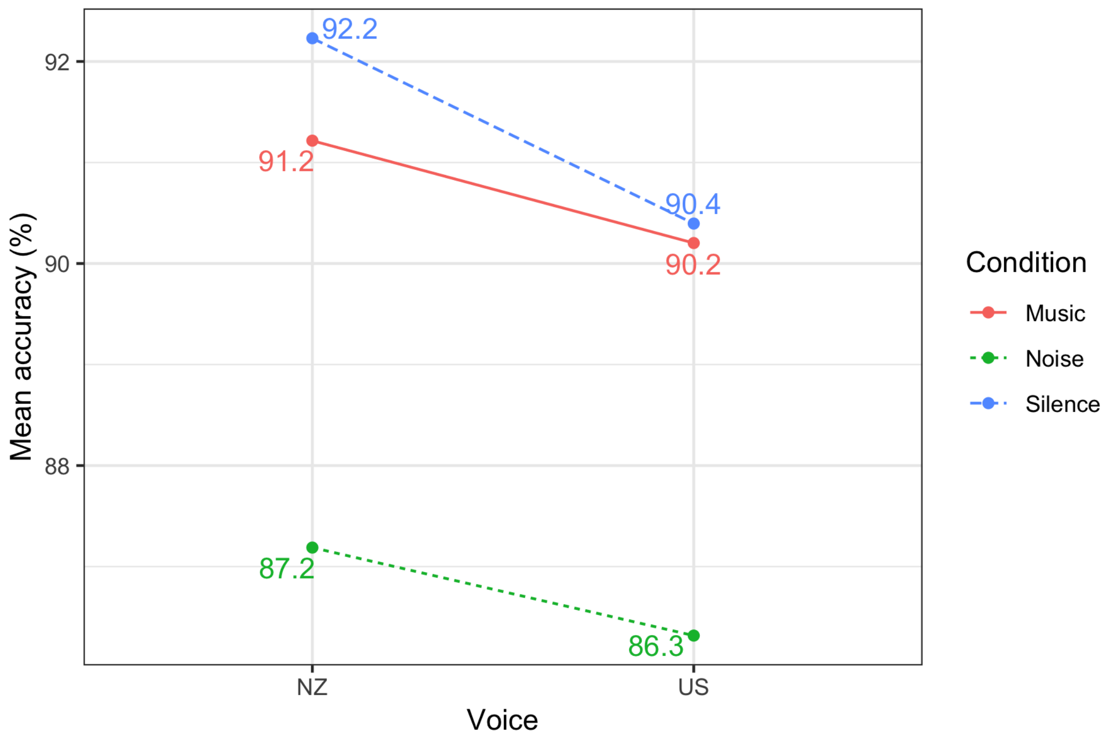


Figure 4.2: Mean percent accuracy for responses to real words, for the NZ and US voices in the three conditions.

To summarise, there was a tendency towards over-acceptance of nonwords in the native (NZ) dialect and over-rejection of real words in the different-region (US) dialect. This is not particularly interesting or surprising in its own right, but does justify the use of only the real word data in the statistical analyses of accuracy which follow. Both of the raw datasets show the same pattern when it comes to the interaction of Voice with Condition: even though noise-masking decreases accuracy more than does music-masking, they interact similarly with the voice distinction. Accuracy to the NZ voice is damaged to a greater extent by masking of either type than is accuracy to the US voice. This may actually be due to idiosyncracies of the voices themselves, rather than any general trend. It may be that the American speaker simply had a more clearly enunciated speech style than the NZ speaker. This is indeed my impression, particularly with respect to various consonant distinctions. For example, even in the Silence condition, the NZ speaker's distinction between /b/ and /v/ was sometimes quite ambiguous, while it was very clear for the American speaker.

Further analysis of the raw accuracy rates, looking at the various sociolinguistic variables in turn, revealed that NURSE was the only variable which seemed to support the Lexical Access Hypothesis. It had a higher mean accuracy rate to the US than NZ voice in all three conditions, but particularly in the music condition (US mean accuracy 2% higher than NZ mean in Noise and Silence, and 8% higher in Music). Attempts at modelling this apparent interaction between Voice and Condition for just the subset of trials containing NURSE, however, did not reach significance.

4.3.1.3 Raw results: Reaction time

The mean reaction time across all 10,571 stimuli (after removal of 229 outliers) was 1042ms. This was longer for nonwords (1089ms, n=5263) than words (996ms, n=5308), and longer for incorrect responses (1154ms, n=1058) than correct responses (1030ms, n=9513). Looking only at correct responses to words, which is the portion of the data used throughout the analyses of reaction time presented below, the mean across all 4756 correct responses to words was 981ms. This was longest in the Noise condition (1010ms), and similar in Music (967ms) and Silence (968ms). Responses were faster to the NZ voice (973ms) than the US voice (989ms).

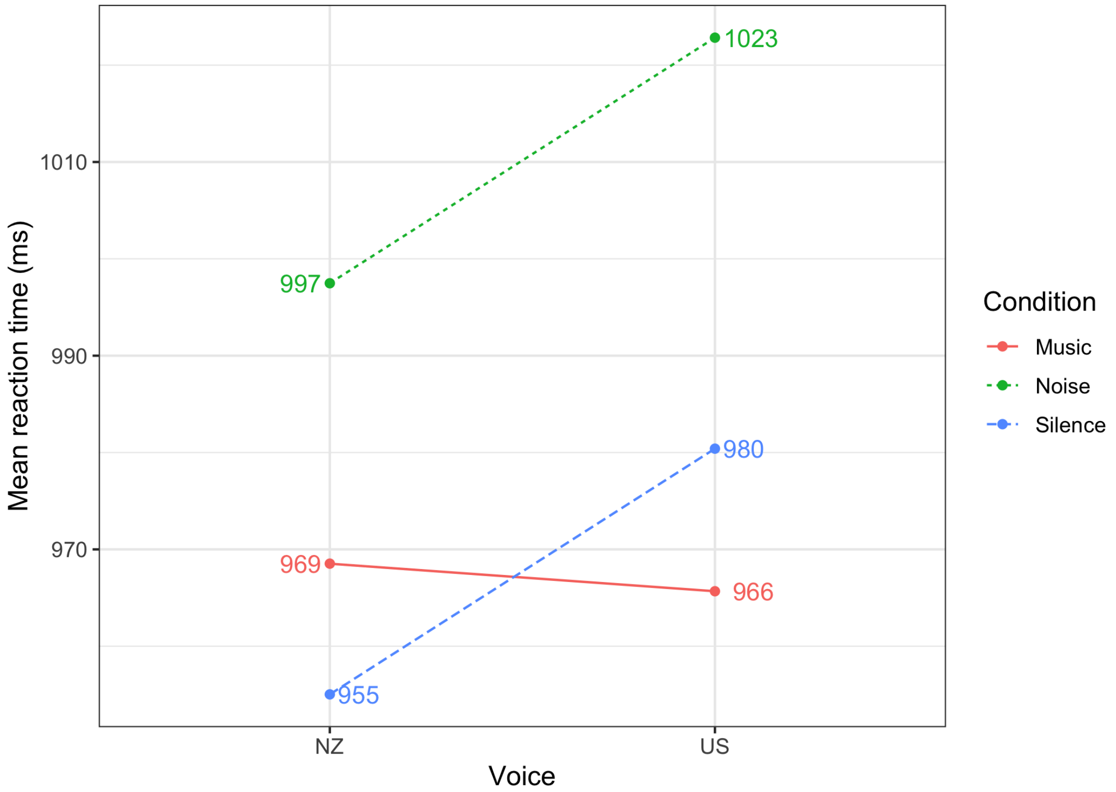


Figure 4.3: Mean reaction time (ms) for correct responses to real words, for the NZ and US voices in the three conditions.

As for the crucial interaction between Condition and Voice which is at the heart of the experiment's design, the raw results for mean reaction time across the two voices and three conditions are shown in Figure 4.3. Participants were faster to the NZ voice than the US voice in Noise (997ms vs. 1023ms) and Silence (955ms vs. 980ms), as expected. For the Music condition the pattern is different, and is in the hypothesised direction: the mean reaction time to the US voice is actually slightly faster than it is to the NZ voice (966ms vs. 969ms).

Looking at the raw data for reaction times across the different variables, once again, it is NURSE which shows the strongest support for the Lexical Access Hypothesis, with the mean RT for the US voice being 12ms faster than the NZ voice in Music, and slower than the NZ voice in Silence and Noise (by 60ms and 62ms, respectively). NORTH also appears to support the hypothesis. Note that in the statistical analyses which follow below, the rhotic variables (NURSE, NORTH, START) are grouped together, while GOAT, LOT and BATH are treated separately, as per the preregistration, since the rhotics have fewer trials.[3]

There are various factors that we expect to affect accuracy and reaction time other than those related to the research questions. It is thus important to build statistical models which can hold the variation in those factors constant so that we can see whether the patterns in the raw data hold. Firstly, in the next section, the accuracy data will be modelled to see if the apparent lack of support for the Lexical Access Hypothesis persists once other variables are taken into account. Then, in Sections 4.3.2.2–4.3.2.6, the RT data will be modelled, to see whether the support for the Lexical Access Hypothesis seen in the raw data is robust.

4.3.2 Statistical Models

It is not uncommon in older publications to see models which do not attempt to account for any variation in the data other than that related to the experiment's predictions. This is no longer accepted practice, and indeed, in this thesis I am attempting to present an approach which covers many different possible approaches to data analysis. This results section, then, is not solely a presentation of results, but also a foray into statistical methods. It should be noted that as with Chapter 2, my presentation style in this chapter blends the results and their discussion together to a large extent as the steps of data analysis unfold. I will present a series of models following the preregistered processes, followed by a model with an allowance for exploratory steps in the process. While the final model may introduce more bias, it also allows me to hone in on important features of the dataset that otherwise go unseen when stringently following a deductive, rather than partially inductive, approach.

Before beginning that process, however, I present here some simple and naive statistical analyses, models with no predictors other than those that speak directly to the research question. The most basic of all would be a linear regression with no random effects structure. If we model RT (from the start of the stimulus, and logged so as to fit the normality assumptions of the model) based on an interaction of Condition by Voice, the interaction approaches significance, trending in the hypothesised direction. Responses to the US voice are facilitated in the Music condition as compared to Silence ($p=0.064$) and when compared to Noise ($p=0.057$). There is no significant difference between Silence and Noise. Note, however, that such a model does not meet the assumption of the linear regression model that observations are i.i.d. (independent and identically distributed), since there are repeated measures for each participant, and each stimulus. We can expect that individuals will have some consistency in their responses, and this means that the many observations relating to that individual are not independent of one another. However, we do not know about, or have predictions about, the specific idiosyncracies of each participant — the variation between participants is to some extent 'random'. This is where linear mixed effects models (LMEMs) become a much more appropriate tool than basic linear regression. LMEMs can handle the fact that there are repeated measures, and account for the random variation associated with each individual, through random intercepts. For example, a random intercept can offset the fact that one participant

[3] The preregistration actually omitted LOT in error, but the intention to group the rhotics together was clear.

tends to respond slowly, and another more quickly. The idiosyncratic way in which individuals respond to other variables in the experiment can also be controlled for, through random slopes. For example, a given person might give quick responses to the NZ voice and particularly slow responses to the US voice. Rather than allowing that person to strongly affect the outcome of the model, a slope for Voice on Participant would be able to control for any idiosyncrasy that is not supported by a group-wide pattern.

In a simple linear mixed effects model, then, we can allow each Subject and each Word to have its own intercept, to control for random variation, and then model the interaction between Condition and Voice in the fixed effects (that is, variables that we have some prediction about). The following model was run: RT.lsc $\sim$ Condition * Voice + (1|Subject) + (1|Word). This model revealed a significant interaction in the expected direction with facilitation for the US voice in the Music condition as compared to Silence ($p=0.047$) and when compared to Noise ($p=0.01$). There is no significant difference between Silence and Noise.

Thus, we see already that while the raw results are reflected in these statistical tests, different analyses can affect the significance of that result. Controlling for variability between Subjects and the experimental stimuli allows the hypothesised interaction to reach significance.

As for the accuracy data, which will be modelled in detail in the next section, a basic generalised linear model (with no random effects) shows no significant interaction of Condition by Voice. Nor does a generalised linear mixed effects model with random intercepts for Subject and Word. In this latter model, the co-efficients are in the hypothesised direction, with less accuracy to the US voice in Noise and Silence than in Music. The difference between Noise and Music approaches a trend ($p=0.106$), while it is nowhere near significance for the difference between Silence and Music ($p=0.355$). There is no difference between Silence and Noise. As we shall see in the next section, the hint of a trend in this model goes away once we account for other variation in the dataset.

The procedures all follow the same major steps that were outlined in Section 3.5.2, but with the addition of a systematic method for deciding how to treat independent variables (IVs), as outlined below. For the preregistered models, the list of variables and interactions tested can be found in the preregistration document, included in Appendix C, though much of this information will also be summarised below.

For most of the independent variables, the preregistration listed two or more options for how the variable could be modelled. For example, it specified that the socio-economic index of a participant (NZSEI) should be treated as categorical, but with the option of either a binary or ternary split of the data, while the ratio of song to speech frequency was preregistered as either a continuous variable or as a binary categorical split of the words based on that ratio. Despite giving these options, the preregistration did not specify a process for deciding between them. In the PCT, the IVs were treated in their most raw form, but for the LDT it was decided that a more principled decision-making process should be used. This tightening of procedure was a response to my developing interest in the concept of limiting analyst bias. Having already completed the statistical analysis of the PCT, I began to see that decisions about how to treat IVs could add extra degrees of freedom to the analysis if strict procedures are not pre-defined.

The following method was decided upon to make this systematic. Firstly, a model was built with the most obviously important IVs in it, through a process of exploratory modelling. To this model, each of the options listed in the preregistration was added, and a log-likelihood comparison was conducted between each pair of minimally different models — that is, between the base model and each of the models with a version of the IV in it. Whichever version of the IV had the lowest p-value in the log-likelihood comparison was used in the model fitting procedure.

The outcome of this procedure will be described for each of the three preregistered models (LDT Models 1–3), along with the choices made for LDT Model 4, which were free from the constraints of the preregistration.

As for the dependent variables, accuracy was a binomial response represented as the log-odds of responding correctly, while reaction time was always logged, scaled and centred (scaling and centring done with R command *scale*). The calculation of RT from the end of stimuli will be introduced in Section 4.3.2.3.

Before describing the outcome of the options modelling process, I present here some variables that did not have options specified in the preregistration, or did not need to be tested systematically in order to choose between the preregistered options. This information applies to all of the preregistered models (LDT Models 1–3).

- Age: The preregistration listed 'age group' without any further specification. The decision was made to treat age as a binary factor with participants under 25 grouped together as 'younger' and those 25 and over as 'older'. See Section 3.3 for details about the ages represented in the sample. In all cases where a binary split was made of ordinal data, including this one, the groups were chosen such that the number of participants in each group was as even as possible.

- Gender: Only binary responses were attested. Gender is thus treated as a 2-level factor.

- Handedness: Only two responses were offered on the questionnaire, so this is treated as a 2-level factor. Note that in the preregistration this variable was marked with an asterisk meaning that interactions should be tested prior to removal of the variable. With only three left-handed participants, this was of questionable merit, but in the spirit of rigidly adhering to the preregistration, interactions were tested.

- Ethnicity: This was preregistered as a '2 or 3 level factor, grouping ethnicities deemed similar in terms of linguistic backgrounds'. It turned out that only two obvious categories emerged, with 33 participants identifying as 'Pākehā', 'New Zealander', 'New Zealand European' or 'European', and three participants identifying with Māori and/or Pasifika ethnicities. A two-way split between NZ European/Pākehā and Māori/Pasifika was thus used for ethnicity.

- Time spent overseas: No options were specified in the preregistration, and all models treated this as a 2-level factor with participants grouped according to whether they had spent less than two years overseas (n=23), or between two and six years overseas (n=13), following the principle of maximising the number of participants in each group mentioned above.

- Genre-liking behaviour: This was simplified into two clusters as described in Section 3.3.

Each of the other variables were subject to a decision making process in each of the models below, and will thus be re-introduced for each model. The next section is the first of four, each discussing the statistical model that resulted from a systematic model fitting procedure. The first, in this section, follows the preregistered model fitting procedure for the accuracy data (LDT Model 1). The preregistered models for the RT data are then presented, giving full details for the one measuring reaction time from the start of the stimulus in the main body of the text (LDT Model 2), and then summarising the results of the model measuring RT from the end of the stimulus (LDT Model 3), the details of which are presented in an appendix. An interim discussion will describe issues with the preregistered model fitting procedures, and finally, another model will be presented for the RT data, using a refined model fitting procedure, free from the constraints of the preregistration (LDT Model 4).

4.3.2.1 LDT Model 1: Preregistered model for accuracy data

The base model to be used for deciding between IV options was as follows: Acc $\sim$ Condition + (1|Subject) + (1|Word). The versions settled upon through the decision-making process are listed below.

Decision-making processes for choosing amongst preregistered versions of IVs for LDT Model 1:

- Socio-Economic Index: Neither version of NZSEI predicted accuracy well, but the 2-level factor ($p=0.63$) was better than 3-level factor ($p=0.94$).

- Time spent listening to music: This was better as a continuous variable ($p=0.11$) than as a 2-level factor ($p=0.15$).[4]

- Proportion of US/NZ artist listening: Three options were tested. For the first two options, the proportions of NZ and US music that a participant said they listen to were tested as continuous variables[5]. In a third option, the relationship between the above two variables was also tested, and was represented as a 3-level factor (USNZMusic), with the levels 'USdom' (participants for whom US artists are dominant in their listening behaviour, n=28), 'Equal' (participants who said the same proportion of their music listening is to US and NZ artists, n=6), and 'NZdom' (those who listen more to NZ than US artists, n=2). As can be seen from the participant numbers represented in each of these levels, this preregistered three-way factor was perhaps poorly conceived, given the dominance of US music in NZ. The log-likelihood comparisons for these three options revealed that amount of US music listened to performed better ($p=0.27$) than amount of NZ music ($p=0.99$) or the relationship between the two ($p=0.59$).

- How surprising is a NZ accent in a song?: The binary split outperformed the continuous variable ($p=0.18$ vs. $p=0.58$).

[4]Continuous versions of these survey items were based on the 5-point Likert scales. While Likert scales are inherently ordinal, not continuous, and the distribution of results tended not to be normal amongst the responses, the treatment of scales as continuous is a parsimonious approach.

[5]Binary splits of these two scales were not tested.

- How much did the stimuli sound like they were sung in the Music conditions?: The continuous variable was very slightly better (p=0.59) than the 2-level factor (p=0.62).

- Lexical frequency: Spoken frequency (based on the Cob Celex frequencies, so actually based on both the written and spoken portions of the Cobuild corpus, p=0.0028) performed slightly better than song frequency (based on just Lyrics Planet, p=0.0034), though both were highly significant.

- 'Songiness': In these first three models, Songiness is defined as the ratio of LyrPlan frequency to Celex frequency. This performed better as a 2-level factor (p=0.2) than as a continuous variable (p=0.84).

- Block: Treating block as a 3-level factor (i.e., SubBlock, which groups together, for example, the first block heard in each voice) performed better (p=0.007) than treating it as a 6-level factor (p=0.044) or as a continuous variable (p=0.18).

Description of LDT Model 1 output: Table 4.2 shows the output of the final GLMER model for the accuracy results. The dependent variable is the log-odds of responding accurately. Positive co-efficients thus correspond to higher predicted likelihood of responding accurately. The model had the following syntax:

LDTmod1 = Acc ~ Condition + Voice + AgeBinary + Gender + Musician + NZsurprisBinary + slxVar + speechFreq.sc[6] + SubBlock + Gender * slxVar + AgeBinary * Voice + AgeBinary * NZsurprisBinary + (1 + speechFreq.sc | Subject) + (1 + Condition | Word)

Most importantly, note that the interaction between Voice and Condition did not reach significance and was thus dropped from the model. There was no support for the hypothesis that participants would have a facilitation in accuracy when hearing a US voice in Music.

The other main effects were as follows:

- Condition: Lower accuracy in Noise than in Music or Silence

- Word frequency: Higher accuracy to high frequency words

- Block: The middle block of each half of the experiment had lower accuracy rates, with the first block of each half having the highest accuracy.

- Musician: In terms of participants, musicians had lower accuracy rates, while both gender and age were involved in interactions.

- Age by Voice interaction: The US voice caused more mistakes for older, but not younger, participants.[7]

[6]When 'sc' is appended to the end of a variable name, it indicates that the variable was scaled and centred.

[7]Note that three-way interactions were tested with Condition*Voice, as per the preregistration, and were not significant.

Table 4.2: Preregistered GLMER model for accuracy (LDT model 1).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.546	0.404	8.780	<0.001
ConditionNoise	-0.681	0.191	-3.564	<0.001
ConditionSilence	0.435	0.253	1.716	0.086
voiceUS	-0.413	0.171	-2.417	0.016
AgeBinaryyounger	0.170	0.292	0.583	0.560
GenderM	0.573	0.318	1.801	0.072
Musiciany	-0.700	0.231	-3.027	0.002
NZsurprisBinarylow	0.835	0.286	2.923	0.003
slxVarGOAT	0.416	0.416	0.999	0.318
slxVarLOT	-0.236	0.415	-0.568	0.570
slxVarrhoticity	-0.160	0.364	-0.440	0.660
spokenFreqs	1.096	0.290	3.781	<0.001
SubBlock2	-0.416	0.131	-3.180	0.001
SubBlock3	-0.280	0.134	-2.092	0.036
GenderM:slxVarGOAT	-0.845	0.383	-2.207	0.027
GenderM:slxVarLOT	-0.328	0.378	-0.867	0.386
GenderM:slxVarrhoticity	-0.896	0.328	-2.736	0.006
voiceus:AgeBinaryyounger	0.475	0.217	2.183	0.029
AgeBinaryyounger:NZsurprisBinarylow	-0.794	0.363	-2.190	0.029

- Age by 'Surprised to hear NZ accent in a song' interaction: Amongst older participants (the reference level for age), those who said they would not be surprised to hear a NZ accent in a song (responding 1–3 on the 5 point scale) were more accurate, however this was not the case for younger participants. The interaction of these two binary variables looks like a clear case of overfitting. There was no motivation for testing an interaction like this. By testing all possible interactions, and by using binary versions of each of these two variables, Type I errors became more likely. This will be discussed further below.

- slxVar by Gender interaction: As for the sociolinguistic variables involved, there is an interaction with gender. For the reference level of Variable (BATH), males were more likely to be accurate than females. For rhoticity and GOAT the opposite pattern held, while the gender difference was minimal for LOT.

Note that reaction time is actually a very good predictor of accuracy, and its omission was one important problem with the preregistration specific to this model. To examine this, further modelling was done with RT included, and it did prove to be highly significant, with longer reaction times predicting lower accuracy. VIFs were tested to ensure multi-collinearity was not problematic. In these models, the interaction of Condition by Voice remained non-significant, however, so further detail about these models is not included.

4.3.2.2 LDT Model 2: RT from start of stimulus, preregistered model fitting procedure

This section presents an analysis of the reaction time data, as measured from the start of each stimulus, strictly following the preregistered model fitting procedure. I will refer to this measure simply as RT, whilst in the next section, when dealing with reaction time as measured from the offset of the stimuli, the shorthand RTend will be used. To begin this section, the outcome of the modelling of IV options is presented below, showing the version of each independent variable used in model fitting. The reader is referred to Section 3.3 for a description of the questionnaire items that inform the categorisation of participants.

These decisions were made using the method described in the previous section, and in this case, the base model against which all options were compared had the following form: RT.lsc ∼ Condition * Voice + slxVar * voice + length.sc + (1|Subject) + (1|Stimuli)

Decision-making processes for choosing amongst preregistered versions of IVs for LDT Model 2:

- NZSEI: A median split was better in comparison to the base model (log-likelihood comparison p=0.19) than a tertile split (p=0.42).

- NZsurpris: This better improved model fit when treated as a continuous variable (p=0.56) than as a binary factor (p=0.99).

- StimSingy: Whether people thought the stimuli in the music conditions sounded like they were sung improved fit better when modeled as a binary factor (p=0.56) than as a continuous predictor (p=0.92).

- MusicListening: This did not improve the base model significantly either as a continuous variable (p=0.82) or as a factor (p=0.74), but the binary split was the better of the two.

- NZ vs. US music listening: Proportion of NZ/US music listened to, and the relationship between them were compared. Amount of NZ music listening did not improve the model (p=0.97), while US music listening did much better (p=0.20). The three-way factor USNZMusic (US-dom, Equal, NZ-dom), despite having sparsity issues, did the best of these three options (p=0.16).

- Lexical frequency: Celex (p=4.208e-05) and LyrPlan (p=7.651e-10) frequencies are both highly predictive of reaction times, with faster responses to high frequency words in both cases. Song frequency was used in model fitting since it had the lower of the two p-values.

- Songiness: The ratio of LyrPlan to Celex did not improve the base model when added as a continuous predictor (p=0.76), but as a binary factor based on a median split of this ratio, it significantly improved the base model (p=0.002). Reaction times were faster to 'speechy' words than to 'songy' words. Including both the sung frequency itself and the ratio of that frequency to the spoken frequency was carefully considered given a suspicion that it would cause collinearity problems. VIF tests were all non-problematic, however, so both Frequency and Songiness were used in model fitting.

- Block: Block could be modeled in multiple ways — as a continuous variable or as a factor, and with two potential ways of grouping comparable blocks together. The decision making process for Block was complicated by the fact that a specific interaction (Block*Trial*Condition) was preregistered, necessitating a more nuanced approach to the decision-making process. Which version of block to use was based on how much it improved model fit when in interaction with Trial and/or Condition.

  The three-way interaction of Block*Trial*Condition was near significantly better than the component two-way interactions for both the factor and the continuous versions of Block (p=0.09 for factor and p=0.07 for continuous).[8] The factor was much better than the continuous version, however, when comparing the Block*Trial interaction to the component main effects (p=0.002 for factor, p=0.47 for continuous). Similarly, treating Block as a factor was better for the Block*Condition interaction (p=0.004 for factor, p=0.11 for continuous). Since the difference between the categorical and continuous options was marginal for the three-way interaction, but large for the two-way interactions, it was decided that Block should be treated as a factor in the model fitting procedure.

  Having decided upon using a categorical version of block, the model output for the 6-level factor was considered, to determine whether the two aspects of Block (BlockHalf and SubBlock) could be sensibly separated — that is, do Blocks 1, 2, and 3 pattern differently from Blocks 4, 5, and 6, or do Blocks 1 and 4, Blocks 2 and 5, and Blocks 3 and 6 behave similarly, despite occurring in the different halves of the experiment? There was no clear pattern according to the half of the experiment in which a trial occurred. There did, however, appear to be a similar pattern for SubBlock number within each half. Log-likelihood comparisons were thus conducted to compare Block6 with SubBlock. Once again comparing models with the three-way interaction to models with the component two-way interactions, SubBlock improved model fit more than Block6 (p=0.049 vs. p=0.095). Block was thus modeled as a three-level factor, grouping together the first, second and third blocks encountered in each half of the experiment.

Once decisions had been made about how to treat each of the independent variables, the model fitting procedure was conducted, pruning non-significant items from an initially maximal model and checking all backward steps with log-likelihood comparisons, checking VIFs at various points in the process, and checking all preregistered interactions. Following this preregistered procedure led to a final model with the following terms:

LDTmod2 = RT.lsc ∼ Condition * Voice + Gender + songFreq.sc + slxVar * Voice + slxVar * length.sc + AgeBinary * Condition + AgeBinary * SubBlock + SonginessBin * Trial50 + SubBlock * Trial50 + (1 | Subject) + (1 | Word)

[8]To ensure that the log-likelihood comparisons were based on minimally different models, models with three-way interactions were compared to models that contained all component two-way interactions. For example, the model with (Block(continuous) + Trial + Condition)^3 was compared to the model with (Block(continuous) + Trial + Condition)^2. The p-value from that log-likelihood comparison then competed with the p-value resulting from a comparison of (Block(6-level factor) + Trial + Condition)^3 to a model with (Block(6-level factor) + Trial + Condition)^2.

Table 4.3: Preregistered LMER model for reaction time (LDT model 2).

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.239	0.149	88.069	-1.608	0.112
ConditionNoise	0.050	0.050	4574.762	1.004	0.315
ConditionSilence	-0.102	0.049	4579.323	-2.062	0.039
voiceUS	-0.103	0.058	4602.202	-1.770	0.077
GenderM	0.426	0.142	32.981	3.000	0.005
sungFreqs	-0.157	0.032	130.076	-4.887	<0.001
slxVarGOAT	-0.133	0.102	174.924	-1.300	0.195
slxVarLOT	0.067	0.108	202.918	0.618	0.537
slxVarrhoticity	-0.025	0.090	178.593	-0.282	0.778
lengths	0.143	0.039	2917.495	3.668	<0.001
AgeBinaryyounger	-0.167	0.140	40.320	-1.199	0.238
SubBlock2	-0.092	0.064	4572.419	-1.430	0.153
SubBlock3	-0.281	0.064	4581.115	-4.392	<0.001
ratioBinaryspeechy	0.031	0.076	251.611	0.405	0.685
Trial50	<0.001	0.002	4573.905	0.216	0.829
ConditionNoise:voiceUS	**0.143**	**0.054**	**4570.686**	**2.634**	**0.008**
ConditionSilence:voiceUS	**0.115**	**0.054**	**4572.961**	**2.152**	**0.031**
voiceus:slxVarGOAT	0.124	0.092	4545.268	1.355	0.176
voiceus:slxVarLOT	0.079	0.080	4404.308	0.987	0.324
voiceus:slxVarrhoticity	0.204	0.068	4684.572	2.987	0.003
slxVarGOAT:lengths	-0.029	0.064	3320.440	-0.449	0.654
slxVarLOT:lengths	0.104	0.067	1728.522	1.559	0.119
slxVarrhoticity:lengths	-0.074	0.047	2937.487	-1.570	0.117
ConditionNoise:AgeBinaryyounger	0.119	0.056	4576.925	2.149	0.032
ConditionSilence:AgeBinaryyounger	0.101	0.055	4578.247	1.857	0.063
AgeBinaryyounger:SubBlock2	0.142	0.055	4574.173	2.577	0.010
AgeBinaryyounger:SubBlock3	0.129	0.055	4575.375	2.344	0.019
ratioBinaryspeechy:Trial50	-0.003	0.002	4576.600	-2.194	0.028
SubBlock2:Trial50	-0.001	0.002	4576.306	-0.563	0.573
SubBlock3:Trial50	0.006	0.002	4580.065	3.124	0.002

The output of this model is presented in Table 4.3, and each term in the model is summarised below. Note that the *lmerTest* package was used to calculate p values for each term in the model. The highest VIF for the model was 7.5, for the interaction between Block and Trial. Crucially, in this model, the interaction between Condition and Voice is significant, with participants' responses to the US voice being facilitated in Music as compared to Silence and Noise.

- Condition by Voice interaction: Responses to the US voice are faster than to the NZ voice in the reference level of Condition (Music), and slower than the NZ voice in Noise and Silence. This can be seen by looking at the main effect for voice in addition to its interaction with Condition. Note that Voice also interacts with slxVar, so the main effect for Voice, and the Condition*Voice interaction are showing co-efficients for the reference level of SlxVar: BATH. When Condition is relevelled in an otherwise identical model, we find that Noise and Silence are not significantly different to one another (p=0.61) in the

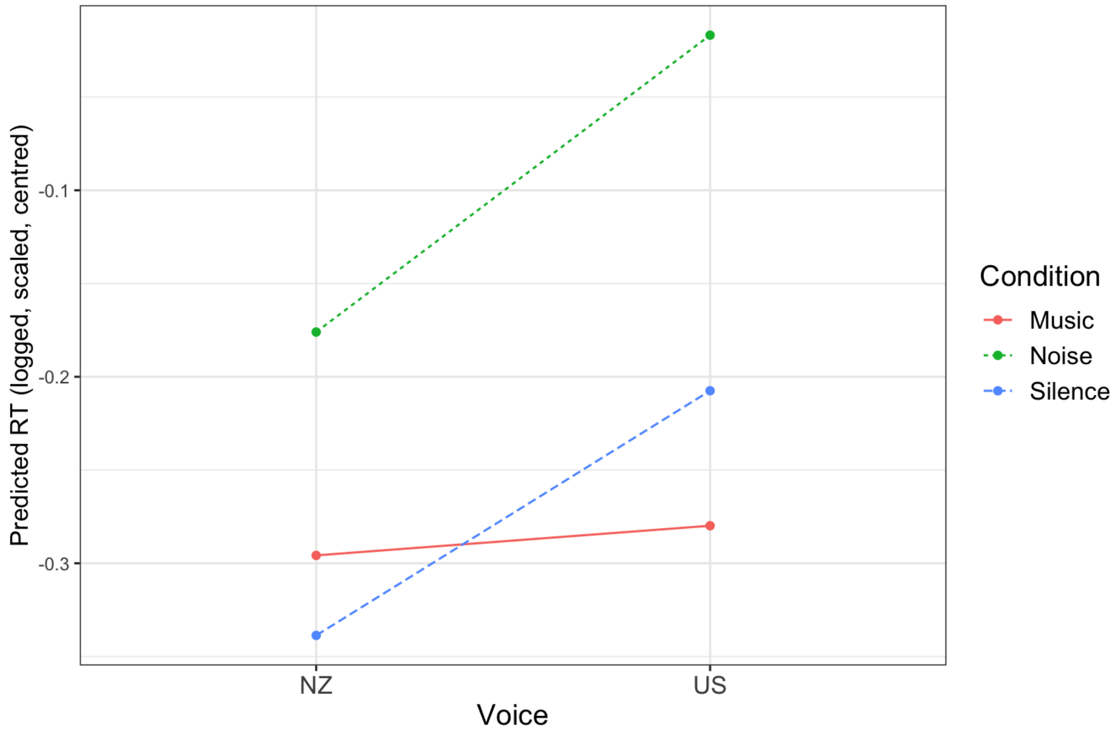


Figure 4.4: Interaction of Condition and Voice in the preregistered model with RT measured from the start of stimuli (LDT model 2).

context of the Condition by Voice interaction. This can be clearly seen when the interaction is plotted, in Figure 4.4, where the lines for Noise and Silence are parallel when comparing between the two voices. The clear facilitation for the US voice in Music can also be seen, with the line for Music crossing over the line for Silence between voices.

- Gender: Males had longer RTs than females.

- Frequency: Higher frequency words had faster RTs.

- Voice by Sociolinguistic Variable interaction: As described above, the effects for Voice depend upon Condition. They also vary according to Sociolinguistic Variable, with the US voice attracting facilitated responses compared to the NZ voice when looking at the reference level of slxVar, which is BATH. This facilitation becomes progressively less when looking at LOT, GOAT and rhoticity, respectively.

- Length by Sociolinguistic Variable interaction: Longer words had slower RTs. While this main effect for length interacts with Sociolinguistic Variable, all levels show the same direction of effect, with RT increasing for longer words, but to differing degrees for different variables.

- Condition by Age interaction: The main effect for Age shows faster responses for younger participants in the reference level for Condition (Music). The interaction shows that while the young group are still predicted to be faster than the older group in all three conditions, the effect is diminished in Noise and Silence.

- Age by SubBlock interaction: Younger participants are faster than older in the reference Block, Block 1. This gap diminishes to almost zero in Blocks 2 and 3.
- Songiness by Trial interaction: There were faster responses to 'songy' words at the start of the block, but this effect diminishes as the block progresses such that by the end of the block responses are faster to 'speechy' words.
- Block by Trial interaction: Participants get faster as they move through the three blocks of each half of the experiment. This interacts with Trial number, however, such that in Block 3 they slow down as the block progresses.

These results will be discussed with respect to the Lexical Access Hypothesis in Section 4.4. In the coming sections, however, I discuss some limitations of the preregistered statistical methods and present further analyses of the reaction time data.

4.3.2.3 LDT Model 3: RT from end of stimulus, preregistered model fitting procedure

The preregistration included a second RT model in addition to the one just presented. This model differs in its dependent variable, by measuring reaction time from the end of the stimulus. While it is typical to measure RT from the start of the trial, the RTend model was included since it was the one presented in Walker and Hay (2011). The same process of careful model fitting was carried out, and the resulting LDT Model 3 showed very similar results to LDT Model 2 presented above. Since it does not add any substantially different findings to those already presented, it is included as Appendix D, rather than here. The detailed description of the modelling procedure and the final resulting model can be found there.

As became apparent in the discussion of LDT Model 2 above (and as is also discussed for LDT Model 3 in the appendix), the modelling procedure outlined in the preregistration had various weaknesses. In the next two sections, these issues will be discussed. While sticking as closely as possible to the preregistered methods, a new model fitting procedure will be introduced to rectify these issues, and a final model will then be presented in Section 4.3.2.6.

4.3.2.4 Interim discussion 1: Problems with the preregistered models

While the two reaction time models presented so far give an unbiased examination of the research question at the heart of the experiment, they also involve some steps that reflect the fact that the preregistration was made without any direct knowledge of how the data would turn out. In this section, the lessons learned from these models will be discussed, setting the scene for a final model of the RT data (LDT Model 4) which will be presented in the next section. It should be emphasised here that the preregistered models remain the most important evidence for the hypothesis of this experiment. The process followed to achieve LDT Model 4 benefits from the ability to respond to the nuances of the data, but loses the objectivity of the preregistered analysis. My reasoning for going into some detail here about the ways in which the preregistration could have been improved, is to guide decisions made in the preregistration of future experiments.

The issues with the preregistered models are outlined below. In each case, I state the problem, and then describe how that problem will be resolved in the model fitting procedure used to determine LDT Model 4. After considering the more basic of these issues, I will present an in-depth discussion of the way the Songiness variable was calculated in the preregistered models, and put forward the rationale for a new calculation of that variable to be used in LDT Model 4.

- The preregistered model fitting procedure worked by removing least significant IVs first, checking all 2way interactions prior to their removal. This was done to avoid throwing out variables that might actually be important in modelling the data once tested in interactions. However, testing the interactions of the least significant IVs first had the unintended effect of actually giving them priority status in the fitting of interactions. This may have led to over-fitting of spurious interactions and Type I errors. Additionally, interactions were *only* tested for non-significant main effects. That is, there was no scope for testing interactions of those effects which were significant throughout the procedure. A more sensible modelling procedure would pre-specify interactions with some *a priori* motivation, and test only those interactions. This method will be employed for fitting LDT Model 4.

- I also became concerned by the practice of splitting participants into two groups for a range of questionnaire items. While deciding between the preregistered options was done in a very structured way, the very idea of testing a lot of interactions with a range of binary factors may invite overfitting. For that reason, in LDT Model 4, I will tend towards using continuous versions of the questionnaire items where possible, and not test interactions if this is not possible. Relatedly LDT Model 4 will not test for any age effects since the sample of participants is not variable enough in terms of age. Differences in exposure to accents in song over apparent time is of great interest, but it cannot be sensibly tested with this cohort of participants.

- Another oversight of the preregistration was that the reaction time of the preceding trial was not included in models. This is a well-established and robust predictor of reaction time in LDTs (Goldinger, 1996a). If the previous trial had a slow response, the present trial is also likely to have a slow response. This IV will be included in LDT Model 4.[9]

- The preregistration did not consider the possibility that the hypothesised effects would occur only early on in the experiment. Based on the results of the PCT, it is reasonable to expect that the effects of expectation will wear off slightly as the participants tune into the voices. The hypothesised result may therefore be stronger near the start of the experiment. To examine this,

[9]The inclusion of PrevRT required a decision about how missing values would be handled, since any observation with missing values in one of the IVs is excluded from LMER models. For the first trial of every block (where there *was* no previous reaction time), and for the 53 trials where the preceding RT was zero (that is, in cases where the participant had not responded to the previous trial), a preceding RT of 994ms was entered. This was the median RT across the entire dataset, that is for words and non-words, and for correct and incorrect responses, which is appropriate since PrevRT is an experimental control based on all trials in the experiment, not just those of interest to this analysis.

a three-way interaction of BlockHalf * Condition * Voice will be tested in the fitting of LDT Model 4.

- The preregistered random effects structure throughout included a random intercept for Word, but actually this should be an intercept for Stimulus nested within Word since there are two recordings per word, one in each voice. This nested random effects structure will be used when fitting the final model.

- The preregistration did not specify that continuous variables such as lexical frequency should be tested for normality and then transformed if necessary. In LDT Model 4, all continuous variables will be scaled and centred (marked with '.sc' at the end of the variable name), and if they are right-skewed they will also be logged (which is the case for RT, PrevRT, Frequency and Songiness).

- Whilst a procedure for avoiding multicollinearity of predictors was included in the preregistration for the model pruning phase, no maximum VIF value was set for the final model. In practice, I worked with a maximum VIF of 10 for the preregistered models, a practice I follow again for LDT Model 4.

- The relationship between the amount of NZ and US listening was preregistered as a factor with three levels, but these levels were very unevenly represented, with only two people listening to more NZ music than US music and six people listening to the same amount of NZ and US music. The remaining 28 participants listened more to US music than NZ music, to varying degrees. To rectify this issue, a new variable is created for LDT Model 4, which I will call USorientation. The aim of this variable is to capture three pieces of information gathered about participants' music listening practices in a single numeric variable. To do this, the NZ listening score is subtracted from the US listening score for each participant and then this value is scaled according to the overall amount of music they listen to. The formula for USorientation, based on three Likert scale responses, is thus: (USmusic – NZmusic) * MusicListening. In this way, those who listen to the most music are given more extreme values, and their preference for US or NZ music is given more weight in the model. This is a more nuanced representation of the data, and it is of a more genuinely continuous nature, being much closer to a normal distribution than the original Likert scales. The maximum value for USorientation is 10, for three individuals who listen to mainly US music, and a lot of it. The minimum is -3, for a NZ oriented listener. The mean is 3.97, reflecting the overall dominance of US music, as already discussed.

- One final limitation of the first three models relates not so much to the preregistration as to choices I made in calculating the lexical frequency IVs. After running LDT Models 1–3, I started to question the method that had been used to calculate SongFreq, SpeechFreq, and the resulting ratio between them. The remainder of this section presents a careful consideration of the rationale for this variable, and puts forward a new method for its calculation, which will be used in LDT Model 4.

4.3.2.5 Interim discussion 2: Lexical frequency across registers

As discussed in various sections above, lexical frequency plays an important role in word recognition (Connine et al., 1993). This subsection will bring issues regarding lexical frequency into focus prior to running LDT Model 4, in order to best motivate the predictions that can be tested by that model. In particular, I will present a more thorough description of the 'Songiness' variable.

The rationale for such a consideration is as follows. We know that frequency effects exist, with faster lexical access for higher frequency words. We also know that these effects must be driven by an individual language user's linguistic experiences. The purpose of any frequency measure drawn from a corpus is to make a guess about the amount of exposure the individual of interest has had to different words. Language varies systematically within a society, and this structured variation no doubt extends to lexical frequency. Such variability needs to be considered, then, when dealing with subtle frequency effects in speech perception experiments.

For example, listeners have encountered words at differing frequencies for different types of speaker. They may learn to expect *trash* from an American speaker and *rubbish* from British and NZ speakers, for example. Reaction times could thus be faster not just to lexical frequency overall, but to context-weighted lexical frequency. Taking a ratio of the Canterbury Corpus to the Buckeye corpus reveals that amongst the most NZ-like words are *sporty*, *carving* and *boating*, while the more US-like words include *profit*, *rafting* and *gotten*. It could be that there are faster RTs to the NZ voice on the former set of words, and to the US voice on the latter. In the same vein, songy words may be accessed faster in the music condition, and speechy words in the non-music conditions. While it is beyond the scope and design of this experiment to test for dialectal differences in lexical frequency, it is worth examining whether register-based differences in frequency affect reaction times. This is the reason the Songiness variable is included in the statistical models.

In the first three models, Songiness was calculated as the ratio of frequencies in the two large corpora, LyricsPlanet and Celex (Cob). This version of Celex includes both CobS (spoken data in the Cobuild corpus) and CobW (written data from Cobuild). Further contemplation about this method for calculating corpus specific frequencies calls into question the validity of including frequencies from CobW, when the whole experiment is geared towards auditory linguistic experiences, rather than experiences with written language. For this reason, before conducting the analysis of the final model, I decided to take a more principled approach to word frequency. Of the 150 real words used in the task, which ones are most likely to occur in song, speech, or writing? Do frequencies in speech and writing correlate strongly with one another or should texts from the written corpus be excluded?

For the purposes of this analysis of lexical frequency, then, Celex is split into CobS and CobW, resulting in a total of six corpora: CobW, CobS, Canterbury Corpus (CC), Buckeye, LyricsPlanet and PoPS (see Section 2.3.3 for a description of each of these corpora). Figure 4.5 shows the relationship between lexical frequency in three different sets of these corpora, for the 150 words used in the LDT. The frequencies have been logged for this display because the distribution in all cases is strongly right-skewed. The value labels shown on the axes, however, have been backtransformed to show the actual occurrences per million words for each point.

Before looking in detail at the plot, let us consider the relationships between these corpora. There is a significant correlation between each pair of corpora for the 150 words under consideration, reflecting a degree of functional uniformity in language use across domains. Interestingly, though, the frequencies of words in speech and writing are more strongly correlated with each other ($\rho = 0.8$) than either one is to song. Writing and song are slightly more strongly correlated ($\rho = 0.44$) than speech and song ($\rho = 0.38$). Even though song seems to be the most distinct mode in terms of lexical frequency, this analysis suggests that it would also be sensible to separate spoken and written sources of lexical frequency for the purposes of LDT Model 4. I will therefore proceed with a version of spoken frequency that excludes the written portion of Celex. Before leaving the topic of register differences in lexical frequency, though, Figure 4.5 will be described in more detail.

On the vertical axis (the z axis in the language of R's *scatterplot3d* package, used to make this plot) is the average of the lexical frequency in PoPS and in LyricsPlanet. The words sitting on the top of the tallest 'sticks' are words which are highly frequent in songs. On the horizontal axis (x axis) is spoken word frequency, based on the average frequency across the spoken portion of Celex (i.e. CobS), the Buckeye corpus and the Canterbury Corpus, reflecting speech across a range of dialects. Words to the right hand side are highly frequent in speech, while on the left hand side, at 0.05 occurrences per million, there are several words that weren't present in any of the spoken corpora. On the axis representing depth (the y axis) is the frequency of each word in the written portion of Celex (i.e. CobW). This dimension is also represented through the coloured shading of the points. This shading, along with the lines connecting each point to its position on the x-y plane can help with interpretation of the position of points in the three-dimensional space.

In any case where a word was missing from the given corpus (for writing) or set of corpora (for speech and song), a constant of 0.05 was added to the occurrences per million value, so that ratios could be calculated between corpora. These 'missing' words can be seen clearly along the 0.05 line on the left, for words which did not occur in the spoken corpora, for just a couple of words at the front of the graph for words that were not in the written corpus, and on a handful of 'short sticks' representing words which were not in the lyrics corpora.

To visually assess how 'songy' a given word is – that is, the extent to which it is over-represented in songs as compared to speech and writing, we can find a clump of sticks that are nearly overlapping, but have different heights. For example, at the front left of the graph, we see that while the word *harpist* did not appear in any of the three corpora, the word *flirty* was absent from speech and writing, but somewhat frequent in songs. If we look at higher frequency words, we see that while *farming* and *dancing* are similarly frequent in speech and writing, *farming* is strongly under-represented in song lyrics, while *dancing* is particularly songy. The word *farming* has a frequency of 26 occurrences per million words in speech, 33 in writing, and just 0.15 in song (coming from three tokens in the LyricsPlanet corpus). The word *dancing*, on the other hand, has 28 occurrences per million in speech, 34 in writing, and 167 in song. The word *farming* is 'speechy', while *dancing* is 'songy'. Other words are 'writy' and 'songy', but not 'speechy'. For example, the word *sorrow* occurs frequently in both writing and song, but did not occur at all in the spoken corpora. The word *rafting*, by comparison, is frequent in speech, but absent from song and rare in writing.

Comparing the Songiness ratios used in LDT Models 2 and 3 (which included just LyricsPlanet and all of Celex) to the new method (using all five corpora, and

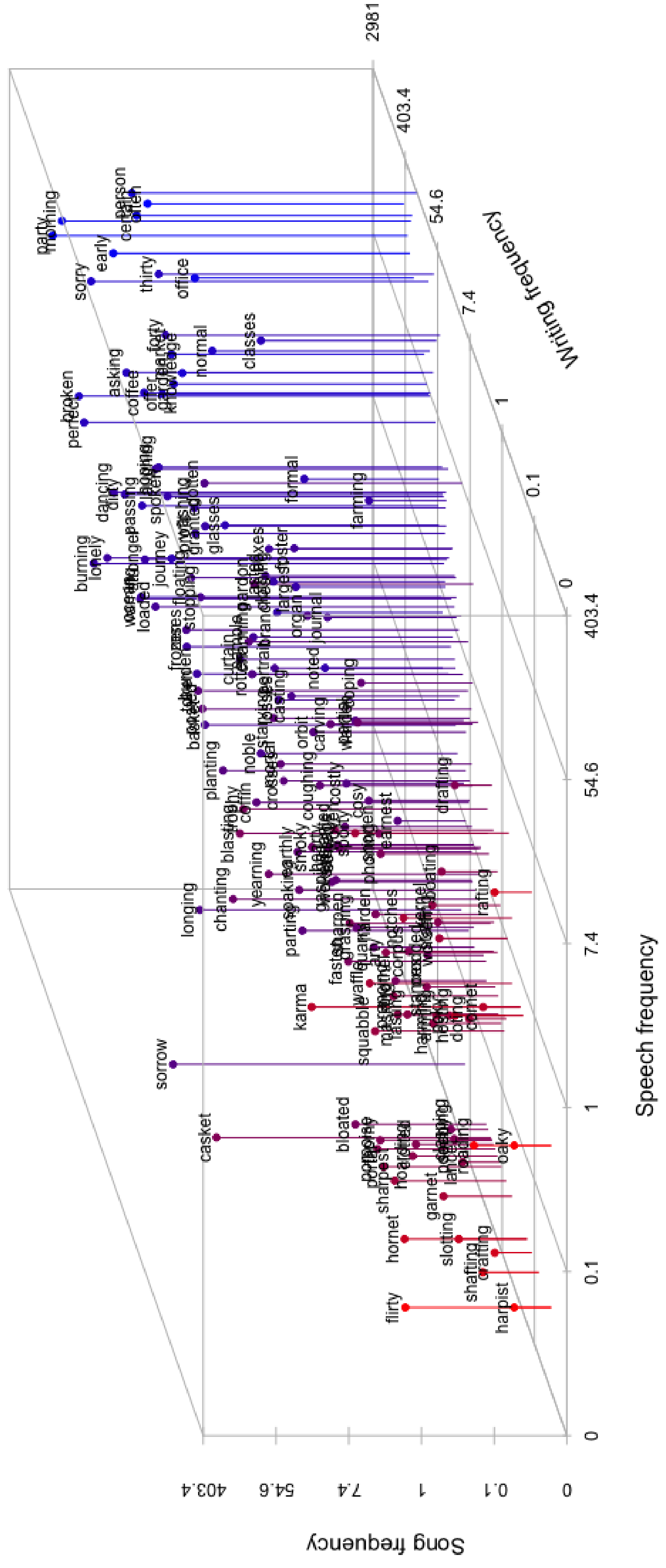


Figure 4.5: Frequency of LDT stimuli in three types of corpora: song, writing and speech. Values shown are the log occurrences per million words with backtransformed axis labels.

excluding CobW), several words change from songy to speechy and vice versa. Words such as *sorrow*, which could perhaps be described as 'literary' words, did not show up as being songy under the first method, since they had a high frequency in Cob. By excluding CobW, these words show up as being more songy. Other examples of this type of word are *casket*, *longing*, *chanting*, *burning* and *warning*. Words which are speechy words but not writy words, by comparison, had their token counts increase as a proportion of the speech corpora once CobW was removed. Examples of words that show up as speechy under the new method include *roading*, *sporty*, *pardon*, *slanted*, *notches*, *gotten* and *sample*. In sum, this new method is a more careful and nuanced way of capturing register based variation in word frequency than the method used in the first models.

With this deeper understanding of how lexical frequency varies across different registers, the following prediction can be made about how these differences might affect the results of the present experiment: Songy words should attract faster reactions in the Music condition than in the Silence or Noise conditions, while speechy words should show the opposite pattern. Thus, an interaction of Songiness with Condition will be tested in LDT Model 4. An interaction of Songiness and Voice would also be theoretically interesting. However, trying to understand not only frequency differences between registers but also between dialects would require a much more carefully selected set of stimuli. This analysis of frequency is, after all, post hoc. The criteria for choosing words for the experiment were based on dialect differences at the phonological level, without any attention paid to lexical frequency ratios. Songiness * Voice will thus not be tested.

4.3.2.6 Final RT Model using refined model fitting procedure (LDT Model 4)

The various decisions made above were implemented in the fitting of LDT Model 4. The maximal model included all interactions to be tested, each of which had a motivated prediction. From this maximal model, pruning was conducted in a linear down-stepping fashion, without adding further terms to the fixed effects at any point in the process. As for the random effects structure, the same process was undertaken as in the previous models: once all fixed effects in the model were significant, all possible slopes were added to the model. As that model did not converge, slopes explaining the least variance were removed until the model converged. First, all slopes explaining less than 0.01 of the variance were removed, then those with variance under .02, and so on.[10] The terms included in the maximal model, along with a prediction for each, are listed below.

- Predictions for main effects not expected to be involved in any interaction:
  - Length — Longer words will have slower RT.
  - Frequency — This is the mean of speech mean frequency and song mean frequency for each word, based on the five corpora discussed above, and excluding CobW. We expect faster RT on higher frequency words.

[10]The nested intercept of Stimuli within Word was attempted but caused convergence issues, and was thus abandoned. However, the slope for Voice | Word was eventually added, meaning that same variance was still assigned to the random effects structure.

- Previous RT — A longer RT in the previous trial should lead to a longer RT on the current trial.
- Gender — Based on LDT Models 2 and 3, males are expected to be slower than females.

• Predictions for potential two-way interactions:

- Condition * Voice — This interaction tests the Lexical Access Hypothesis, that RTs will be facilitated for the US voice in Music as compared to Silence and Noise. Note that this is also tested in three three-way interactions, listed below.
- Musician * Condition — Musicians are expected to have faster reaction times than non-musicians in the Music condition. [11]
- Songiness * Condition — As outlined in the discussion of lexical frequency above, it is expected that the more songy a word is, the faster its RT will be in the Music condition, and the slower its RT will be in Silence and Noise.
- Songiness * USorientation.sc — Based on the significant interaction of Songiness with USNZMusic in LDT Model 3, it is expected that US oriented listeners will have facilitated access to more songy words.
- Trial50 * SubBlock — This interaction has been consistent in the other models. Participants get faster as they go through each half of the experiment but slow down in the last block of each half.

• Predictions for potential three-way interactions:

- Condition * Voice * BlockHalf — The hypothesised effect may be stronger in the first half of the experiment.
- Condition * Voice * USorientation.sc — The hypothesised effect may be stronger for those who listen mainly to US music, and lots of it.
- Condition * Voice * StimSingy (as a 2-level factor) — The hypothesised effect is expected to be weaker for participants who did not feel like the voices were singing in the Music conditions.

After progressing through the model fitting procedure, pruning non-significant terms, or those which caused problems with multi-collinearity or convergence, LDT Model 4 was reached. The maximum VIF in the model was 7.1, for the three-way interaction between Condition, Voice and StimSingy. The model output is shown in Table 4.4, and had the following syntax:

[11] Musicians may simply be more motivated and engaged than non-musicians in the music condition. Additionally, there is a rationale for this expectation based on musical ability. RTs are facilitated when participants know when the next response will be required (see Nobre and Rohenkohl, 2014, for a review), and while all three conditions have the same stimulus onset asynchrony, the Music condition clearly demarcates that temporal period with a range of rhythmic information to which the listener might entrain. Another body of literature shows that musicians can predict intervals more accurately than non-musicians Repp (2007). The combination of these two findings would predict relatively faster reaction times for Musicians in the Music condition.

LDTmod4 = RT.lsc ∼ Length.sc + Freq.sc + PrevRT.lsc + Gender + Musician * Condition + Trial50.sc * SubBlock + Condition * Voice * USorientation.sc + Condition * Voice * StimSingy + (1 | Subject)
+ (1 + Voice | Word)

Table 4.4: Final LMER model for reaction time (LDT model 4).

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.336	0.116	44.809	-2.889	0.006
Length.sc	0.106	0.020	216.371	5.359	<0.001
Freq.lsc	-0.217	0.027	141.127	-7.960	<0.001
PrevRT.lsc	0.101	0.012	4573.655	8.280	<0.001
GenderM	0.352	0.133	30.794	2.646	0.013
ConditionNoise	0.051	0.057	4514.695	0.889	0.374
ConditionSilence	-0.001	0.057	4525.784	-0.008	0.994
Musiciany	-0.222	0.157	35.057	-1.408	0.168
SubBlock2	-0.035	0.028	4557.750	-1.243	0.214
SubBlock3	-0.052	0.028	4530.882	-1.852	0.064
Trial50.sc	-0.014	0.019	4533.432	-0.746	0.456
Voiceus	0.083	0.061	1458.254	1.365	0.173
StimSingysingy	0.010	0.134	41.224	0.075	0.941
USorientation.sc	0.010	0.066	41.638	0.153	0.879
ConditionNoise:Musiciany	0.158	0.069	4549.820	2.286	0.022
ConditionSilence:Musiciany	0.116	0.069	4535.417	1.683	0.092
SubBlock2:Trial50.sc	-0.023	0.027	4546.485	-0.859	0.390
SubBlock3:Trial50.sc	0.076	0.027	4536.914	2.853	0.004
ConditionNoise:Voiceus	0.153	0.080	4531.100	1.913	0.056
ConditionSilence:Voiceus	-0.025	0.079	4533.279	-0.318	0.751
ConditionNoise:StimSingysingy	0.016	0.079	4529.108	0.200	0.841
ConditionSilence:StimSingysingy	-0.135	0.077	4536.358	-1.759	0.079
Voiceus:StimSingysingy	-0.049	0.076	4524.548	-0.651	0.515
ConditionNoise:USorientation.sc	-0.049	0.038	4511.689	-1.283	0.199
ConditionSilence:USorientation.sc	<0.001	0.038	4532.921	0.010	0.992
Voiceus:USorientation.sc	-0.074	0.038	4505.867	-1.968	0.049
ConditionNoise:Voiceus:StimSingysingy	**-0.010**	**0.109**	**4544.284**	**-0.094**	**0.925**
ConditionSilence:Voiceus:StimSingysingy	**0.237**	**0.107**	**4544.812**	**2.217**	**0.027**
ConditionNoise:Voiceus:USorientation.sc	**0.116**	**0.054**	**4517.856**	**2.154**	**0.031**
ConditionSilence:Voiceus:USorientation.sc	**0.039**	**0.053**	**4545.366**	**0.737**	**0.461**

The significant effects in this model, as shown in Table 4.4, are listed below, along with references to relevant interaction plots.

- Main effects with no interaction:
  - There are significant main effects for Length.sc, Freq.sc and Gender showing the same patterns as were seen in LDTmod2, fast responses to short and frequent stimuli, and slower responses from male participants. In addition, there is a highly significant effect of PrevRT.lsc. If the previous trial had a slower response, the present trial is also more likely to have a slower response.
- Two-way interactions:
  - Condition * Voice — The Lexical Access Hypothesis is once again supported. This time, however, it is embedded in two three-way interactions.

The co-efficients for the two-way interaction shown in the model output table are therefore only relevant to the reference level/value of StimSingy (notSingy) and USorientation.sc (0), respectively. The two-way interaction must be interpreted in the context of its higher-order interactions, which are presented below, and bolded in Table 4.4.

– Condition * Musician — Musicians have a speed advantage over non-musicians when responding in a musical context. This interaction is shown in Figure 4.6.

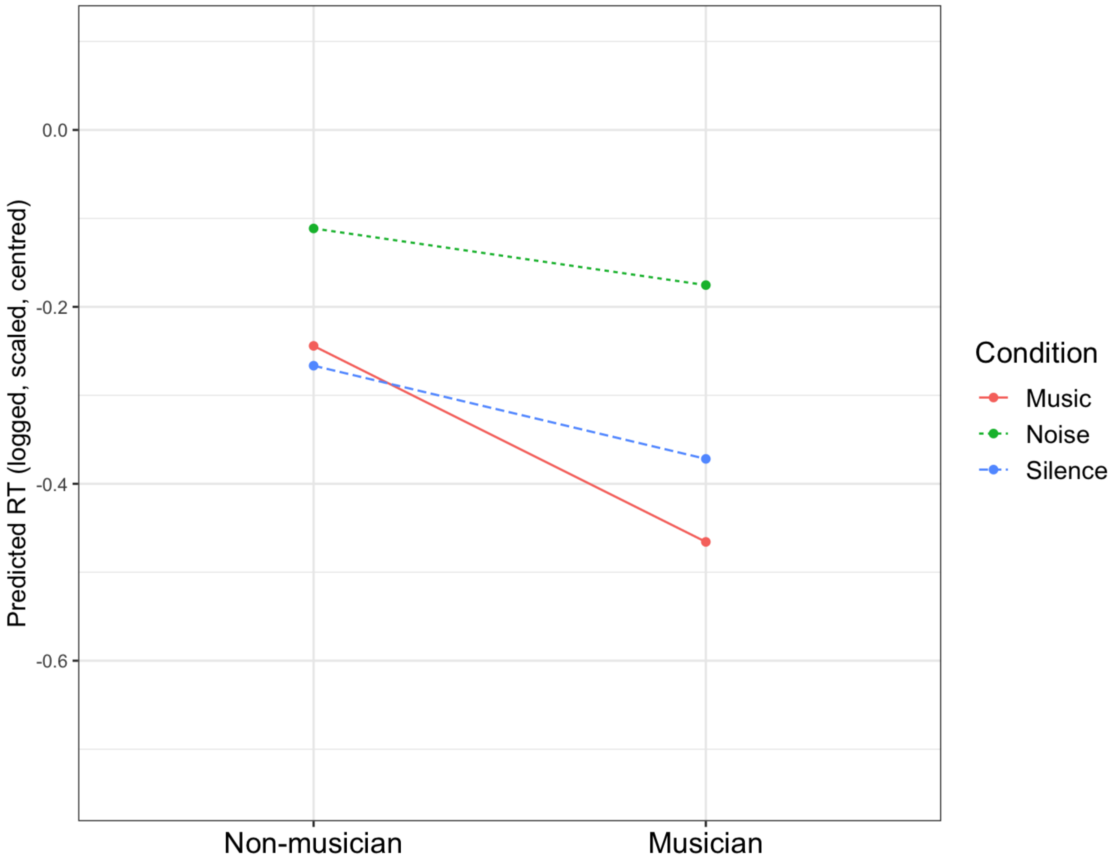


Figure 4.6: Interaction of Condition with whether or not the participant is a Musician.

– Condition * Songiness — This predicted interaction was nowhere near significance, with a log-likelihood comparison p-value of 0.93 after the interaction was dropped, confirming that the larger model including the interaction was not justified.

– Songiness * USorientation.sc — This predicted interaction did not reach significance. However, the co-efficient for the interaction of Songiness with Condition when comparing Noise to Music had a p-value of 0.2, and the direction of that co-efficient suggested that US-oriented listeners were faster to songy than speechy words. The log-likelihood comparison of the models with and without this interaction had a p-value of 0.33, however, confirming the non-significance of this term.

– Trial50 * SubBlock — This interaction was significant, with the same pattern as that seen in the prior models. Participants tend to get faster as they go through each half of the experiment, but slow down in the last block of each half. This interaction is shown in Figure 4.7.

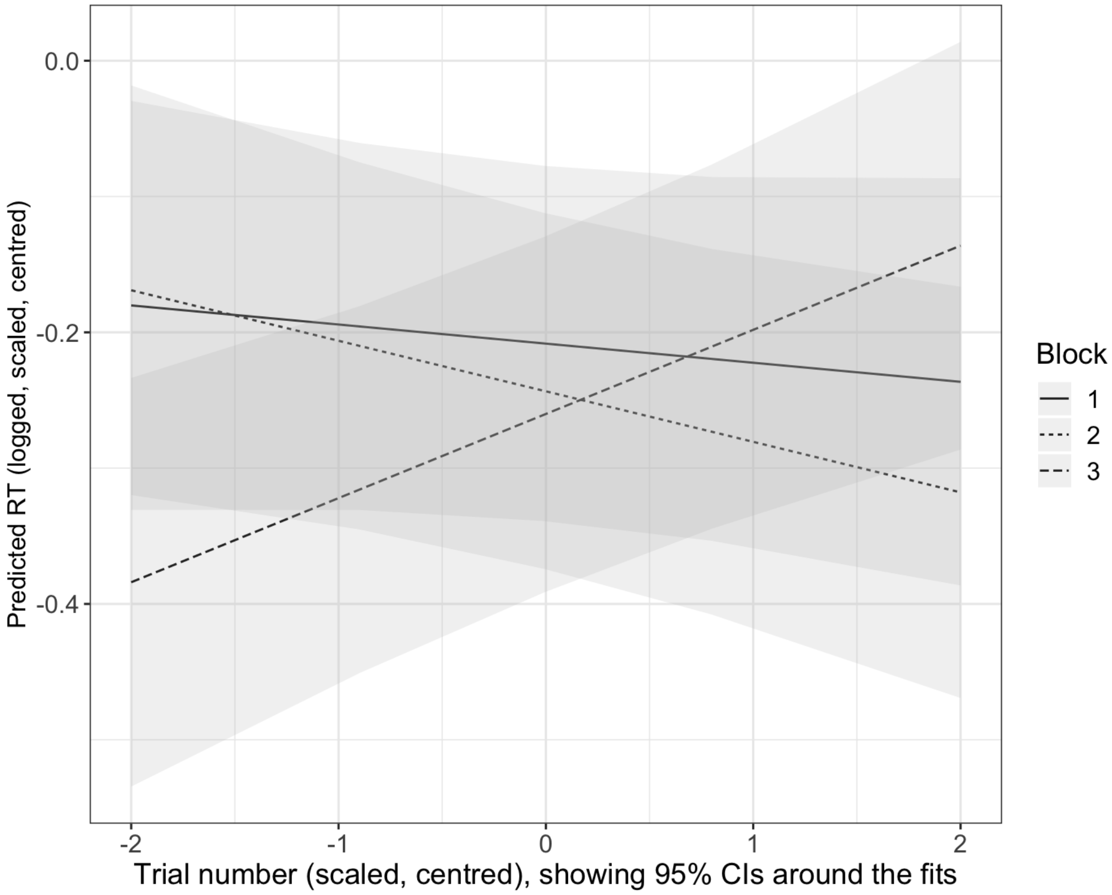


Figure 4.7: Interaction of Block number with Trial number.

- Three-way interactions:
  - Condition * Voice * BlockHalf — This interaction was dropped for reasons of multicollinearity. When included in the initial, maximal model, the highest VIF (for Voiceus:BlockHalf2) was 53. Once the interaction was dropped, the problem of high VIFs was solved.
  - Condition * Voice * USorientation.sc — This interaction was significant when comparing Noise to Music, and in the expected direction, with US-oriented listeners having a greater facilitation to the US voice in Music than NZ-oriented listeners. It should be noted that it is only the difference between Noise and Music that is carrying this interaction; the comparison of Silence to Music in this three-way interaction is non-significant. The interaction is shown in Figure 4.8, including all three conditions, but any difference in slope between Music and Silence across participant groups should be interpreted with caution, since they are based on a non-significant term in the model. Noise and Silence are not significantly

different to one another in the context of this interaction (Condition-Noise(ref=Silence) * VoiceUS * USorientation.sc: Estimate = 0.0765; p = 0.15). As can be seen by comparing these values to those in the Table 4.4, however, the difference between Noise and Silence is actually closer to significant than is the difference between Music and Silence. This provides further cause for caution in the interpretation of this interaction.

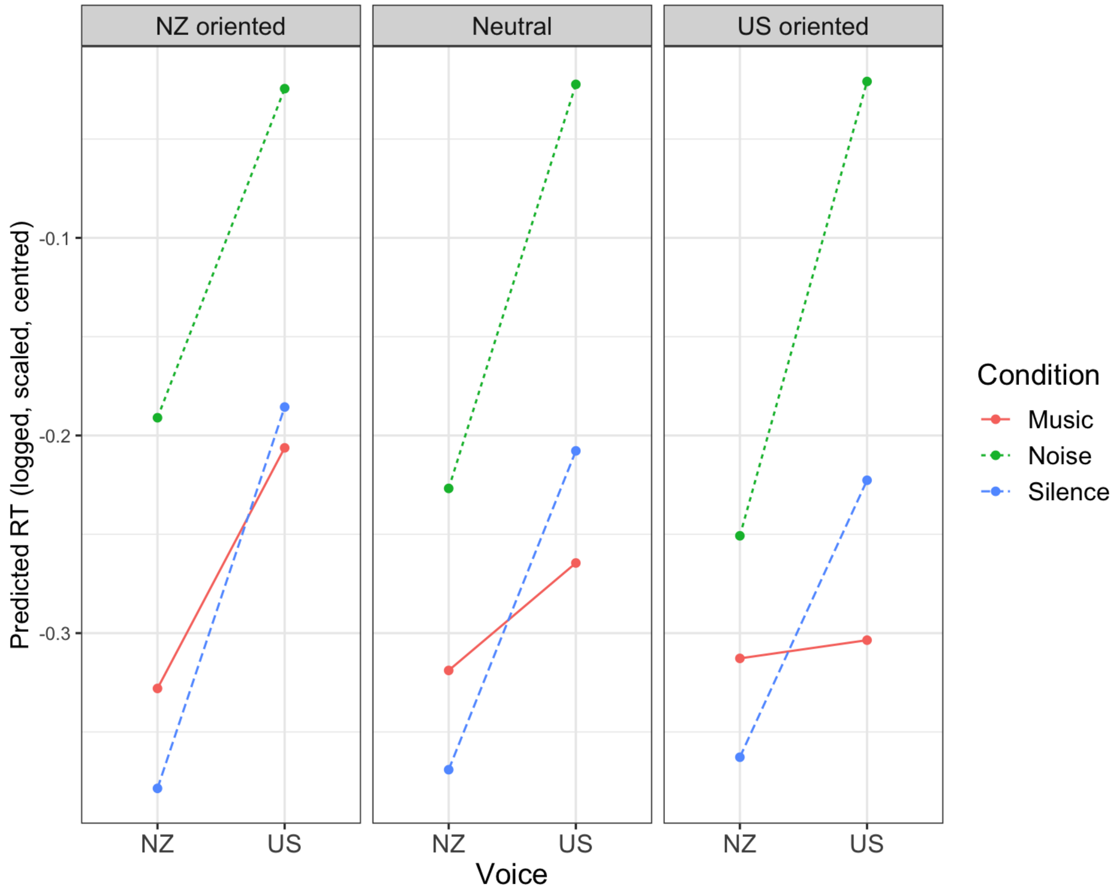


Figure 4.8: Interaction of US orientation with Condition and Voice in LDT Model 4, showing Q1, median and Q3 of USorientation, labelled as NZ oriented, neutral and US oriented participants, respectively.

– Condition * Voice * StimSingy[12] — The facilitation for the US voice in Music is weaker for participants who did not think the voices sounded like they were singing. In this case, it is the difference between Silence and Music that is carrying the interaction, shown in Figure 4.9. Differences in slope between Noise and Music across participant groups should therefore be interpreted with caution. Also problematic to any interpretation of

[12]Note that due to my concerns about over-fitting when performing binary splits on the participants, I ran another full model fitting procedure using a continuous version of StimSingy, despite this scale being very non-normal, with most responses being either 1 or 2 on the Likert scale. All of the interactions shown in the final model were also significant in the model which resulted from that procedure. That model, in fact, ended up being identical except for minor differences in random effects structure.

this interaction is the fact that a version of the model with reordered levels of Condition revealed a significant difference between Noise and Silence (ConditionNoise(ref=Silence):VoiceUS:StimSingysingy; Estimate = -0.248, p=0.022). This can be interpreted as follows: For those who thought the stimuli sounded like singing, the US voice attracted relatively faster responses in Noise than in Silence, just as it did for Music vs. Silence. This weakens the potential for any claim that this is particularly about music rather than masking. The result (when considering those who found the stimuli to sound like singing) is reminiscent of the accuracy results, where we saw accuracy to the NZ voice being hurt to a greater extent than the US voice by masking of either type (Music, Noise).

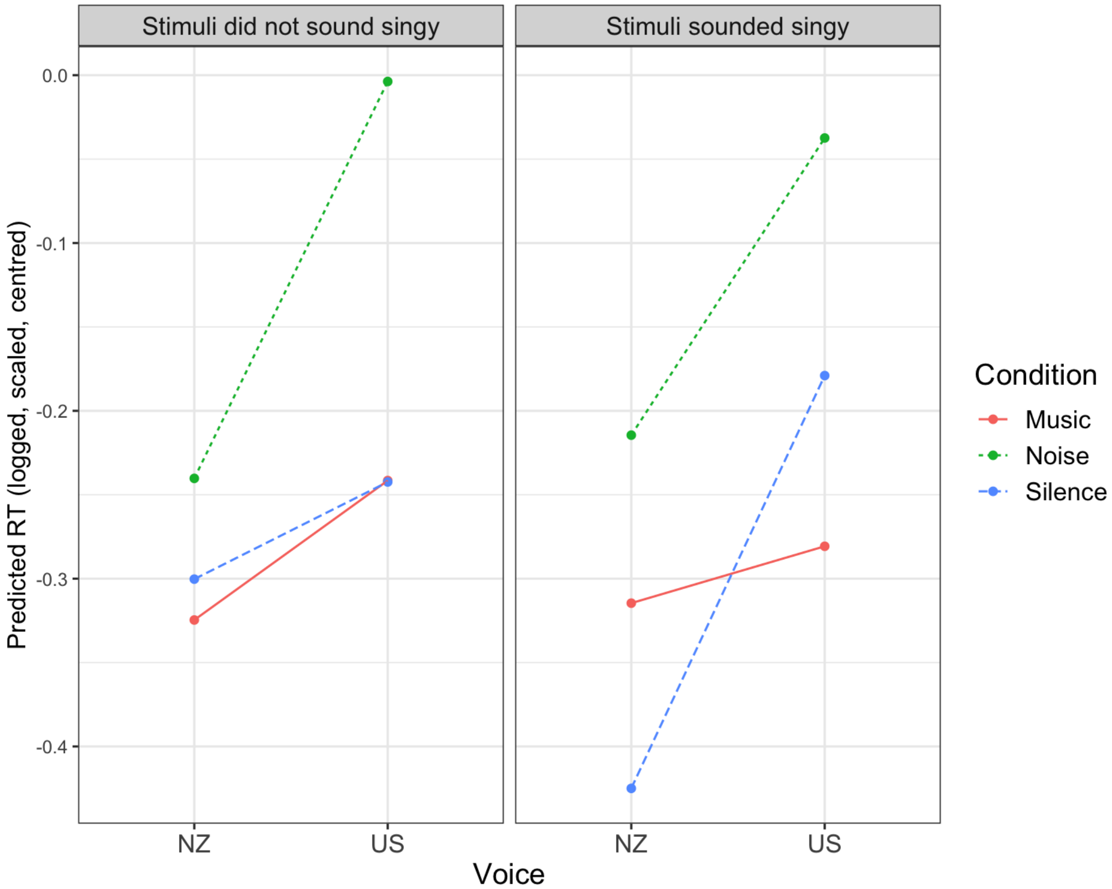


Figure 4.9: Interaction of whether the stimuli sounded like singing or not with Condition and Voice in LDT Model 4, showing participants who circled 1 (didn't sound like singing) on the left and those circling 2–4 (sounded somewhat like singing) on the right.

4.3.3 Perception of BATH, rhoticity, LOT, and GOAT

Before concluding the results section of this chapter, I present here a brief exploratory analysis of the subsets of data relating to the four sociolinguistic variables that were analysed in Chapter 2, and included in all of the stimuli for this experiment.

Simple mixed effects models were tested for each of the subsets of data for the four sociolinguistic variables included in the stimuli. The models had as fixed effects the interaction of Condition by Voice, and a main effect for the length of the soundfile and the sung frequency of the word, with random intercepts for Subject and Word. The RT result discussed above appears to be carried most strongly by the rhoticity variable. The Condition by Voice interaction is significant in that subset, with a significant facilitation of the US voice in Music vs. Silence, and trending in the same direction for Music vs. Noise.

The interaction of Condition with Voice does not significantly improve model fit for BATH or LOT, but does for the GOAT data, where there is a facilitation for the US voice in Music vs. Noise, but no clear trend for Music vs. Silence.

Since the vowel variables all have less data than the rhoticity variable, another model comparison was done for the BATH, LOT and GOAT data combined. In this model, the interaction of Condition with Voice improves model fit near-significantly (log-likelihood $p=0.075$). There is significant facilitation for the US voice in Music vs. Noise (co-efficient t-value = 2.26), and while non-significant, the difference between Music and Silence is in the expected direction ($t=0.91$).

These results fit well with those presented in Chapter 2, confirming the status of rhoticity as a key marker of the US singing style, but challenging the idea that BATH is the shibboleth I had thought it to be.

4.3.4 Further investigation of participant differences

During modelling, various participant groupings were attempted but modelling was limited by the complex nature of the hypothesis. To analyse differences specifically with respect to whether or not participants supported the Lexical Access Hypothesis, each person's data was distilled into a single number. This calculated their mean facilitation for the NZ voice over the US voice in each condition, and then calculated whether that facilitation was lower in the Music condition than in Silence and Noise. Positive values mean the Lexical Access Hypothesis was supported: NZ facilitation was less in music than non-music. Across the 36 participants, the average of these values is 29.4, reflecting the overall finding that the hypothesis was supported. A total of 21 participants had positive values and 15 had negative values. This value was then used as a dependent variable in a series of linear regressions to test for participant effects. The best model had just one significant IV. If a participant listed classical and/or jazz music as a favourite genre, they were less likely to support the Lexical Access Hypothesis.

4.4 General Discussion of LDT results

By looking at the results through a range of analyses, a clear picture has emerged. The various perspectives will be drawn together in the next chapter, and discussed with reference to the wider research questions of the thesis, and alongside the results from Chapters 2 and 3. Given the inclusion of many points of discussion already presented above, and the general discussion still to come in the next chapter, I will keep my comments in this section brief.

> The [exemplar] model is consistent with the standard assumption that

> reaction times for phonological and lexical decisions reflect the time required for activation to build up and cross a decision threshhold. Thus, the model is consistent with, and can even serve to elucidate, results on the speed of phonological and lexical decisions.(Pierrehumbert, 2001, p. 6)

As a reflection of activation, the results of this lexical decision task provide clear support for the role of context-relevant memories in raising the availability of congruent lexical items for processing. New Zealand listeners were faster to a US voice than to their native dialect in the context of song. If a 'native' dialect is the one most deeply and completely learnt, then these results encourage us to consider: a) the role that music has played in the sound experiences of these individuals; b) the role that context plays in the structuring of their memories of language; and c) whether these NZ participants may actually be native listeners of SPMSS, despite perhaps only rarely engaging in song production themselves.

Unlike the strong results for RT, the models determined that there was no significant interaction between Condition and Voice for accuracy. This may be partly due to a ceiling effect, and perhaps the slightly more discrete nature of accuracy rates as compared to reaction time. As mentioned earlier, even though Walker and Hay (2011) did find a significant congruence effect for both RT and accuracy, results were more consistent and robust for RT than accuracy in the experiments reported by Hay et al. (2019). Note that in the raw results for accuracy we see the importance of including Noise as a control condition, not just Silence. Without it, there would have appeared to be facilitation of accuracy for the US voice in Music. However, the same US facilitation also applies in Noise as compared to Silence, and may have been related to the relative clarity of the two voices in a masked context.

The final model brought some interesting inter-participant differences to light: those who listened mainly to US music, and a lot of it, supported the Lexical Access Hypothesis most strongly. The careful statistical approach taken in LDT Models 1–4 limited my ability to explore individual variation, but further investigation of participant differences singled out jazz and classical music listeners as particularly unlikely to support the hypothesis. These findings suggest that consumption of commercial pop is crucial for a 'native-like' acquisition of SPMSS.

While the findings about music listening patterns foreground the role of experience, another interaction foregrounds the role of salience. Participants who experienced the illusion of listening to singing in the music conditions, even though the stimuli were identical across conditions, were more likely to support the Lexical Access Hypothesis. In Chapter 1, I argued that a listener's exemplar space is activated in a way which best matches the current context. Participants who had the subjective experience that the voices were singing, rather than speaking (cf. Falk et al., 2014), may have thus activated more strongly their memories of song, resulting in greater activation of phonetic variants that matched the US voice. Taking these two interactions together, the findings connect the historical component of individual experience — the encoding of memory — to the reinstantiation (Danker and Anderson, 2010) of those memories by the context.

4.4.1 Directions and extensions

There are various ways the results from this experiment could be strengthened. As was mentioned in the procedure section, multiple cues were given to prime the idea that participants would be encountering singing in the music conditions. Additionally, the fact that the voices were labelled with nationalities and blocked together, rather than randomised within blocks, allows listeners to form a coherent and persistent frame of reference for that voice. Thus, the experiment may be tapping into higher-level expectations associated with dialect labels and abstractions. The difference between the abstract and episodic contributions to the effect could potentially be disentangled by presenting the voices in randomised rather than blocked order, so that it is harder for participants to maintain persistent ideas about who is speaking. If the effect disappeared, this would tell us that it is driven at least to some extent by concepts at a more general level than exemplars of individual words.

The significant result presented in this chapter begs to be replicated in other contexts. There are several paths this could take. Different participant populations would offer opportunities to explore the issues studied here. Comparing the results of Chapter 2 to the findings of O'Hanlon (2006) suggests that own-accent rap is much more prevalent in Australian than NZ hip hop. Would Australian listeners show the RT facilitation to a US voice at different rates depending on their amount of hip hop they have listened to? Could the use of hip hop instrumentation abolish the effect? With respect to genre, a range of extensions are possible. For example, listeners may have facilitated processing of a Jamaican voice in the context of a reggae music background (see Gerfer, 2018), while a Southern British English voice might be easier to process in the context of choral music (see Wilson, 2017).

Beyond behavioural tasks such as the one presented here, it could be revealing to measure event-related brain potentials in response to phonetic stimuli that are or are not congruent with a musical prime. In the final section of this chapter, I will explore this possibility.

4.5 The Potential of Cognitive Neuroscience in a Sociophonetics of Popular Music

Behavioural experimentation is limited in two fundamental ways. Firstly, it requires a task, which is usually not ecologically valid (categorising words from nonwords is not a task people engage in very often when processing language in everyday circumstances). Secondly, it measures a point in the language perception process which is quite delayed, it cannot assess the subtleties of early stages of auditory processing and lexical access. In the final section of this chapter, before drawing together the production and perception findings of this thesis in Chapter 5, I will briefly present the results of a pilot study which suggests it may be possible to modify the LDT and analyse event-related potentials (ERPs) in future work. First, I present some of the most relevant literature in this area.

4.5.1 Measuring congruence through event-related potentials

The development of techniques to measure ERPs has allowed the study of context effects in memory to flourish. Kutas and Federmeier (2011) review a compelling range of evidence for the role of expectancy and congruence in lexical access. The review article focuses on one of the most commonly studied ERP (event-related potential) components, the N400 (a negative peak of electrical activity reaching its maximal amplitude about 400 ms after a stimulus), which is greater when encountering a semantically incongruent or unexpected word. The N400 is inversely proportional to the cloze probability of a word, and thus proportional to the informativity of a word (Shannon, 1948).

Van Berkum et al. (2008) is of particular relevance to the present project. A greater N400 effect was found for words that are incongruent with speaker characteristics, on several dimensions including the age, sex and perceived class of a speaker. For example, the sentence ‘I can’t go to sleep without my teddybear in my arms’ elicited a greater N400 to the word *teddybear* when spoken by an adult than by a child.

A range of other studies have found related results using a range of methodologies, and analyzing a range of ERP components. Foucart et al. (2015) found a posterior late positive potential for an incongruent speaker vs. a congruent speaker. Martin et al. (2015) found a very late negativity (in the 700–900ms band) for lexical access when vocabulary was not congruent with accent, such that a greater negativity was found for example to the word *vacation* in a sentence spoken by a British voice and a greater negativity to the word *holiday* in an American voice. Loudermilk (2013) found N400-like responses to a mismatch in the sociolinguistic variable (ING), such that a greater negativity was found for a Southern voice using the velar variant than the alveolar variant, and vice versa for a Californian voice. Conrey et al. (2005) analysed ERP measurements of participants with and without the PIN/PEN merger in US English. They found that speakers of the merged dialect had a reduced late positive component to incongruent stimuli than did those participants for whom the vowels are distinct.

In another study which investigated the role of habituation to surprising stimuli, Nieuwland and Van Berkum (2006) provide evidence that N400 incongruity effects can fade as a listener becomes accustomed to a context violation. In short stories which violated the animacy requirements of verbs, for example with statements like ‘the yacht cried’, they found that while the first sentence of this kind in a short story elicited a significant N400, the surprisal effect quickly diminished and was absent by the fifth repetition. A further study showed the reverse scenario. By creating a discourse appropriate understanding of the inanimate object, no N400 was elicited upon encountering the contextually-appropriate statement ‘the peanut was in love’, while an N400 was evoked by ‘the peanut was salted’. Thus, expectation is updated according to discourse context.

The N400 has been shown to reflect semantic processes, be they linguistic or otherwise (as reviewed by Kutas and Federmeier, 2011). It is elicited, for example, by pairs of incongruent pictures. Extending this research paradigm to music, Daltrozzo and Schön (2009) recorded EEG from participants as they carried out a visual lexical decision task. Each word or nonword was immediately preceded by a

1 second excerpt of music. For the real words, this excerpt was either congruent or incongruent, with a range of excerpt-word pairs initially suggested by expert musicians and then selected for use in the experiment according to agreement levels of ratings by non-musicians. Music–word incongruence was associated with a (late) N400 effect, supporting earlier findings that had used longer (10s) extracts of music (Koelsch et al., 2004).

Taken together, these studies provide evidence that congruence effects are specific to expectations about speaker styles and are based on the language an individual has encountered, at phonological, lexical and semantic levels. Furthermore, our expectations shift dynamically with our understanding of the current situation. Different linguistic levels appear to be associated with different ERP components, and there are complicating factors such as habituation to novelty (van den Brink et al., 2012).

The above studies look into the effect of incongruence on event-related potentials in a range of ways. Knowledge of phonetic differences between singing and speech could potentially be assessed with N400 measurements, hypothesising that a sung NZ accent would be associated with ERP components found in prior studies to be markers of voice–content incongruence.

4.5.2 Auditory processing of words in music: An ERP pilot study

Doing a full-scale ERP experiment was out of the scope of this thesis project, but some inroads were made. Five participants completed an early version of the LDT with their EEG signal recorded. Unfortunately, issues with the timing of the stimuli meant that this data could not be analysed for facilitation to the US voice in the music condition (or surprisal to the NZ voice in music). But this pilot data does lay some groundwork as a proof of concept for running an experiment like this in future. It is unclear from the literature whether auditory ERPs to words can be recorded when there is music playing concurrently. There is consensus that the N1–P2 auditory response is later and of reduced amplitude for words presented in noise, but no similar studies have been found for the N1–P2 response to words heard in music.

EEG recordings for the five pilot participants used the standard 10–20 layout, with a 32 channel electrode array recorded through the BioSemi system, which also collected time-stamp information from E-Prime. Reference electrodes were placed on the earlobes. While three of the participants had poor quality signals, the clean data from the remaining two participants was averaged across the two voices in each condition and time-aligned to the start of the stimuli.

Figure 4.10 shows the grand average auditory response for 600 trials. The Silence condition shows a typical N1–P2 complex, which is generated in the primary auditory cortex in response to the onset of a sound. For a review of literature on N1 see Näätänen and Picton (1987) and for P2 see Crowley and Colrain (2004). The Noise condition (blue) shows the delayed N1 and P2 that have been reported in prior studies. In the Music condition (red), the N1–P2 complex seems to be delayed to a similar degree as it is in Noise, with an even greater reduction in amplitude. This data from two participants suggests that with a longer experiment with a large number of trials, it may be feasible to detect N400 even in the context of ongoing music.

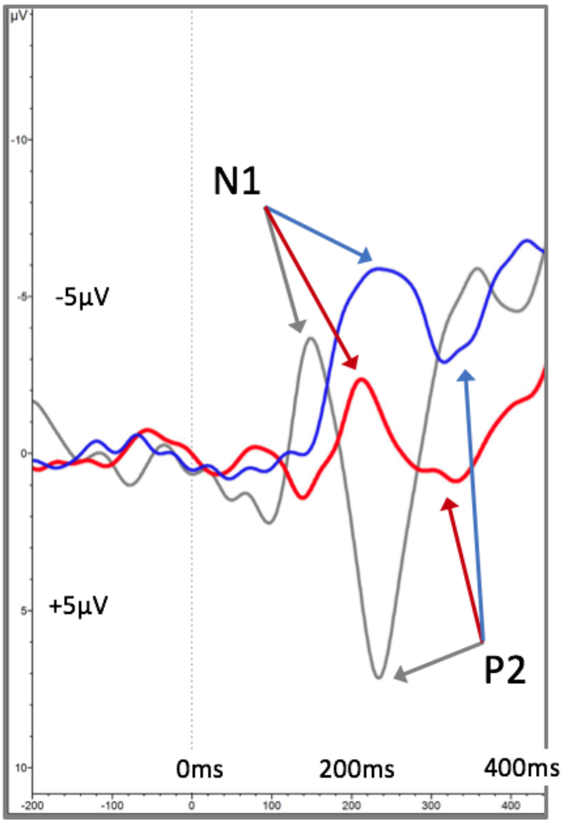


Figure 4.10: Grand average auditory event-related potential responses for 100 trials across three conditions for two participants (600 trials total). Responses during the Silence condition are in grey, the Noise condition is shown in blue, and responses during the Music condition are shown in red. N1 and P2 for each condition are labelled.

What is particularly promising about the ERP methodology is that there is actually no need for the LDT. Participants could simply be presented with the words in different contexts. There is also scope for EEG to measure salience pop out effects, which should be associated with a P3a surprisal response (Polich, 2007). By measuring EEG while participants simply listen to words, with attention (for example involving sporadic memory tests) or without attention (watching a silent movie), we can assess the extent to which certain words are more expected, or those which violate expectations and cause surprisal responses. Using the same stimuli as those in the lexical decision task, but without the nonwords or the requirement for response, measures of P300 and N400 responses may provide evidence that contextually specific phonetic expectations can affect early stages of lexical access.

With this modest addendum to the lexical decision task, I draw to a close the empirical part of this thesis. Across three chapters, I have presented various types of evidence that when occurring in song, both speech production and speech perception are powerfully affected. Drawing the findings together, we can now return to the question that sparked this whole endeavour, and consider whether we are closer to understanding why New Zealanders might find it difficult to sing the way they speak.

Chapter 5

General Discussion and Conclusion: Language Style as Memories in Context

The basic finding of this thesis will be common sense to most readers, at least upon some reflection: New Zealand vocalists predominantly sing in US-derived accents. Due to its dominance, and given the fact that it involves levelling of a wide range of US dialectal differences, I have referred to this style as the Standard Popular Music Singing Style (SMPSS), rather than continuing to describe it as American.

The adoption of SPMSS by New Zealand artists comes as no surprise, but the other core argument of this thesis is less intuitive. This relates to the motivations of those singers, and the processes leading to this outcome. I have argued that the SPMSS variants are part of a responsive and automatic language style, and in this final chapter, I further my argument that this automaticity is due to the tendency of the language system to hone in on context-relevant exemplars of words. Evidence for this perspective comes from the perception experiments presented in the preceding chapters, which found that New Zealanders expect SPMSS in song and that they adjust their perceptual vowel space accordingly. They also find it easier to process a US voice in a music context where it is contextually congruent. Taken together, these findings suggest that New Zealanders are 'native-like' in SPMSS when listening to popular music. The PoPS corpus revealed exceptions to the SPMSS norm, in the form of New Zealand English features, which were described as involving conscious styling. These initiative styles, I argue below, can also be well accounted for within an exemplar theory approach.

In this final chapter, I will summarise the results of the corpus study and the perception experiments, in Sections 5.1 and 5.2, respectively, bringing both together in Section 5.3. Turning to sociolinguistic themes in Section 5.4, I propose that some New Zealanders may be bidialectal, with a native-like command of both NZE and SPMSS. I then look again at the debate around media influence in sound change and what a sociophonetics of popular song might contribute, before considering the key concepts of indexicality, authenticity and salience in light of the results. Finally, I will return to the role of memory in our experience of songs, bringing all of these themes together in a discussion of how a sociophonetics of popular song can be conceived of within exemplar theory, in Section 5.5.

5.1 Summary of Corpus Results

To study the influence of music on speech production, the Phonetics of Popular Song corpus was created, including a sample of NZ and US pop and hip hop vocalists, balanced in each location by ethnicity, age and gender. I begin this section with a recap of the key findings from the analysis, before stepping through how those findings relate to the Dominance, Accuracy, Genre and Salience Hypotheses. The majority of NZ artists analysed in the PoPS corpus were shown to embrace SPMSS across all of the variables studied in detail: BATH, non-prevocalic /r/, linking /r/ and LOT, and also in the variables given a more cursory analysis: GOAT, DRESS and TRAP. They generally did so with quantitative accuracy to the model, though there were also exceptions. A few rappers evidenced a clear intention to adopt their 'own accent' across multiple variables. Other artists in both pop and hip hop, that showed a moderate desire to use New Zealand English (NZE), tended to do so on the more salient variables (realisation of BATH with PALM and avoidance of rhoticity), and at sites of contextual prominence. No strict salience hierarchy between variables could be established, however. Examination of how features clustered together across different artists showed that patterns of clustering are themselves subject to indexical processes. There is an apparent trend, for example, for young pop artists to use SPMSS variants for BATH and rhoticity in combination with NZE (or perhaps British English) variants for LOT. That variant may be a case of own-accented singing, or a change in SPMSS reflecting a mainstreaming of influence from British English.

Linking /r/ differs from other variables in its high rate of realisation in both European American and Pākehā speech styles, whilst occurring at lower rates in African American and Māori/Pasifika speech. Linking /r/ varied more by ethnicity than genre, contrary to the Genre Hypothesis, with African American and European American vocalists differing from each other in both pop and hip hop. A difference between Māori/Pasifika and African American vocalists was also found. While both groups avoid linking /r/, they do so in ways that reveal different approaches to hiatus resolution, with African American artists using the strong boundary marker [ʔ] and Māori/Pasifika artists using high rates of VV sequences, that is, showing a tolerance for vowel hiatus. These approaches are at opposite ends of the continuum of hiatus resolving strategies outlined by Uffmann (2007). This distinction would be hidden by a binary analysis of linking /r/ as simply present or absent. The results supported recent findings that the avoidance of linking /r/ is a feature of young Pasifika Englishes in New Zealand (Gibson, 2016). Hiatus resolution is strongly constrained by stress and rhythm, and the analysis revealed how songwriters and rappers generally align spoken stress patterns with musical beat structures.

An auditory analysis of the GOAT vowel showed that the fronting variant of this vowel, characteristic of spoken New Zealand English, is all but absent in NZ song and rap. It tends to be used only by those who also employ more salient features. Overall, the Accuracy Hypothesis was largely supported, suggesting a native-like command of SPMSS by NZ artists. The F1/F2 results for the LOT vowel (after removing four singers with an objectively determined intention to use NZE), and the F1 results for DRESS and TRAP showed that NZ singers are statistically indistinguishable from US singers. There was also, however, some counter-evidence to this hypothesis, with Pākehā female pop singers using more linking /r/ than their European American counterparts. This was interpreted as an intention to display rhoticity. Linking positions may provide a phono-opportunity (Coupland, 1985) for this display. Alternatively, the result may reflect an additive effect of linking /r/ being present in both NZE and SPMSS. To summarise the results with respect to the hypotheses of the corpus study:

- Dominance Hypothesis: SPMSS will be prevalent in NZ pop and hip hop. This was strongly supported — SPMSS is the norm in NZ pop and hip hop.

- Accuracy Hypothesis: when adopting SPMSS, NZ performers will be accurate to the model. This hypothesis was supported by the results for non-prevocalic /r/, LOT, and also DRESS and TRAP, though there were hints of an intention to display SPMSS /r/-fulness by female Pākehā pop singers' overshot rates of linking /r/.

- Genre Hypothesis: pop will be homogeneous while hip hop styles will reflect the speech communities of performers. This third hypothesis was generally supported. Pop music showed a high degree of homogeneity with respect to singers' ethnic and geographic backgrounds, while hip hop had greater diversity, as predicted. However, there were also cases where ethnic identity came through in pop music, particularly for non-prevocalic /r/, where both African American and Māori/Pasifika pop singers appeared to adopt speech community patterns rather than SPMSS. This may reflect a genre distinction that inadvertently occurred in the selection of these ethnicities.

- Salience Hypothesis: artists will better perform their identity goals on sociolinguistically salient variables and at contextually salient sites. This hypothesis required a bootstrapping approach to be tested, so results are inherently exploratory. However, there was basic support for the idea. NZ identity appeared to be expressed more through the avoidance of rhoticity and realisation of BATH with PALM than it was through rounded LOT or fronting GOAT. The latter only attracted NZ variants from artists who also used NZE on BATH and avoided rhoticity. The only exception was Youmi Zouma, the one pop group who could clearly be described as 'indie'. This finding thus supports those of Coddington (2004), that NZE GOAT is a feature of alternative/indie music in NZ. Several anecdotal examples also pointed to greater use of NZE at sites of contextual salience such as in lower frequency words or at the end of a musical phrase, providing some new support for the ideas put forward by Yaeger-Dror (1991).

Overall, the findings strengthened the claims of Gibson and Bell (2012) that American-derived forms are so normative in popular music that it requires intention for NZ singers and rappers to use their spoken dialect in song. The question which frames this thesis asks 'why is this the case?'. In Chapter 1, the global spread of US popular music was described with reference to cultural imperialism, and the uptake of AmE by non-US singers was likened to the spread of the English language through nineteenth century colonial expansion. Against this backdrop, SPMSS developed into a form which is not taken up by singers alone, but which is also embedded in the experiences of listeners.

5.2 Summary of Perception Results

The perception experiments were designed to explore this phenomenon from the perspective of the general population of music-listening New Zealanders. Both experiments were preregistered, including detailed methods for statistical analysis of results. Chapter 3 presented a phonetic categorisation task (PCT). Thirty-six native speakers of New Zealand English listened to stimuli that ranged from *bed* to *bad* in the context of Music, Noise and Silence, and decided which word they heard.

- Reference Frame Hypothesis: listeners will classify ambiguous stimuli as *bad* more often in the speech conditions, reflecting expectations of the raised NZE short front vowels. They will respond *bed* more often in the Music condition, in expectation of the more open short front vowels characteristic of SPMSS.

There was significant support for the Reference Frame Hypothesis in all aspects of the data analysis. These findings show that listeners are primed by a pop music context to expect SPMSS. Participants expected more open vowels in singing than speech for the words *bed* and *bad*, in line with SPMSS and not NZE. This result occurred despite participants having been told that the speaker/singer would be a New Zealander. The analysis of F1 in DRESS and TRAP vowels in the PoPS corpus provided clear motivations for participants' expectations. NZ spoken TRAP is almost entirely overlapping with sung DRESS (as realised by a singer of either NZ or US origin). A vowel in this F1 region is thus highly ambiguous if there is no contextual information to guide the listener. The result was especially strong at the start of the experiment, and gradually reduced. This is interpreted as follows: in the absence of a voice, the macro-context is used to determine expectations. Once speech is encountered, it provides more reliable cues for the prediction of upcoming words. Thus, as listeners tuned into the phonetics of the vowel continuum itself, they became less affected by the presence of music.

The distinction between DRESS and TRAP rests mainly on an F1 difference, and the result could therefore conceivably have been based on a simple rule in the minds of listeners that singing involves a more open jaw setting than speech. To determine that the results are actually based on expectations specifically about dialectal norms, a lexical decision task (LDT) used naturally recorded stimuli. A US and a NZ speaker produced words that contrasted on a range of vowel and consonant variables which distinguish NZE from SPMSS, and are independent from jaw opening. Participants classified words and non-words that were heard in Music, Noise, or Silence.

- Lexical Access Hypothesis: participants will be faster and/or more accurate to the US voice when it occurs in music and slower/less accurate to the NZ voice in the Music condition.

Through a series of analyses, the Lexical Access Hypothesis was robustly supported with respect to reaction times, while no support for the hypothesis was found in the accuracy data. Replicating prior research, participants were much faster to correctly identify a stimulus as a real word when it was spoken in their native dialect, NZE, in the typical experimental conditions of background noise or silence. Participants were faster, however, to the US voice when it occurred in a musical context.

This interaction held up across multiple data analyses, providing significant support for the Lexical Access Hypothesis with respect to reaction time.

In both experiments, a range of strategies were used to convince participants that they would hear singing in the Music condition. In the Music conditions, the instructions explicitly told participants they would hear singing, the stimuli were presented on the beat, and at a stable pitch which was tonally congruent with the music. These factors were designed to support the illusion of song, even though the vocal stimuli were identical in the music and non-music conditions. As shown by responses to the questionnaire, the illusion of song was not experienced equally by all participants. Importantly, the degree to which they reported hearing the stimuli as sung predicted their reaction times in the LDT. Rather than there being a solely mechanistic effect of background music on expectations, those who perceived the stimuli as sung had a stronger boost in facilitation for the US voice in music.

An interaction was also found between reaction times and the type and amount of music that listeners are exposed to in their daily lives. Those with overtly US-orientated music listening practices were more likely to display increased facilitation for the US voice in music. This result, more than any other, drives home the connection between the corpus and perception results of this thesis since the singers and rappers analysed in Chapter 2 are also music consumers. By exploring how music listening affects speech perception, we can gain insights about the patterns that may exist amongst an individual's episodic memories, and these patterns will have a central role to play in determining targets for speech(/song) production.

5.3 Bringing Production and Perception Together

In Chapter 1, I emphasised the importance of the relationship between production and perception to a sociophonetics of popular music. This section thus considers the results of the two parts of the thesis as a whole. Exposure to music in each person's 'life in sound' (Kraus and White-Schwoch, 2015) was shown in the LDT to play a role in speech perception. If a participant had listened to a lot of US music, and not much NZ music, the hypothesis was more strongly supported. As shown in Chapter 2, SPMSS forms the central tendency in the phonetics of NZ music, though exceptions also exist.

Exceptions to SPMSS are rare enough, and genre-specific enough, that they may only be encountered by a certain type of music-listener, who actively seeks out NZ music. The subset of participants in the LDT who have experience with NZE in song may be more able to quickly process the NZ voice in music. This may be in part due to stronger encoding of experiences with sung NZE, through a 'novelty pop-out' effect (Hay et al., 2018). Those listeners who have encountered no (or very few) instances of NZE in song, however, may have a harder time with this incongruent voice–context pairing, accounting for the overall significance of the result found in the LDT.

An exemplar theory account of these results would predict word-specific effects in both production and perception, particularly with respect to skews in lexical frequency that are structured clearly by context. It has now been well established that language users have such knowledge at least for skewed distributions of lexical frequency for speakers of different ages and genders (Walker and Hay, 2011; Hay et al., 2019). While those experiments specifically used ratios of lexical frequencies in their design, the studies presented in this thesis did not structure lexical frequency ratios into their methods. I did, however, consider the possibility of such effects during analysis.

In the corpus study, a near-significant effect of lexical frequency on rates of non-prevocalic /r/ was found, as discussed in Section 2.5.3.1. New Zealand artists were more likely to realise /r/ on words that are over-represented in song lyrics as compared to spoken corpora. While neither part of this thesis was designed explicitly to test effects of lexical frequency, such effects provide central evidence for the role of episodic memory in language processing and are thus of great interest. In the LDT, similar effects were searched for, with the expectation that facilitation for the US voice would be strongest on the most songy words. However, no effects of lexical frequency ratios were found. There are several possible reasons for this:

- Firstly, the words in the LDT were less songy overall than the words represented in the non-prevocalic /r/ data. The median songiness ratio in the LDT stimuli was 0.98, whereas the median ratio in the rhoticity dataset was 1.62.

- The hypothesis of the LDT is much more complex than the binary analysis of whether /r/ is realised or not. It depends on variation in reaction times with respect to an interaction of Condition with Voice spread across multiple blocks. Detection of a frequency effect in this task may thus require much higher statistical power.

- The LDT was not designed to test questions of skews in lexical frequency. In order to examine this issue, a balanced set of songiness ratios would need to be included in the design.

Despite these possible explanations, the absence of a result here could also be interpreted as evidence that word frequency is kept track of only *once per word*, i.e. independently of context. That is, the words *dancing* and *farming* may simply be treated as having the same frequency as one another, despite their different contexts of usage. Such a finding would provide evidence against the model used to explain all of the results in this thesis. There is evidence elsewhere, however, that our tracking of lexical frequencies includes information about relative rates of usage (Needle and Pierrehumbert, 2018). Since the LDT was not designed to test claims about this issue, I will continue to assume that we have the ability to learn that *dancing* is more frequent than *farming* in the context of song, even though they have the same frequency in speech (see Section 4.3.2.5).

By taking a joint approach to the production and perception of song, we gain insights that can reinforce each other. In this thesis, knowledge of production patterns helped to make accurate predictions about perception. An understanding of speaker-listeners' memories of music along with knowledge of their identity goals should also be able to predict their production. This idea will be explored in Section 5.5.

5.4 Connections to Sociolinguistics

In this section, I explore the results of this thesis with respect to relevant concepts and debates in sociolinguistics. I begin by proposing that the situation of SPMSS in

NZE may represent a case of bidialectalism, and consider this claim with reference to Hazen (2001). I then consider how the present results might speak to the debate on the role of the media in sound change. Sociolinguistic perspectives on authenticity, awareness and indexicality will lead us back to the role of memory in language representation.

5.4.1 NZE and SPMSS: Stable bidialectalism?

I stated above that the participants of the LDT appear to be 'native-like' in SPMSS when listening to a voice in the context of music. While I did not begin this project with bidialectism in mind, it seems it may be a central concept for exploring the sociophonetics of popular music in English-speaking countries outside of the USA. In one sense, being fluent in two dialects is analogous to bilingualism, while in another it can be viewed as a kind of intra-individual dialect contact, and thus subject to levelling processes (Chambers, 1992). In 2001, Hazen stated that '[i]n sociolinguistics, despite extensive work on language variation, no one has seriously investigated whether humans are capable of maintaining two dialects in the same ways they can maintain two languages' (p. 88). Ultimately he argues that they probably cannot. Noting that bidialectalism does not simply mean receptive knowledge of more than one dialect, Hazen argues that true bidialectalism would need to show mutually exclusive usage of a wide range of complex phonetic and phonological distinctions between D_1 and D_2.

The research presented in Chapter 4 suggests a grasp of such complexity in these NZ music-listener's D_2, SPMSS. Speeded reaction times in lexical decision are facilitated in a person's native-dialect. The flipping of the RT pattern between music and non-music conditions is therefore striking. Listeners appear to have more direct links to SPMSS than NZE sound structures in the context of music. The overall support for the Accuracy Hypothesis found in the corpus analysis, that NZ singers conform closely to US artists, further suggests popular music as a potential site of bidialectalism for New Zealanders. Hazen (2001, pp. 96–97) asked a series of questions that a researcher should ask if they suspect a person is bidialectal. I respond to each one below, in italics:

- Can a speaker fully acquire a second dialect and maintain the language variation patterns of the first dialect? *While this was not tested directly, it seems unlikely to me that NZ singers lose their command of NZE in speech. The same response applies to the next question.*

- Can a speaker who has acquired a second dialect in another region come back to the home region and continue to convince native speakers that the first dialect is authentic?

- If a speaker produces D_1 and then acquires features of D_2, will that speaker acquire those features with the qualitative and quantitative constraints as a native speaker? *This question has been explored at length in Chapter 2, and the signs are positive, though answering this question would require us to look at more complex phonological phenomena. Can New Zealanders with the* NEAR – SQUARE *merger unmerge accurately in song? Do they acquire the rule to tense* TRAP *before nasals, along with the lexical exceptions to that rule (Sneller et al., 2019)?*

- Will the speaker be able to switch between sets of dialect features instead of mixing linguistic features from two dialects in a single production? *For the majority of singers, the switch involves a very stark contrast. Mixing only seems to occur when NZ singers try to produce own-accent singing.*

- Can the speaker produce both sets of language variation patterns in unpracticed conversation? Even in practiced conversation? *This is an interesting question which could be a good diagnostic of whether NZ singers have a native-like command of SPMSS. Can they produce novel utterances in SPMSS (given a backing track to sing along to) or is their command limited to well-rehearsed passages.*

- Can a speaker switch more than sociolinguistic stereotypes? Can the speaker switch less salient markers or indicators? *Yes. This was clearly demonstrated in Gibson (2010b), and was the main evidence for the foundational assumptions of this thesis.*

There is one question that Hazen may not have thought to ask, but which seems relevant here. In a case of bidialectalism, do the pressures involved in maintaining mutually exclusive command of the two dialects lead to inhibitory effects? That is, could the two dialects be so contextually restricted that the speaker can only produce each dialect in its appropriate context. I add this question to Hazen's list:

- In a case where each dialect is restricted to use in a particular context, is the speaker unable to produce D_1 in $Context_2$, and vice versa? *This is the definitive property of the bidialectalism I propose to exist for New Zealand singers.*

At the core of Hazen's exploration of bidialectalism is the mutual exclusivity criterion: 'for Speaker C to be bidialectal between Dialects A and B, Speaker C would need to produce the features of both A and B in a mutually exclusive manner' (Hazen, 2001, p. 92). I will argue below, in the context of an exemplar theory approach, that it is mutual exclusivity between conversation and popular music contexts that leads to mutual exclusivity in language representations, and as a consequence, bidialectalism between speech and song.

5.4.2 The role of the media in sound change

A contentious debate in sociolinguistics is the role (or lack thereof) of the media in the diffusion of sound change. As discussed above, SPMSS may be better thought of as stable bidialectalism than as involving any kind of sound change as traditionally conceived.[1] Stuart-Smith et al. (2013) argued that TV dramas play a role in the diffusion of phonological features in the U.K., noting that 'viewers can and do become highly engaged emotionally and psychologically with the characters and their stories' (Stuart-Smith et al., 2013, p. 506). There can be no doubt that music listeners engage emotionally and psychologically, and often physically, through dance, with popular music. The traditional sociolinguistic view, however, has been that the only changes attributable to the media are lexical.

[1]The situation has many parallels to diglossia (Ferguson, 1959), even though it does not meet several of the original criteria set out by Ferguson.

> If the electronic media influenced phonology significantly, everyone in the British Isles would now have an American accent, or at least there would be progress in that direction.(Trudgill, 2014, p. 216)

This debate brings us back to the role of awareness. Trudgill states that phonological diffusion typically happens below the level of conscious awareness, but changes in lexis involve awareness, the kind of 'act of noticing' (Woolard, 2008) that inevitably happens upon encountering a novel word. What I propose here is that SPMSS is a dimension of the English language which varies so strongly from its geographically bounded forms that it acts as a distinct sub-system. And while I have perhaps not made this explicit, I also expect that SPMSS is a genuine instance of a language variety and will reflect at least some of the processes that operate on spoken varieties, including the existence of structured heterogeneity in phonetics/phonology, and a propensity to change over time as a result of the formation and morphing of social groups, even if such changes do not feed back into *speech* communities. Given that celebrity is at the core of the music industry, there may be a special role for highly prominent characterological figures (e.g. Ed Sheeran, see Section 2.17) in the transmission of new features (e.g. LOT rounding). This could tell us something about how linguistic change works, in a very different setting.

Change from below is defined by Labov (2011, p. 305) as 'the gradual development of the linguistic system in the speech community, driven by factors internal to that community. Yet relations between speech communities are present in the background throughout and sometimes emerge to take center stage'. The interaction between speech communities that occurs through popular music is one of these cases of dialect contact. Musicians form communities of practice (Watts and Andres Morrissey, 2019), despite their geographic dispersion, which could be referred to as *song communities.*

A central claim of this thesis is that the adoption of one's own accent in a song context requires great cognitive effort for a language user who has never heard that accent in a song. The large cognitive divide separating popular music from conversation allows for relatively distinct linguistic sub-systems to remain stable, in a diglossia-like relationship. However, there are several reasons to think that this divide could decrease in the future. Firstly, the putative stability of SPMSS only arose in the first place out of cultural dominance and monopolisation of the creative industries by record companies. Now that social media and mobile technology have broken down barriers to content generation and sharing, we may actually witness a turn towards centrifugal forces in the sociophonetics of song. In the language of Schneider's 2007 dynamic model, song communities might enter Phase Four, leaving the homogeneity of the mother country to enter the phase of endonormative stabilisation that some hip hop communities have already embarked upon. Alternatively, the inertia may continue. Whatever happens, it seems to me that building the study of singing into the ongoing development of both laboratory phonology and sociolinguistics would provide the fields with an excellent tool for analysing the way our ability to process speech variation relies on co-occurrence patterns in the world around us.

5.4.3 People and meanings: authenticity, awareness and indexicality

As Alim (2002, p. 300) stated, 'Hip Hop artists, by the very nature of their circumstances, are ultraconscious of their speech. As members of the HHN, they exist in a cultural space where extraordinary attention is paid to speech'. This point was made in order to challenge Labov's 1972 early conception of the relationship between attention to speech and standardness. Alim's data on copula deletion showed, by comparing hip hop lyrics to interview data, that increased attention to speech led to greater use of variants associated with African American Street Culture.

To be a hip hop artist in New Zealand involves a kind of dual identity. Rappers belong to the Hip Hop Nation (HHN) as well as to the streets of South Auckland, for example. In Chapter 2, I sometimes referred to the latter type of belonging as one of 'New Zealand' identity, as a shorthand to distinguish NZ from the USA. To think of a South Auckland based Māori rapper's use of a NZE variant as an act of 'national identity', however, would be off the mark. As a group exposed to structural racism from Pākehā majority power structures in New Zealand, 'nationality' is too broad a term for the much more specific identity of tangata whenua (Māori people, the people of the land). It is through non-belonging to the dominant culture of their region, and through 'parallels of oppression' (Zemke-White, 2008, p. 109), that Māori have a greater claim to membership in the HHN. Conversely, Pākehā rappers, implicated with the dominant majority, may feel they need to earn their place in the HHN, perhaps through affiliation with 'the streets'. A Pākehā rapper, from a middle-class suburban background, however, has recourse to fewer dimensions through which to establish belonging in the HHN, foremost amongst which is to 'keep it real' and represent their background honestly (cf. McLeod, 1999; Cutler, 2014).

'Authenticity', however, is a process and not an entity, and this process is often coerced for commercial gain. The IFPI's global music industry report for 2018 (IFPI, 2018) focuses on several breakthrough artists. These artists are lauded for 'using their own voice', 'flying the flag of their mother-tongue', or 'tapping into their roots'. The commercialisation of 'the real' makes the fusing of person and character (Coupland, 2011) highly complex. It is a task that involves the management of a wide array of subtle and complex indexical relationships.

Indexicality has been central to sociolinguistics, albeit under varying names, since at least Labov's (1972) distinction between indicators, markers and stereotypes, which from the outset invoked awareness or lack thereof as a key determining feature of different kinds of variable. While indexical processes and abstraction of social categorisations can happen in the absence of awareness (e.g. in markers), high levels of awareness add feedback dynamics to the indexical system (e.g. in stereotypes) (cf. Wedel and Fatkullin, 2017).

Phonetic variants can take on different social meanings in different contexts (Campbell-Kibler, 2011). For example, a low F2 in LOT might mean 'authentic NZ identity', 'British singer-songwriter genre', or even 'clear singing style'. On the other hand, a variant can be so generalised in its attachment to a well-enough defined context that former social meanings may begin to lose their relevance (cf. Irvine, 2001, on erasure). In particular, I would hypothesise that over the second half of the twentieth century, geographic place associations became backgrounded in commercial popular music, through accent homogeneity. This is a phenomenon defined by Squires (2014, p. 44) as *indexical bleaching*. I will draw heavily on this concept in Section 5.5.

> '[I]ndexical bleaching happens through repetition in use, and the outcome is a feature that ceases to carry the marked indexical meaning that once accrued to it'

As performers navigate the complex indexical space of popular music phonetics, they inevitably face conflicting identities (Trudgill, 1983), and must resolve these conflicts by choosing which aspects of social meaning are to be foregrounded in their presentation of self. Such decisions are dependent on motivations related to the tension between emulating esteemed performers (amongst whom there will also be conflicts) and authentication goals, and are mediated by the dynamics of self-awareness and awareness of phonetic variation. Thus, authenticity, awareness and indexicality are all central to a sociophonetics of popular music.

5.5 Structure vs. Agency in a Context-Sensitive Language System

> '[T]he performance forms of a community tend to be among *the most memorable*, repeatable, reflexively accessible forms of discourse in its communicative repertoire'. (Bauman, 2005, p. 149, emphasis added)

When I quoted the above passage in my 2011 article on parody in performances by the Flight of the Conchords, I was highlighting how the memorability of performed texts helped to reinforce characterological figures in the cultural psyche — in its 'communicative repertoire'. I draw attention to it here in order to consider how the union of words and melody in song can enhance memorability, and the impact that this may have on the issues considered in this thesis around language representation and processing.

It has been established for some time that the general music-listening public have detailed memory for both absolute pitch and rhythm in familiar songs (Halpern, 1988, 1989). As Kraus and White-Schwoch (2015, p. 645) state, auditory processing is 'always on and cannot be volitionally turned off'. The tendency for songs to get 'stuck-in-your-head' has been exploited in applied linguistics (Murphey, 1992), and studies on the speech-to-song-transformation have shown how automatically we can make music out of the human voice, given enough repetition (Falk et al., 2014). Summarising evidence from fMRI studies of people as they imagine music, Zatorre (2012) states that we co-opt perceptual mechanisms as we retrieve vivid musical memories.

The above studies come from multiple disciplines, but all emphasise the fact that song is highly memorable. By imagining a piece of music, we draw on representations in auditory memory, and can call forth specific detail about the objective properties of a recording. This occurs for the general music-listening public, not just those with 'perfect pitch'. These findings provide good evidence for the storage of detailed exemplars of sonic experiences, and thus support exemplar theories of language in which phonetic memories are highly detailed, and, like music, can be called forth by imagining (Johnson et al., 1999). It is this focus on our memories of voices in speech and in song that provides the framework for the final section of this chapter.

5.5.1 Visualising an exemplar space

> 'The perception of difference is the basis of categorization' (Wedel and Fatkullin, 2017, p. 77).

I close this chapter, and the dissertation, with a section drawing together the concepts of indexicality, awareness, authenticity, and memory. The content of this section is part discussion, and part narrative. The narrative suggests an answer[2] to the question: 'why might it be difficult for New Zealanders to sing how they speak?', and then carries on to propose a mechanism by which New Zealanders might overcome that difficulty, as a small minority have been shown to do in Chapter 2. The material below is inspired by the schematic of an exemplar store from Todd et al. (2019) that was reproduced in Figure 1.2. I will step through five different scenarios, examining different ways of looking at this exemplar store (in Figures 5.1–5.5), whilst describing how the structures of memory contribute to the predicament of the NZ speaker–listener whose dialect is rarely represented in sung form.

5.5.2 Parameters for an imaginary exemplar store

To explore the idea that the processing of a speech segment depends strongly on its context, a dataset of exemplars was created, as described in this section. Using this simulated data, I will consider how labels may emerge or fall away over time in the mind of a fictional NZ speaker–listener, who I will refer to as *the agent*. The visualisation focuses on the memory store itself, and also on labels which might emerge from it. While my discussion focuses more on perception, it is assumed that activities of both perception and production are constantly shaping the store of exemplars. This exemplar store is not a closed loop like the one represented in the Todd et al. (2019) model: the agent is exposed to a wide range of other voices in various contexts.

The data used for the visualisation come from F2 values for the LOT vowel from three different sources. I take the mean values from the males in the PoPS corpus, that were presented in Section 2.7, for each place of origin and genre. For NZ speech, I take the mean LOT value for males born after 1970 in the Canterbury Corpus, and for US speech, I take the F2 value at the centre of the ellipse for LOT presented for Western male speakers in Clopper et al. (2005). For the purposes of illustration, I will simulate normal distributions around each of these means (using a constant standard deviation of 100Hz). This will create a simulated exemplar store in the mind of the agent, representing a collection of her encounters with the F2 dimension of the vowel in LOT words.

As a New Zealander, the large majority of the agent's experiences with this vowel are in the context of talking to other New Zealanders. I represent this skewed distribution in the simulated exemplar space by including extra tokens in the NZ conversation cell. The number of exemplars in each cell of the imagined exemplar space, and the mean F2 value for each cell, is shown in Table 5.1. The combined

[2] This 'answer' is really just my own non-technical description of exemplar theory as adopted in sociophonetics (Hintzman, 1986; Goldinger, 1998; Johnson et al., 1999; Pierrehumbert, 2001; Hay et al., 2006a; Foulkes and Docherty, 2006; Johnson, 2006; Pierrehumbert, 2006; Drager, 2010; Hay and Foulkes, 2016; Racz et al., 2017; Todd et al., 2019), along with sociolinguistic treatments of indexicality (Silverstein, 2003; Eckert, 2008).

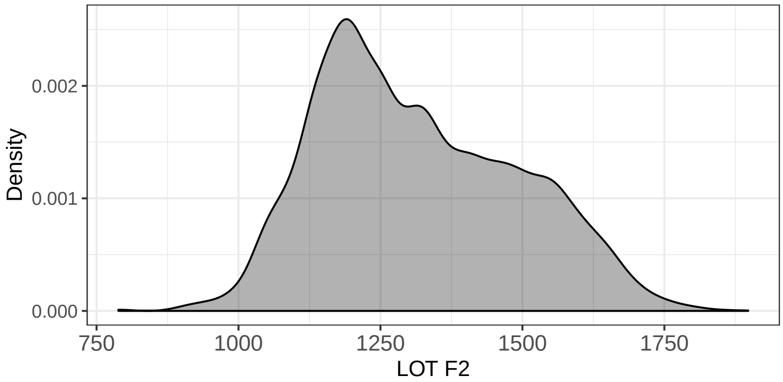


Figure 5.1: Distribution of experiences with the F2 of LOT for a NZ speaker-listener, with knowledge of all socio-contextual information hidden.

Table 5.1: Mean LOT F2 values for each cell in the simulated exemplar space of a NZ speaker-listener, with the number of exemplars in brackets. Tokens come from NZ and US speakers/singers, in the PoPS corpus (for rap and song), and from the Canterbury Corpus and Clopper et al. (2005) for conversation.

	Conversation	Rap	Song
NZ	1202Hz (8000)	1435Hz (1000)	1556 (1000)
US	1450Hz (1000)	1469Hz (1000)	1566Hz (1000)

density distribution of these 13,000 exemplars of LOT is shown in Figure 5.1. This distribution has one clear peak (representing the plentiful encounters in conversation just mentioned), with a slightly right-skewed distribution. This distribution is similar in concept to the distribution for each of DRESS and TRAP in the Todd et al. (2019) model, in that any systematicity according to speaker characteristics or context is hidden (cf. Pierrehumbert, 2002, p. 115).

The plots below all focus on just this single acoustic dimension, F2, and just this one phonological category: LOT. The density distributions in the above plot, and the next two plots below could be compared to the 'Exemplar Store' section of the Todd et al. (2019) model reproduced in Figure 1.2. The data represented here would correspond to just one of the phoneme clusters, with the F2 of LOT corresponding to the gradient shading that fills each exemplar in that schematic. In the present simulated exemplar space, there are also two dimensions of socio-contextual information (Country and Context) represented within each exemplar, which is where this example departs from the Todd et al. (2019) model. In this demonstration, the agent is able to consider input from other speakers, and to consider co-occurrences with other acoustic and non-acoustic dimensions of a speech signal.

5.5.3 Revealing systematicity with socio-contextual information

I begin adding complexity to the agent's exemplar store with an unrealistic scenario: at the time of encoding, the agent knew (consciously or not) the place of origin of the speaker, for every exemplar of LOT stored. This scenario is illustrated in Figure

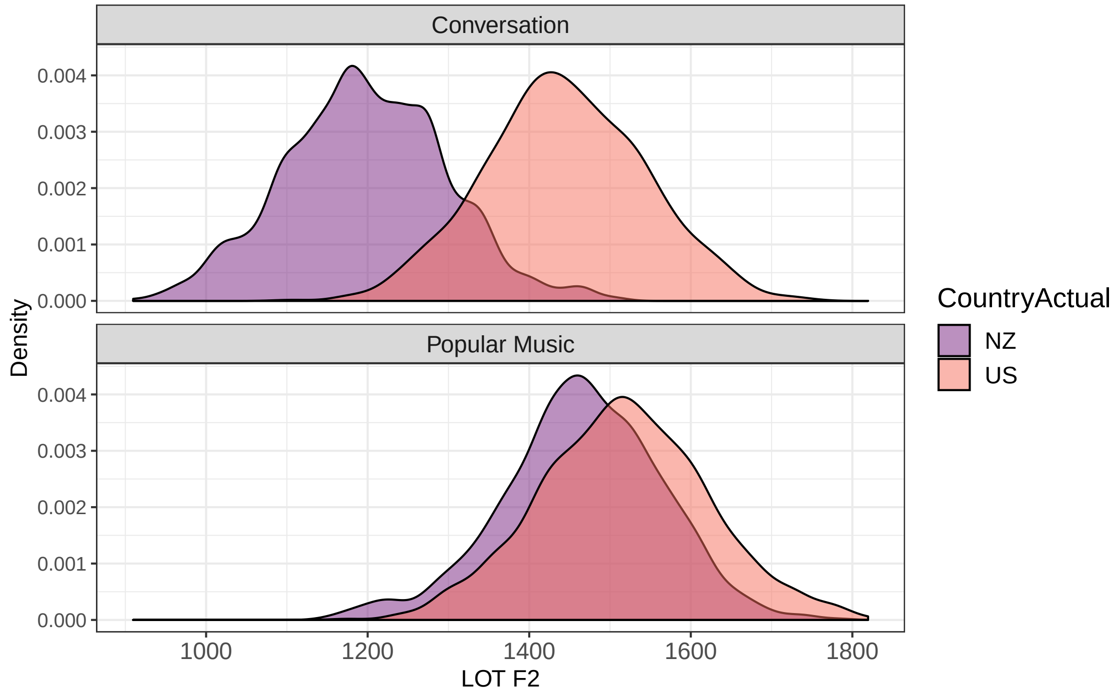


Figure 5.2: Simulated F2 distributions for LOT in the exemplar store of the agent, based on normal distributions around the means from PoPS, Canterbury Corpus and Clopper et al. (2005). In an unrealistic scenario, all exemplars are tagged with the speaker/singer's place of origin.

5.2. In this scenario, the agent also kept track of whether each token of LOT was sung or spoken. Unlike place of origin, that distinction is encoded in the acoustics of the vowel token itself, on a range of dimensions not shown here. Foremost amongst these would be the presence or absence of accompanying instrumentation (cf. Pufahl and Samuel, 2014). Therefore, unlike place of origin, it is quite reasonable to assume that each exemplar is labelled in a meaningful way as speech or song. The multi-dimensionality of the exemplar space will be emphasised throughout this section.

Through the addition of the Country and Context dimensions, systematicity that was hidden in Figure 5.1 has been revealed. The agent can see that New Zealanders in conversation have lower F2 in LOT words than do US speakers in conversation, or singers from either country.

In a more realistic scenario, the agent would not always know which singers come from which country, though the agent may have semantic knowledge about some singers. There are some famous singers that the agent knows are American, for example. There are perhaps a few that she has read articles about online, and some others that she has seen playing live at a NZ music festival. But many of the LOT vowels the agent hears in songs come from singers she knows nothing about. To assign those exemplars with a Country label, she must be able to make a guess about where they come from based on properties within the signal itself. This is difficult, because the singers that she *does* have reliable labels for, have overlapping distributions of F2 (and overlapping distributions on a range of other dimensions).

The task of assigning a country label to an exemplar is easy in conversation — NZ and US speakers form well-defined peaks. It is also functionally relevant to the agent, on a range of other dimensions, whether she is talking to a New Zealander or

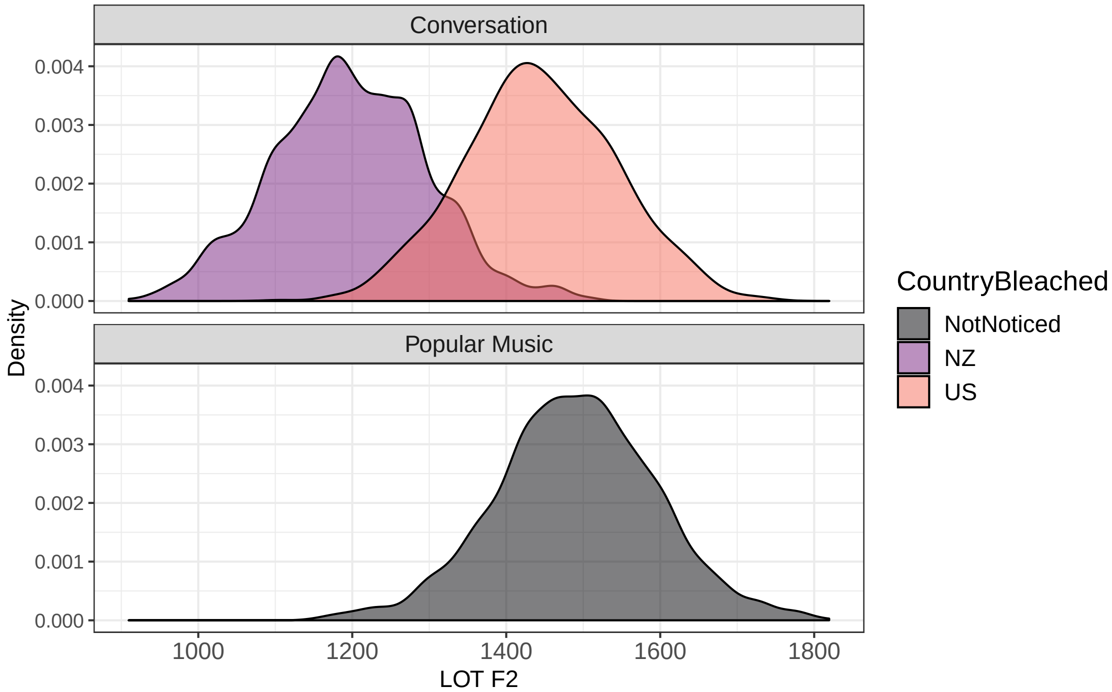


Figure 5.3: Simulated F2 distributions for LOT in the exemplar store of a NZ speaker-listener, based on normal distributions around the means from PoPS, Canterbury Corpus and Clopper et al. (2005). In a scenario of complete indexical bleaching in the context of popular music, the agent stops paying attention to the relationship between a singer's place of origin and the F2 of their LOT vowels. The agent still sees obvious utility in tracking this information in the context of conversations, however, and continues to do so. In spoken contexts, a LOT vowel with a high F2 'sounds American', but in popular music it does not.

to an American. In conversation, she can quickly determine that a new interlocutor comes from New Zealand through the co-occurrence of a raised DRESS and a fronting GOAT vowel, among numerous other indices, in the vicinity of the LOT token being encountered. This strengthens the indexical connection between the NZE variant of each of those vowels and the NZ label. Cues in nearby variables are no more helpful in *song*, however, than the token of LOT itself. The distributions between NZ and US singers overlap in each lexical set. DRESS and TRAP are open in songs no matter where the singer is from, and the F2 of GOAT always decreases from nucleus to offglide. Even though the agent could not describe these acoustic patterns, her mind stores and learns from the patterns she encounters, without her awareness.

Eventually, she (or her statistical learning mechanism) learns to stop assigning the Country label to sung tokens of LOT as they enter her exemplar store. Whether a person is from NZ or the US is relevant in face to face conversations, but it is not worth keeping track of in songs since it doesn't predict F2. Through a process of *indexical bleaching* (Squires, 2014), the LOT vowels no longer 'sound American' when sung by a US singer, nor do they 'sound NZ' when sung by someone she knows to come from NZ. They all just sound like singing. This indexically bleached exemplar space is shown in Figure 5.3, and can be summarised as follows:

Indexical bleaching — the agent stops paying attention to the relationship between a singer's place of origin and the quality of their vowels.

5.5.4 Socio-contextual scales, not categories

The reason I simulated a large number of points in this hypothetical exemplar space was so that we could visualise not just distributions, but *clouds* of exemplars. In such a visualisation, I avoid the tendency to treat acoustic variables as gradient, and socio-contextual variables as categorical. Already we have seen how a binary categorisation of NZ and US speakers is insufficient. There is not only a scale from 'very NZish' to 'very American', but there is also a space where the dimension itself ceases to be relevant, as represented by the grey 'NotNoticed' area in Figure 5.3. Treating socio-contextual dimensions as gradient leads to a consideration of how subtle, or how vast, the cognitive differences between two levels of a variable might be.

Across the massive array of dimensions of differentiation abstracted from experience, there is variability in the magnitude of the difference represented. For example, the distinction between *angry* and *frustrated* may form a dimension of differentiation in the mind, on a slight, yet gradient, scale. In the same way, while we might have a speaker-characteristic label called 'Age', with the poles *old* and *young*, we could also have numerous other dimensions of difference to represent age. For example, there might be enough characterological figures in our cultural imagination to carve out a meaningful opposition between a *twenty-something* and a *thirty-something*. Even this subtle social distinction should be conceived of as granular, with prototypical and in-between exemplars. Imagine the fMRI patterns that might be elicited by the faces of twenty-somethings and thirty-somethings. While we could measure this social knowledge through behavioural tasks, it might be difficult to measure in the scanner. Neurophysiological differences in activation between experiences with speech and song, on the other hand, can be readily measured. In sum, rather than thinking of socio-contextual information as a set of discrete labels, I conceive of

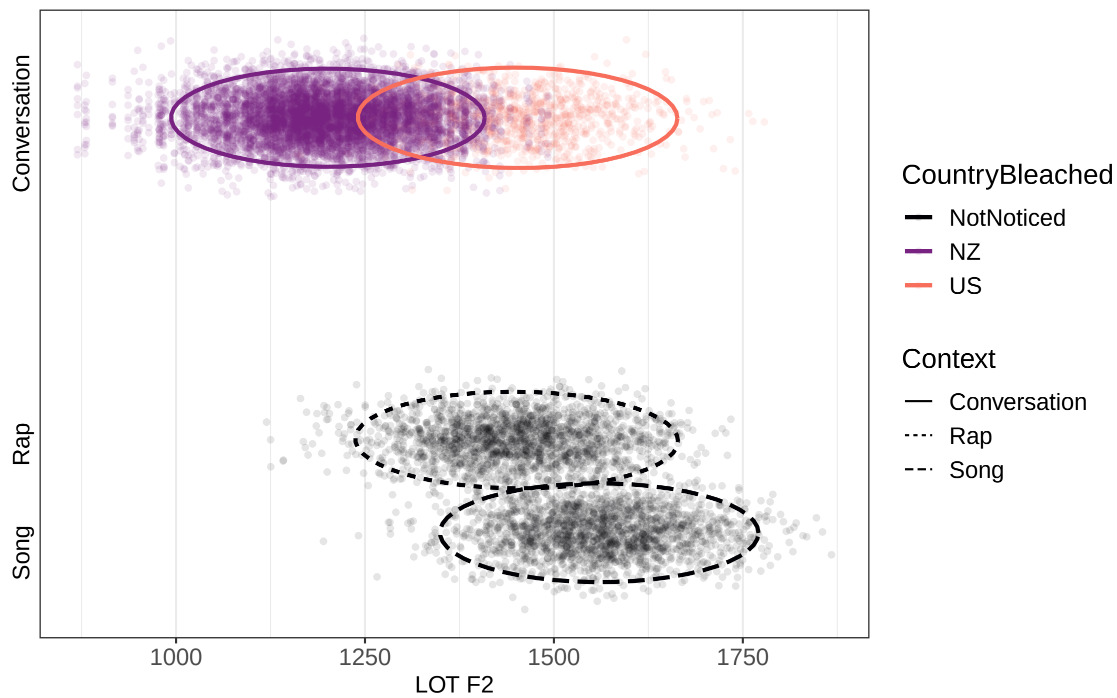


Figure 5.4: Simulated F2 distributions for LOT in the exemplar space of a NZ speaker-listener, based on normal distributions around the means from PoPS, Canterbury Corpus and Clopper et al. (2005). Indexical bleaching scenario, with the Conversation–Singing dimension represented as gradient.

such information as capturing gradient relationships. These are stored in the mind, through similar mechanisms as those keeping track of gradient acoustic properties, through multi-dimensional arrays:

Dimensions of differentiation — The agent keeps track of where exemplars sit on a number of gradient scales. These scales span between two qualities or entities deemed by the statistical learning mechanism to be related (perhaps through a process not dissimilar to the one used to abstract continuous word representations in Bojanowski et al., 2016).

5.5.5 From distributions to clouds

To represent the above point of view in the agent's exemplar store, Figures 5.4 and 5.5 use a different visual format. The exemplar *store* becomes an exemplar *space*, and a wide gap between conversation and song is represented. This allows us to think about what a cline from singing to speech actually consists of. It is instructive to imagine where a poem, a sermon, a Gregorian chant, or a lullaby might sit on such a scale. Many dimensions of differentiation would represent various aspects of the singing–speech opposition: the presence of background music; the complexity of harmonic structures; the stability and height of a voice's pitch; the isochrony of syllables; the prevalence of rhyme. For now, I make a simple three-way distinction. Popular music is hugely different to conversation, and within popular music, pop is a bit more 'singy' than hip hop.

In Figure 5.4, we return to the agent's experiences with the F2 of LOT, with the same scenario of complete indexical bleaching, but now with the conversation–

singing dimension represented as gradient. For the first time, we can see the systematicity by genre that was previously hidden. LOT has a slightly lower F2 in rap than in song. Pausing our examination of the agent's experiences with LOT for a moment, the next section considers the exemplar space represented in Figure 5.4 in the context of the framing question of this thesis.

5.5.6 Why is it hard to sing how you speak?

The preceding sections have highlighted a number of possible reasons why it might be difficult for New Zealanders to sing how they speak. Figure 5.4, I argue, represents four key reasons why the adoption of NZE sound structures in song might take conscious effort.

- Empty indexical space: the gap between conversation and music creates a cognitive gulf of some kind that is difficult to cross.[3].

- The homogeneity of singing accents: the large difference between speech and song, and the homogeneity of acoustic realisations of tokens within the song cluster enter into a feedback dynamic that supports and entrenches the division.

- Indexical bleaching: speaker biographical information is difficult to determine from the acoustics, and the acoustics do not robustly predict speaker characteristics. This constitutes another feedback dynamic, causing the indexical fields in song and speech to pull away from one another. Each context develops its own distinct internal systematicity.

- Bidialectalism: Once again, this is both a result and a cause, and is related to indexical bleaching. Individuals can produce mutually exclusive dialects according to context, which further detaches the speech and song indexical fields. While innovations away from SPMSS are constantly occurring, they occur in a space where the pressures of levelling are heightened by this mass bidialectalism. Variants that are marked will quickly be levelled through the central principle of deterministic contact situations: '*the survival of majority forms*' (Trudgill, 2004, p. 114, emphasis in original).

While the global system as a whole seems to be stuck at Phase Three of Schneider's 2007 dynamic model (see Section 1.5.1), its very dependence on the feedback dynamics described above may make its stasis fragile. The bubbling out of centrifugal forces (Bakhtin, 1981) is constantly present, and it occurs through the agency of individuals. The question that has framed this project focused on the ways in which New Zealanders are *consrained*, on the perceived 'difficulty' of using NZE in song. In seeking an answer to that question, the project has grown into an exploration of how memory structures not only constrain us to repeat ourselves, but also provide us with the tools for re-invention. The creation of something new by recombining the structures of what came before (e.g. through bricolage, Hebdige, 1979) is at the

[3]I have hard-wired this into the graph simply to demonstrate the point. Some evidence for this position was presented in Section 1.2, though that literature related more to the distinctions between music and language cognition than to differences specifically between singing and speech.

heart of performance (Bauman and Briggs, 1990). The structuring force of normativity is important, but it is not the end of the story. In the next section I return to the illustration of the agent's exemplar space, and consider the processes through which own-accent singing and rap might emerge.

5.5.7 Pop-out effects, feedback dynamics and indexical fusion

Up until this point, the agent has been exposed almost exclusively to SPMSS and HHNL in her music-listening experiences. In this section, I consider what might happen when she encounters a track by one of the own-accent rappers described in Chapter 2 (the individuals shown in the last few rows of Table 2.17), David Dallas, for example. She is struck by a series of NZE variants, including a new exemplar for LOT that has an F2 much lower than she is used to hearing in popular music. As an outlier to the distribution, it *pops out*, attracting her attention. It is surrounded by tokens of other variables which also pop out, and these tokens all activate the cloud of exemplars inhabiting the portion of exemplar space indexed as 'NZ Conversation'.

While outliers that do not make any sense to the statistical learning mechanism may be rejected as errors, this low F2 LOT vowel is passed to the agent's consciousness because it is incongruent *in a meaningful way*. It reaches out across empty indexical space to exemplars that match it in unexpected ways. The token grabs the agent's attention (perhaps eliciting a P3 response, Polich, 2007) and causes an *act of noticing* (Woolard, 2008). It is then encoded strongly, and persists in memory (Sumner et al., 2014). An indexical link begins to form between these strongly encoded tokens and a cloud of exemplars that connect (multi-modal) episodic memories of David Dallas to a range of dimensions of differentiation across the multi-dimensional exemplar space, including the Country dimension, previously dismissed as irrelevant to popular music.

Because of this act of noticing, the indexical bleaching that had happened below the level of awareness is slightly reversed. There is a meaningful link between Country and F2 in rap. On a behavioural level, the agent acts on a positive evaluation of the song and artist, and starts listening to it regularly, along with the rest of David Dallas' music, and music by related artists. Her exposure to NZE in popular music enters into a process of 'feedback over cycles of perception, categorization, and reproduction' resulting in an evolving 'self-referential system' (Wedel and Fatkullin, 2017, p. 77). By attracting conscious awareness, the connections between F2, Country and Context move to higher orders of indexicality, and become amenable to conscious processes of assessment and categorisation. Through such processes, the agent may *fuse* other dimensions of differentiation onto this indexical package. For example, the agent may implicate the Authenticity dimension in her assessment of these low F2 LOT tokens, as labelled on Figure 5.5. This process can be defined as follows:

Indexical fusion — Through agency, the speaker-listener focuses their attention on connections between dimensions of difference not previously noticed. Feedback dynamics emerge, and a dimension of differentiation forms connections with an existing indexical field, either raising its dimensionality, or bleaching other dimensions of difference from the field.

Within the exemplar space shown in Figure 5.5, the strongly encoded 10% of LOT

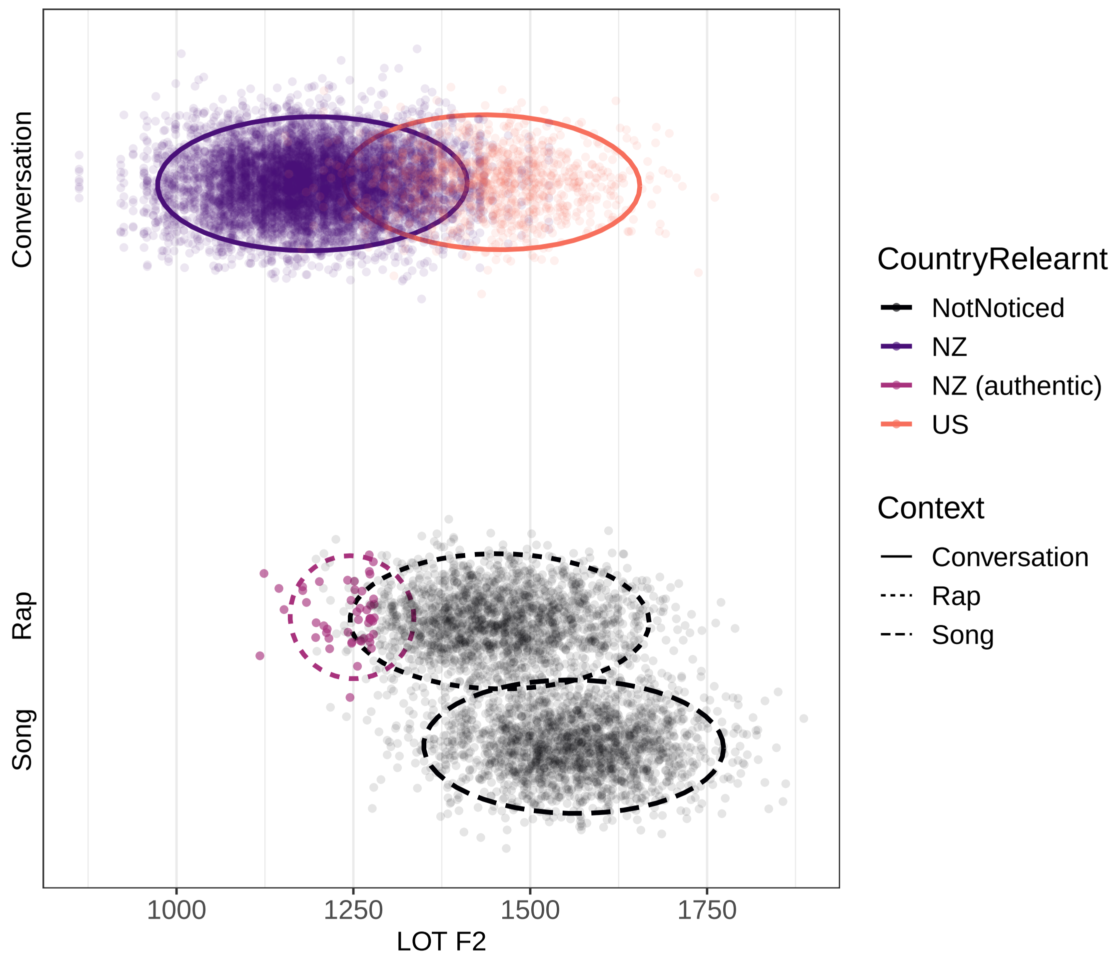


Figure 5.5: Simulated F2 distributions for LOT in the exemplar space of a NZ speaker-listener, based on normal distributions around the means from PoPS, Canterbury Corpus and Clopper et al. (2005). Outlier exemplars with low F2 *pop out*, attracting attention and ultimately evaluation. Their 'New Zealandness' becomes salient, and as a result, the tokens form indexical connections with the ideology of authenticity. The Authenticity dimension of differentiation begins to make connections with, and draw closer to, the indexical field displayed.

vowels with the lowest F2 have formed indexical links with a new 'NZ (authentic)' category label. Through further stages of assessment, incoming tokens may start to be tagged on the opposite end of the Authenticity dimension, leading to the ideology of 'a fake American accent'.

Indexical bleaching gives way to indexical fusion and the agent now has the tools required for own-accent rap, at least with respect to the F2 of LOT. She has developed a *motivation* to be 'authentic', and *awareness* of how this desired social meaning maps onto acoustic space. When selecting a production target, she does not select randomly from the distribution of LOT vowels in rap, but focuses her attention on the sub-portion of that cloud which maps onto the desired pole of the authenticity dimension. In this way, she exerts *control* over her production of this variable. Through targeted selection, she performs an initiative act of identity, and stylises NZE in the rap context.

Other vowels may have remained at the stage of indexical bleaching, though the authenticity dimension will now continue to make connections wherever NZ-like variants *pop out* in music contexts. This, like the processes described above to explain why it is *difficult* to use NZE in song, introduces feedback into the indexical system. For the agent in this scenario, the feedback leads to rapidly increasing awareness of how NZE differs from SPMSS and HHNL. In this agent, then, we see the emergence of a centrifugal force (Bakhtin, 1981). How this force interacts with and affects the exemplar spaces of other agents depends on the unique exemplar store that each one of them possesses.

5.5.8 Mergers and splits of non-linguistic categories

The purpose of the above example was to visualise indexical processes operating in a gradient multi-dimensional exemplar space. The ideology that emerged for the fictional agent in this narrative is, in fact, the very same one which was discussed in Chapters 1 and 2: the ideology that makes singers want to fuse person and character (Coupland, 2011) — to 'be real'. The feedback dynamics which led to the ideology were initially sparked by an act of noticing. This act of noticing occurred because a contextually incongruent token *popped out* of the signal and attracted attention.

The final state of the simulated exemplar space above showed signs of re-attaching place meanings to the acoustics of the sung LOT vowel. This seems to be a plausible account of the steps a NZ artist might go through on their way to own-accent singing or rap. If it were happening to *a lot of artists*, *a lot of the time*, then we would see a mass shift from SPMSS/HHNL to NZE. Of course, the degree to which categories merge, form or split is affected by the degree to which the category boundaries are robustly supported. One reason that this mass shift has not happened, then, could be that the cluster of exemplars shown as 'NZ (authentic)' in Figure 5.5 may not be well enough differentiated from other nearby tokens to allow for the emergence of the category label shown. It may be that the cluster of productions in popular music are, overall, so robustly different from productions in speech that splits within it are difficult to support.[4]

[4] A computational implementation of the processes discussed here could perhaps draw on the modelling of mergers and splits described by Harrington et al. (2018, pp. 716–718).

5.6 Concluding Remarks

By studying New Zealanders' production and perception of phonetic variation in song, this thesis has explored both sociolinguistic and psycholinguistic issues. The results of the corpus analysis demonstrated the prevalence of (AmE-derived) Standard Popular Music Singing Style, whilst hip hop demonstrates the influence of (AAE-derived) Hip Hop Nation Language, along with a greater tendency to own-accent phonetic styles. The phonetic categorisation task showed that listeners shift their phonological reference frame according to context-induced expectations. When listening to a song, they expect to hear SPMSS, and adjust the perceptual boundaries between phonetic categories accordingly.

The lexical decision task revealed a striking finding. Average music-listening New Zealanders are as quick to process a US voice in song as they are to process a NZ voice in non-musical contexts. They behave like native-speakers of American English in song, and like native-speakers of New Zealand English elsewhere. This result is strongest for people who listen mainly to US music, and who believed the voice in the experiment, when set to music, actually sounded like it was singing. This highlights the importance of individual experience in the context-dependent bidialectalism which the corpus and perception results, as a whole, suggest some New Zealanders possess.

The results provide further evidence that SPMSS is the default style in the context of song, in both production and perception. Cases of own-accent singing and rap are exceptions to the norm, and represent initiative acts of identity. In this final chapter, I have presented an illustration of how such acts of identity might take place. Language users form connections between ideological constructs (including motivations) and acoustic/socio-contextual dimensions, through a process of indexical fusion. Through *awareness* of these relationships, the social agent can focus on a relevant sub-portion of their acoustic memories and *control* their speech (or singing, or rap) production.

A sociophonetics of popular music has a great deal to offer both laboratory phonology and sociolinguistics. It suggests that we should give as much consideration to the contextual aspects of our experiences as we do to linguistic and social indexical knowledge. As a case study of context-dependent style, the study of singing can provide insights into both cognitive and social aspects of language variation.

References

Adank, P., Smits, R., and van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, 116(5):3099–3107.

Agha, A. (2003). The social life of cultural value. *Language and Communication*, 23:231–273.

Agha, A. (2005). Voice, footing, enregisterment. *Journal of Linguistic Anthropology*, 15(1):38–59.

Alim, H. S. (2002). Street-Conscious Copula Variation in the Hip Hop Nation. *American Speech*, 77(3):288–304.

Alim, H. S. (2009). Intro. Straight outta Compton, straight aus Munchen: Global linguistic flows, identities, and the politics of language in a Global Hip Hop Nation. In Alim, H., Ibrahim, A., and Pennycook, A., editors, *Global linguistic flows: hip hop cultures, youth identities, and the politics of language*. Routledge, New York and London.

Alim, H. S., Ibrahim, A., and Pennycook, A. (2009). *Global linguistic flows: hip hop cultures, youth identities, and the politics of language*. Routledge, New York and London.

Andres Morrissey, F. (2008). Liverpool to Louisiana in one lyrical line: Style choice in British rock, pop and folk singing. In Locher, M. A., editor, *Standards and norms in the English language*, book section 10, pages 195–218. Mouton de Gruyter, Berlin.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481.

Babel, A. M. (2016). *Awareness and control in sociolinguistic research.*

Babel, M. (2012). Evidence for Phonetic and Social Selectivity in Spontaneous Phonetic Imitation. *Journal of Phonetics*, 40(1):177–189.

Bakhtin, M. (1981). *The Dialogic Imagination.*

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1).

Bauman, R. (2005). Commentary: Indirect indexicality, identity, performance. *Journal of Linguistic Anthropology*, 15(1):145–150.

Bauman, R. and Briggs, C. L. (1990). Poetics and Performance as Critical Perspectives on Language and Social Life. *Annual Review of Anthropology*, 19:59–88.

Beal, J. C. (2009). "You're Not from New York City, You're from Rotherham": Dialect and Identity in British Indie Music. *Journal of English Linguistics*, 37(3):223–240.

Bell, A. (1984). Style as Audience Design. *Language in Society*, 13(2):145–204.

Bell, A. (2001). Back in style: reworking audience design. In Eckert, P. and Rickford, J. R., editors, *Style and Sociolinguistic Variation*, book section 9, pages 139–169. Cambridge University Press, Cambridge.

Bell, A. (2011). Leaving Home: De-europeanisation in a post-colonial variety of broadcast news language. *Standard Languages and Language Standards in a Changing Europe. Oslo, Norway: Novus*, pages 177–198.

Bell, A. and Gibson, A. (2008). Stopping and Fronting in New Zealand Pasifika English. *University of Pennsylvania Working Papers in Linguistics: A Selection of Papers from NWAV 36*, 14(2):42–53.

Bell, A. and Gibson, A. (2011a). Staging language: An introduction to the sociolinguistics of performance. *Journal of Sociolinguistics*, 15(5):555–572.

Bell, A. and Gibson, A. (2011b). *The Sociolinguistics of Performance: Theme Issue of the Journal of Sociolinguistics (15:5)*. Wiley, London: UK.

Bell, B. A., Ferron, J. M., and Kromrey, J. D. (2008). Cluster size in multilevel models: the impact of sparse data structures on point and interval estimates in two-level models. *JSM Proceedings, Section on Survey Research Methods*, pages 1122–1129.

Biewer, C. (2015). *South Pacific Englishes: a sociolinguistic and morphosyntactic profile of Fiji English, Samoan English and Cook Islands English*, volume 52.;G52;. John Benjamins Publishing Company, Amsterdam.

Bigham, D. S. (2010). Correlation of the Low-Back Vowel Merger and TRAP-Retraction. *University of Pennsylvania Working Papers in Linguistics*, 15(2).

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Boersma, P. and Weenink, D. (2019). *Praat: doing phonetics by computer [Computer program]. Version 6.1.04*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information.

Borsky, S., Tuller, B., and Shapiro, L. P. (1998). "How to milk a coat:" the effects of semantic and acoustic information on phoneme categorization. *J Acoust Soc Am*, 103(5 Pt 1):2670–6.

Bourdieu, P. (1991). *Language and Symbolic Power*. Polity Press, Cambridge.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Brooks, W. (1982). Theory and Method: On being tasteless. *Popular Music*, 2:9–18.

Bucholtz, M. and Hall, K. (2005). *Identity and Interaction: A Sociocultural Linguistic Approach*, volume 7.

Campbell-Kibler, K. (2011). The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, 22(3):423–441.

Campbell-Kibler, K. (2012). The Implicit Association Test and sociolinguistic meaning. *Lingua*, 122(7):753–763.

Carmichael, K. (2017). Displacement and local linguistic practices: R-lessness in post-Katrina Greater New Orleans. *Journal of Sociolinguistics*, 21(5):696–719.

Chambers, J. K. (1992). Dialect acquisition. *Language*, 68(4):673–705.

Clark, L. (2018). Priming as a Motivating Factor in Sociophonetic Variation and Change. *Topics in Cognitive Science*, 10(4):729–744.

Clopper, C. G., Pierrehumbert, J. B., and Tamati, T. N. (2010). Lexical neighborhoods and phonological confusability in cross-dialect word recognition in noise. *Laboratory Phonology*, 1(1).

Clopper, C. G. and Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32:111–140.

Clopper, C. G., Pisoni, D. B., and de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America*, 118(3):1661–1676.

Clopper, C. G., Tamati, T. N., and Pierrehumbert, J. B. (2016). Variation in the strength of lexical encoding across dialects. *Journal of Phonetics*, 58:87–103.

Clopper, C. G. and Walker, A. (2017). Effects of Lexical Competition and Dialect Exposure on Phonological Priming. *Language and Speech*, 60(1):85–109.

Clyne, M. G. (1997). Pluricentric Languages and National Identity: An Antipodean View. In Schneider, E., editor, *Englishes around the World: Studies in honour of Manfred Görlach. Volume 2: Caribbean, Africa, Asia, Australasia*, pages 287–300. John Benjamins Publishing Company.

Coddington, A. (2004). *Singing as we Speak? An Exploratory Investigation of Singing Pronunciation in New Zealand Popular Music*. master's thesis, University of Auckland.

Collister, L. B. and Huron, D. (2008). Comparison of Word Intelligibility in Spoken and Sung Phrases. *Empirical Musicology Review*, 3(3).

Connine, C. M., Titone, D., and Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(1):81–94.

Conrey, B., Potts, G. F., and Niedzielski, N. A. (2005). Effects of dialect on merger perception: ERP and behavioral correlates. *Brain Lang*, 95(3):435–49.

Coupland, N. (1985). 'Hark, hark, the Lark': Social Motivations for Phonological Style-Shifting. *Language & Communication*, 5(3):151–171.

Coupland, N. (2003). Sociolinguistic authenticities. *Journal of Sociolinguistics*, 7(3):417–431.

Coupland, N. (2011). Voice, place and genre in popular song performance. *Journal of Sociolinguistics*, 15(5):573–602.

Crowley, K. E. and Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: age, sleep and modality. *Clinical neurophysiology*, 115(4):732–744.

Cutler, C. (2014). *White Hip Hoppers, Language and Identity in Post-Modern America*. Taylor & Francis.

Daltrozzo, J. and Schön, D. (2009). Is conceptual processing in music automatic? An electrophysiological approach. *Brain Research*, 1270:88–94.

Danker, J. F. and Anderson, J. R. (2010). The ghosts of brain states past: remembering reactivates the brain regions engaged during encoding. *Psychol Bull*, 136(1):87–102.

De Vos, M., Thorne, J. D., Yovel, G., and Debener, S. (2012). Let's face it, from trial to trial: Comparing procedures for N170 single-trial estimation. *NeuroImage*, 63(3):1196–1202.

D'Onofrio, A. (2018). Personae and phonetic detail in sociolinguistic signs. *Language in Society*, 47(04):513–539.

Drager, K. (2006). From Bad to Bed: The Relationship Between Perceived Age and Vowel Perception in New Zealand English. *Te Reo*, 48.

Drager, K. (2010). Sociophonetic Variation in Speech Perception. *Language and Linguistics Compass*, 4(7):473–480.

Drager, K. (2011). Speaker Age and Vowel Perception. *Language and Speech*, 54(1):99–121.

Duncan, D. (2017). Australian singer, American features: Performing authenticity in country music. *Language & Communication*, 52:31–44.

Eberhardt, M. and Freeman, K. (2015). 'First things first, I'm the realest': Linguistic appropriation, white privilege, and the hip-hop persona of Iggy Azalea. *Journal of Sociolinguistics*, 19(3):303–327.

Eckert, P. (2000). *Linguistic Variation as Social Practice*. Blackwell, Oxford.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, 41(1):87–100.

Fahy, K. M., Lee, A., and Milne, B. J. (2013). New Zealand socio-economic index 2013. Report, Statistics New Zealand.

Falk, S., Rathcke, T., and Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*, 40(4):1491–1506.

Fancourt, D. and Perkins, R. (2018). The effects of mother–infant singing on emotional closeness, affect, anxiety, and stress hormones. *Music & Science*, 1:205920431774574.

Ferguson, C. A. (1959). Diglossia. *Word*, 15(2):325–340.

Finegan, E. and Biber, D. (1994). Register and social dialect variation: An integrated approach. In Biber, D. and Finegan, E., editors, *Sociolinguistic Perspectives on Register*, pages 315–347. Oxford University Press, Oxford.

Flanagan, P. J. (2019). 'A Certain Romance': Style shifting in the language of Alex Turner in Arctic Monkeys songs 2006–2018. *Language and Literature*, 28(1):82–98.

Floccia, C., Goslin, J., Girard, F., and Konopczynski, G. (2006). Does a Regional Accent Perturb Speech Processing? *Journal of Experimental Psychology: Human Perception and Performance*, 32(5):1276–1293.

Foucart, A., Garcia, X., Ayguasanosa, M., Thierry, G., Martin, C., and Costa, A. (2015). Does the speaker matter? Online processing of semantic and pragmatic information in L2 speech comprehension. *Neuropsychologia*, 75:291–303.

Foulkes, P. (1997). Rule inversion in a British English dialect - a sociolinguistic investigation of [r]-sandhi in Newcastle upon Tyne. *University of Pennsylvania Working Papers in Linguistics: A Selection of Papers from NWAV 25*, 4(1):259–270.

Foulkes, P. and Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, 34:409–438.

Foulkes, P. and Hay, J. (2015). The Emergence of Sociophonetic Structure. In MacWhinney, B. and O'Grady, W., editors, *The Handbook of Language Emergence*.

Fromont, R. and Hay, J. (2012). LaBB-CAT: an Annotation Store. In *Proceedings of Australasian Language Technology Association Workshop*, pages 113–117.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). The Identification and Discrimination of Synthetic Vowels. *Language and Speech*, 5:171–189.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1):110–125.

Gerfer, A. (2018). Global reggae and the appropriation of Jamaican Creole. *World Englishes*.

Gibson, A. (2005). Non-Prevocalic /r/ in New Zealand Hip-Hop. *New Zealand English Journal*, 19:5–12.

Gibson, A. (2010a). New Zealand identity in popular music: Vowel differences between singing and speaking. In Johnson, H., editor, *Many Voices: Music and National Identity in Aotearoa/New Zealand*, book section 11, pages 111–121. Cambridge Scholars Publishing, Newcastle upon Tyne.

Gibson, A. (2010b). *Production and Perception of Vowels in New Zealand Popular Music*. Master's dissertation, Auckland University of Technology.

Gibson, A. (2011). Flight of the Conchords: Recontextualizing the voices of popular culture1. *Journal of Sociolinguistics*, 15(5):603–626.

Gibson, A. (2016). Samoan English in New Zealand: Examples of consonant features from the UC QuakeBox. *New Zealand English Journal*, 29&30:25–50.

Gibson, A. and Bell, A. (2010). Performing Pasifika English in New Zealand: The case of bro'Town. *English World-Wide*, 31(3):231–251.

Gibson, A. and Bell, A. (2012). Popular Music Singing as Referee Design. In Hernández-Campoy, J. M., editor, *Style-Shifting in Public: New Perspectives on Stylistic Variation*, pages 139–164. John Benjamins.

Giddens, A. (1979). *Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis*. University of California Press, Berkeley and Los Angeles.

Godden, D. R. and Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, 66(3):325–331.

Goffman, E. (1981). *Forms of Talk*. Blackwell, Oxford.

Goldinger, S. D. (1996a). Auditory Lexical Decision. *Language and Cognitive Processes*, 11(6):559–568.

Goldinger, S. D. (1996b). Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5):1166–1183.

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, 105(2):251–279.

Goldinger, S. D. and Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*, 31(3):305–320.

Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A., and Trudgill, P. (2004). *New Zealand English: Its Origins and Evolution*. Studies in English Language. Cambridge University Press, Cambridge.

Gordon, R. L. (2010). *Neural and behavioral correlates of song prosody*. PhD thesis, ProQuest Dissertations Publishing.

Gordon, R. L., Magne, C. L., and Large, E. W. (2011). EEG Correlates of Song Prosody: A New Look at the Relationship between Linguistic and Musical Rhythm. *Frontiers in psychology*, 2:352.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, 87(1):1–51.

Grossberg, S. and Kazerounian, S. (2016). Phoneme restoration and empirical coverage of Interactive Activation and Adaptive Resonance models of human speech processing. *The Journal of the Acoustical Society of America*, 140(2):1130–1153.

Hagen, M., Kerkhoff, J., and Gussenhoven, C. (2011). *Singing your accent away, and why it works.*

Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold, London.

Halpern, A. R. (1988). Perceived and Imagined Tempos of Familiar Songs. *Music Perception: An Interdisciplinary Journal*, 6(2):193–202.

Halpern, A. R. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition*, 17(5):572–581.

Harrington, J., Kleber, F., Reubold, U., Schiel, F., and Stevens, M. (2018). Linking Cognitive and Social Aspects of Sound Change Using Agent-Based Modeling. *Topics in Cognitive Science*, 10(4):707–728.

Harrington, J., Pouplier, M., and Reinisch, E. (2019). Introducing abstraction, diversity, and speech dynamics. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 10(1).

Harrison, A. K. (2008). Racial Authenticity in Rap Music and Hip Hop. *Sociology Compass*, 2(6):1783–1800.

Hasson, U., Egidi, G., Marelli, M., and Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, 180:135–157.

Hawkins, S. (2014). Situational influences on rhythmicity in speech, music, and their interaction. *Philos Trans R Soc Lond B Biol Sci*, 369(1658):20130398.

Hay, J. (2018). Sociophonetics: The Role of Words, the Role of Context, and the Role of Words in Context. *Topics in cognitive science.*

Hay, J., Drager, K., and Gibson, A. (2018). Hearing r-sandhi: The role of past experience. *Language*, 94(2):360–404.

Hay, J. and Foulkes, P. (2016). The evolution of medial /t/ over real and remembered time. *Language*, 92(2):298–330.

Hay, J. and Maclagan, M. (2012). r/-sandhi in early 20th century New Zealand English. *Linguistics*, 50(4):745–763.

Hay, J., Maclagan, M., and Gordon, E. (2008). *New Zealand English.* Dialects of English. Edinburgh University Press, Edinburgh.

Hay, J., Nolan, A., and Drager, K. (2006a). From Fush to Feesh: Exemplar Priming in Speech Perception. *The Linguistic Review*, 23:351–379.

Hay, J., Podlubny, R., Drager, K., and McAuliffe, M. (2017). Car-talk: Location-specific speech production and perception. *Journal of Phonetics*, 65:94–109.

Hay, J. and Sudbury, A. (2005). How rhoticity became /r/-sandhi. *Language*, 81:799–823.

Hay, J., Walker, A., Sanchez, K., and Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLoS One*, 14(2):e0210793.

Hay, J., Warren, P., and Drager, K. (2006b). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34:458–484.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., and LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, 139:83.

Hazen, K. (2001). An introductory investigation into bidialectalism. *University of Pennsylvania Working Papers in Linguistics*, 7(3):8.

Hebdige, D. (1979). *Subculture: the meaning of style.* Routledge, London.

Heinrich, A., Knight, S., and Hawkins, S. (2015). Influences of word predictability and type of masker noise on intelligibility of sung text in live concerts. *The Journal of the Acoustical Society of America*, 138(4):2373–2386.

Hess, M. (2009). *Hip Hop in America: A Regional Guide.* Greenwood Press.

Heuer, S. (2017). *"He's like Sheffield's Elvis" – A Diachronic Analysis of the Phonetic Performance of Alex Turner Then and Now.* PhD thesis.

Hickey, R. (2017). Analysing Early Audio Recordings. In Hickey, R., editor, *Listening to the Past: Audio Records of Accents of English*, pages 1–12. Cambridge University Press, Cambridge, U.K.

Hickok, G. (2010). The role of mirror neurons in speech and language processing. *Brain and Language*, 112(1):1–2.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Hintzman, D. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review*, 93(4):411–428.

Hoar, P. (2012). *Hearing the World: Audio Technologies and Listening in New Zealand, 1879–1939*. PhD thesis, University of Auckland.

Hoch, L., Poulin-Charronnat, B., and Tillmann, B. (2011). The influence of task-irrelevant music on language processing: syntactic and semantic structures. *Frontiers in psychology*, 2:112.

Holt, L. L. (2006). Speech categorization in context: Joint effects of nonspeech and speech precursors. *The Journal of the Acoustical Society of America*, 119(6):4016–4026.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Hymes, D. H. (1968). The Ethnography of Speaking. In Fishman, J. A., editor, *Readings in the Sociology of Language*, pages 99–138. Mouton, The Hague.

IFPI (2018). *Global Music Report 2018*.

Irvine, J. T. (2001). Style as distinctiveness: The culture and ideology of linguistic differentiation. In Eckert, P. and Rickford, J. R., editors, *Style and Sociolinguistic Variation*. Cambridge University Press, Cambridge.

Jaffe, A. (2015). Staging language on Corsica: Stance, improvisation, play, and heteroglossia. *Language in Society*, 44(2):161–186.

Jansen, L. and Westphal, M. (2017). Rihanna Works Her Multivocal Pop Persona: A Morpho-syntactic and Accent Analysis of Rihanna's Singing Style. *English Today*, pages 1–10.

Jesse, A. and Massaro, D. W. (2010). Seeing a singer helps comprehension of the song's lyrics. *Psychon Bull Rev*, 17(3):323–8.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K. and Mullennix, J. W., editors, *Talker variability in speech processing*, pages 145–166.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity in phonology. *Journal of Phonetics*, 34:485–499.

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, 27(4):359–384.

Jung, O. (2012). On the Lombard effect induced by vehicle interior driving noises, regarding sound pressure level and long-term average speech spectrum. . *Acta Acustica United with Acustica*, 98(2):334–341.

Kennedy, M. (2006). *Variation in the Pronunciation of English by New Zealand School Children*. Ma, Victoria University of Wellington.

Kim, J. (2016). Perceptual Associations Between Words and Speaker Age. *Laboratory Phonology*, 7(1):18.

Kim, J. and Drager, K. (2017). Sociophonetic realizations guide subsequent lexical access. In *Interspeech*.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., and Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature neuroscience*, 7(3):302–307.

Konert-Panek, M. (2017a). Americanisation versus Cockney: Stylisation in Amy Winehouse's singing accent. In Kennedy, V. and Gadpaille, M., editors, *Ethnic and Cultural Identity in Music and Song Lyrics*, pages 77–94. Cambridge Scholars.

Konert-Panek, M. (2017b). Overshooting Americanisation. Accent stylisation in pop singing–acoustic properties of the bath and trap vowels in focus. *Research in Language*, 15(4):371–384.

Konert-Panek, M. (2018). *Singing accent Americanisation in the light of frequency effects: LOT unrounding and PRICE monophthongisation in focus*, volume 16.

Kraus, N. and White-Schwoch, T. (2015). Unraveling the Biology of Auditory Learning: A Cognitive–Sensorimotor–Reward Framework. *Trends in Cognitive Sciences*, 19(11):642–654.

Kutas, M. and Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62(1):621–647.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13).

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, D.C.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Labov, W. (2011). *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors*. Wiley.

Labov, W., Ash, S., and Boberg, C. (2006). *Atlas of North American English*. Mouton de Gruyter, Berlin.

Ladefoged, P. and Broadbent, D. E. (1957). Information Conveyed by Vowels. *Journal of the Acoustical Society of America*, 29(1):98–104.

Le Page, R. B. and Tabouret-Keller, A. (1985). *Acts of Identity: Creole-based Approaches to Language and Ethnicity*. Cambridge University Press, Cambridge.

Legare, C. H. and Nielsen, M. (2015). Imitation and Innovation: The Dual Engines of Cultural Learning. *Trends in Cognitive Sciences*, 19(11):688–699.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Loudermilk, B. (2013). Psycholinguistic Approaches. In Bayley, R., Cameron, R., and Lucas, C., editors, *The Oxford Handbook of Sociolinguistics*. Oxford.

Low, B. E. (2007). Hip-Hop, Language, and Difference: The N-Word as a Pedagogical Limit-Case. *Journal of Language, Identity & Education*, 6(2):147–160.

Luce, P. A. and Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, 26(4):708–715.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear & Hearing*, 19(1):1–36.

Macfarlane, A. H. and Macfarlane, S. (2018). Toitū te mātauranga : valuing culturally inclusive research in contemporary times. *Psychology Aotearoa*, 10(2):71–76.

Mageau, M., Mekik, C., Sokalski, A., and Toivonen, I. (2019). Detecting Foreign Accents in Song. *Phonetica*, pages 1–19.

Marsden, S. (2017). Are New Zealanders "rhotic"? *English World-Wide*, 38(3):275–304.

Marsden, S. and Holmes, J. (2014). Talking to the Elderly in New Zealand Residential Care Settings. *Journal of Pragmatics: An Interdisciplinary Journal of Language Studies*, 64:17–34.

Martin, C. D., Garcia, X., Potter, D., Melinger, A., and Costa, A. (2015). Holiday or vacation? The processing of variation in vocabulary across dialects. *Language, Cognition and Neuroscience*, pages 1–16.

McGowan, K. B. (2015). Social Expectation Improves Speech Perception in Noise. *Language and Speech*, 58(4):502–521.

McLeod, K. (1999). Authenticity Within Hip-Hop and Other Cultures Threatened with Assimilation. *Journal of Communication*, 49(4):134–150.

Meyer, D. E. and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

Mitchell, T. (2008). Doin' damage in my native language: The use of "resistance vernaculars" in hip hop in France, Italy, and Aotearoa/New Zealand. *Popular Music and Society*, 24(3):41–54.

Mithen, S. (2011). *The Singing Neanderthals: The Origins of Music, Language, Mind and Body.* Orion Publishing Group.

Mufwene, S. (1996). The founder principle in creole genesis. *Diachronica*, 13:83–134.

Murphey, T. (1992). The Discourse of Pop Songs. *TESOL Quarterly*, 26(4):770–774.

Näätänen, R. and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425.

Needle, J. M. and Pierrehumbert, J. B. (2018). Gendered associations of English morphology. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 9(1).

Nieuwland, M. S. and Van Berkum, J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18:1098–1111.

Nobre, A. C. and Rohenkohl, G. (2014). Time for the Fourth Dimension in Attention. In Nobre, A. C. and Kastner, S., editors, *The Oxford Handbook of Attention.* Oxford University Press, Oxford.

NZHerald (2016). *Aaradhna claims racism, gives away Tui at NZ Music Awards.* Radio New Zealand.

O'Hanlon, R. (2006). Australian Hip Hop: A Sociolinguistic Investigation. *Australian Journal of Linguistics*, 26(2):193–209.

Oware, M. (2014). (Un)conscious (popular) underground: Restricted cultural production and underground rap music. *Poetics*, 42:60–81.

Oxenham, A. J. and Plack, C. J. (1998). Suppression and the upward spread of masking. *The Journal of the Acoustical Society of America*, 104(6):3500–3510.

Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(2):309–328.

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, 4.

Patel, A. D. (2012). The OPERA hypothesis: assumptions and clarifications. *Annals of the New York Academy of Sciences*, 1252:124–128.

Pennycook, A. (2007). Language, Localization, and the Real: Hip-Hop and the Global Spread of Authenticity. *Journal of Language, Identity & Education*, 6(2):101–115.

Pennycook, A. and Mitchell, T. (2009). Hip hop as dusty foot philosophy: Engaging locality. In Alim, H. S., Ibrahim, A., and Pennycook, A., editors, *Global linguistic flows: hip hop cultures, youth identities, and the politics of language*, pages 25–42. Routledge, New York and London.

Peretz, I. and Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, 6(7):688–691.

Peretz, I., Gagnon, L., Hebert, S., and Macoir, J. (2004). Singing in the Brain: Insights from Cognitive Neuropsychology. *Music Perception*, 21(3):373–390.

Peterson, G. E. and Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, 24(2):175.

Pichler, P. and Williams, N. (2016). Hipsters in the hood: Authenticating indexicalities in young men's hip-hop talk. *Language in Society*, 45(4):557–581.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. . P. J. H., editor, *Frequency Effects and Emergent Grammar*. John Benjamins, Amsterdam.

Pierrehumbert, J. (2002). Word-specific phonetics. In Gussenhoven, C. and Warner, N., editors, *Laboratory Phonology VII*. Mouton de Gruyter, Berlin.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, 34(4):516–530.

Pierrehumbert, J. B. (2016). Phonological Representation: Beyond Abstract Versus Episodic. *Annual Review of Linguistics*, 2(1):33–52.

Pierson, L. L., Gerhardt, K. J., Rodriguez, G. P., and Yanke, R. B. (1994). Relationship between outer ear resonance and permanent noise-induced hearing loss. *American Journal of Otolaryngology*, 15(1):37–40.

Pinnow, E. and Connine, C. M. (2014). Phonological variant recognition: Representations and rules. *Language and speech*, 57(1):42–67.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Podlubny, R. (2019). *Acoustic Convergence: Exploring the Influence of Ambient Noise on Speech Production.* PhD thesis.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10):2128–2148.

Politzer-Ahles, S. and Piccinini, P. (2018). On visualizing phonetic data from repeated measures experiments with multiple random effects. *Journal of Phonetics*, 70:56–69.

Pollitzer, D. (2009). *Grammatical Features in Pasifika English.* Magisterarbeit, University of Regensburg.

Potter, J. (1998). *Vocal Authority: Singing Style and Ideology.* Cambridge University Press, Cambridge.

Poulin-Charronnat, B., Bigand, E., Madurell, F., and Peereman, R. (2005). Musical structure modulates semantic priming in vocal music. *Cognition*, 94(3):B67–78.

Pufahl, A. and Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cogn Psychol*, 70:1–30.

PWC (2018). *Economic contribution of the music industry in New Zealand 2017*.

Racette, A. and Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory & Cognition*, 35(2):242–253.

Rácz, P. (2013). *Salience in Sociolinguistics*. De Gruyter Mouton, Berlin/New York.

Racz, P., Hay, J. B., and Pierrehumbert, J. B. (2017). Social Salience Discriminates Learnability of Contextual Cues in an Artificial Language. *Front Psychol*, 8:51.

Repp, B. H. (2007). Tapping to a Very Slow Beat: A Comparison of Musicians and Nonmusicians. *Music Perception*, 24(4):367–376.

Sackett, S. J. (1979). Prestige Dialect and the Pop Singer. *American Speech*, 54(3):234–237.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294):1926–1928.

Schellenberg, E. G. and Peretz, I. (2008). Music, language and cognition: unresolved issues. *Trends in Cognitive Sciences*, 12(2):45–46.

Schneider, E. (2007). *Postcolonial English: Varieties around the World*. Cambridge University Press.

Schneider, E. and Kortmann, B. (2004). *A handbook of varieties of English: a multimedia reference tool*, volume 1. Mouton de Gruyter, Berlin New York.

Schneider, E. W. (2003). The Dynamics of New Englishes: From Identity Construction to Dialect Birth. *Language*, 79(2):233–281.

Schulze, K., Vargha-Khadem, F., and Mishkin, M. (2012). Test of a motor theory of long-term auditory memory. *Proceedings of the National Academy of Sciences*, 109(18):7121–7125.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656.

Shuker, R. and Pickering, M. (1994). Kiwi rock: Popular music and cultural identity in New Zealand. *Popular Music*, 13(3):261–278.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language and Communication*, 23:193–229.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366.

Simpson, P. (1999). Language, culture and identity: With (another) look at accents in pop and rock singing. *Multilingua*, 18(4):343–367.

Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, 10(1):2465.

Sneller, B., Fruehwald, J., and Yang, C. (2019). Using the Tolerance Principle to predict phonological change. *Language Variation and Change*, 31(1):1–20.

Squires, L. (2014). From TV Personality to Fans and Beyond: Indexical Bleaching and the Diffusion of a Media Innovation. *Journal of Linguistic Anthropology*, 24(1):42–62.

Squires, L. (2018). Genre and linguistic expectation shift: Evidence from pop song lyrics. *Language in Society*, pages 1–30.

Stæhr, A. and Madsen, L. M. (2017). 'Ghetto language' in Danish mainstream rap. *Language & Communication*, 52:60–73.

Starks, D., Gibson, A., and Bell, A. (2015). Pasifika Englishes in New Zealand. In Williams, J. P., Schreier, D., Trudgill, P., and Schneider, E., editors, *Further Lesser-Known Varieties of English*. Cambridge University Press, Cambridge, UK.

Staum, L. (2008). *Experimental Investigations of Sociolinguistic Knowledge*. PhD thesis, Stanford University.

Stone, R. E. E., Cleveland, T. F., and Sundberg, J. (1999). Formant Frequencies in Country Singers' Speech and Singing. *Journal of Voice*, 13(2):161–167.

Strand, E. A. (1999). Uncovering the Role of Gender Stereotypes in Speech Perception. *Journal of Language and Social Psychology*, 18(1):86–100.

Stuart-Smith, J., Pryce, G., Timmins, C., and Gunter, B. (2013). TELEVISION CAN ALSO BE A FACTOR IN LANGUAGE CHANGE: EVIDENCE FROM AN URBAN DIALECT. *Language*, 89(3):501–536.

Sumner, M., Kim, S. K., King, E., and McGowan, K. B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, 4:1015.

Sweetland, J. (2002). Unexpected but authentic use of an ethnically–marked dialect. *Journal of Sociolinguistics*, 6(4):514–538.

Taylor, C. (2011). *Power to Represent: The Spatialized Politics of Style in Houston Hip Hop*. PhD thesis, Rice University.

Todd, S., Pierrehumbert, J. B., and Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, 185:1–20.

Trudgill, P. (1983). *On Dialect: Social and Geographical Perspectives*. Blackwell, Oxford.

Trudgill, P. (1986). *Dialects in contact*, volume 10. B. Blackwell, New York, Oxford.

Trudgill, P. (2004). *New-Dialect Formation: The Inevitability of Colonial Englishes*, volume 1. Edinburgh University Press, Henderson, Auckland;Edinburgh;.

Trudgill, P. (2014). Diffusion, drift, and the irrelevance of media influence. *Journal of Sociolinguistics*, 18(2):213–222.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual review of psychology*, 53(1):1–25.

Uffmann, C. (2007). Intrusive [r] and optimal epenthetic consonants. *Language Sciences*, 29(2):451–476.

Van Berkum, J. J., van den Brink, D., Tesink, C. M., Kos, M., and Hagoort, P. (2008). The neural integration of speaker and message. *J Cogn Neurosci*, 20(4):580–91.

van Besouw, R. M., Howard, D. M., and Ternstrom, S. (2005). Towards an understanding of speech and song perception. *Logoped Phoniatr Vocol*, 30(3-4):129–35.

van den Brink, D., Van Berkum, J. J., Bastiaansen, M. C., Tesink, C. M., Kos, M., Buitelaar, J. K., and Hagoort, P. (2012). Empathy matters: ERP evidence for inter-individual differences in social language processing. *Soc Cogn Affect Neurosci*, 7(2):173–83.

van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (1988). Effects of Noise on Speech Production: Acoustic and Perceptual Analyses. *Journal of the Acoustical Society of America*, 84(3):917–928.

Villarreal, D., Papp, V., Clark, L., Hay, J., and Watson, K. (2019). *Telling a new story with old data: Random-forest classification of non-prevocalic (r) in Southland New Zealand English [2019 VALP presentation]*.

Walker, A. and Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, 2(1).

Warren, P., Gibson, A., and Hay, J. (2017). The sound of women in New Zealand English. In Marra, M. and Warren, P., editors, *Linguist at Work: Festschrift for Janet Holmes*. Victoria University Press, Wellington.

Watts, R. J. and Andres Morrissey, F. (2019). *Language, the Singer and the Song: The Sociolinguistics of Folk Performance*. Cambridge University Press, Cambridge.

Wedel, A. and Fatkullin, I. (2017). Category competition as a driver of category contrast. *Journal of Language Evolution*, 2(1):77–93.

Wells, J. C. (1982). *Accents of English*. Cambridge University Press, Cambridge.

Wenger, E. (1998). *Communities of Practice*. Cambridge University Press, Cambridge.

Werker, J. F. and Hensch, T. K. (2015). Critical Periods in Speech Perception: New Directions. volume 66, pages 173–196. ANNUAL REVIEWS, PALO ALTO.

Werner, V. (2018). *The Language of Pop Culture.* Taylor & Francis.

Westphal, M. (2018). Pop Culture and the Global Spread of Non-Standardized Varieties of English: Jamaican Creole in German Reggae Subculture. In Werner, V., editor, *The Language of Pop Culture*, pages 95–115. Routledge, New York.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.*

Williams, Q. (2017). *Remix Multilingualism: Hip Hop, Ethnography and Performing Marginalized Voices.* Bloomsbury Publishing.

Wilson, G. (2017). Conflicting language ideologies in choral singing in Trinidad. *Language & Communication*, 52:19–30.

Woolard, K. A. (2008). Why dat now?: Linguistic-anthropological contributions to the explanation of sociolinguistic icons and change. *Journal of Sociolinguistics*, 12(4):432–452.

Yaeger-Dror, M. (1991). Linguistic evidence for social psychological attitudes: Hyperaccommodation or (r) by singers from a Mizrahi background. *Language and Communication*, 11(4):309–331.

Yang, C. (2016). The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language.

Yang, J. H. (2018). 'I Want to be New and Different. Anything I'm Not.' Accent-Mixing in Singing. *Australian Journal of Linguistics*, 38(2):183–204.

Zatorre, R. J. (2012). Beyond auditory cortex: working with musical thoughts. *Annals of the New York Academy of Sciences*, 1252(1):222–228.

Zemke-White, K. (2008). Nesian styles (re)present r'n' b: The appropriation, transformation and realization of contemporary r'n'b with hip hop by urban pasifika groups in aotearoa.

Appendix A

PoPS Corpus: Detailed List of Songs

Artist	Vocalist(s) Analysed	Song	Ethnicity
Māori/Pasifika female pop			
Aaradhna		Forever Love	Indian/Samoan
Anika, Boh And Hollie	Boh Runga	Be Mine	Chinese Malaysian/Māori
Bic Runga		Close Your Eyes	Chinese Malaysian/Māori
Brooke Fraser		Kings And Queens	Fijian/Pākehā
Brooke Fraser		The Dead Don't Dance (IV Fridays)	Fijian/Pākehā
Brooke Fraser		Therapy	Fijian/Pākehā
Clicks	Anna Coddington	Resolution	Māori
Diaz Grimm feat. Raiza Biza, Iva Lankum	Iva Lankum	Foreigners	Samoan/Chinese/Pākehā
Jackie Thomas		It's Worth It	Māori
Jackie Thomas		Until The Last Goodbye	Māori
Ladi6		Beffy	Samoan
Ladi6		Royal Blue	Samoan
Parri$	Parris Goebel	Friday	Samoan
Ria feat. Jagarizzar	Ria Numia	Knocking	Samoan
Ria feat. Spawnbreezie	Ria Numia	Winner	Samoan
Ria Hall		Love Will Lead Us Home	Māori
Ria Hall feat. Che Fu	Ria Hall	Tell Me	Māori
Shapeshifter feat. Anika Moa	Anika Moa	Blazer	Māori
Theia		Roam	Māori /Pākehā (PC)
Theia		Treat You	Māori /Pākehā (PC)
Pākehā female pop			
Chelsea Jade		Ride Or Cry	Pākehā (PC)
Ginny Blackmore		Under My Feet	Pākehā
Jamie Mcdell		Fly Honeys	Pākehā
Kimbra		Sweet Relief	Pākehā
Kylie Price		Here With Me	Filipino/Spanish/English/New Zealand (PC)
Ladyhawke		A Love Song	Pākehā
Lili Bayliss		Tainted Love/Sweet Dreams	NZ Pākehā (PC)
Lorde		Yellow Flicker Beat	Pākehā
Lorde		Perfect Places	Pākehā
Lorde		Royals	Pākehā
Mae Valley	Multiple (Abby, Hannah)	Brightside	Pākehā
Nakita		In The Water	Swiss/English (PC)
Sachi feat. Nīka	Nīka	Shelter	Pākehā
Stack And Piece Feat Helen Corry	Helen Corry	Burning Out	Pākehā (PC)
Youmi Zouma	Christie Simpson	December	Pākehā /NZ European (PC)
Māori/Pasifika male pop			
Beau Monga		King And Queen	Māori
Benny Tipene		Lanterns	Māori
Benny Tipene		This Is Where Love...	Māori
Dennis Marsh And Friends	Dennis Marsh	Christmas In New Zealand	Māori
Jason Kerrison		You Want Me As Me	Māori
Jordi Webber		I'll Be Loving You	NZ Māori European (PC)
Kora	Multiple (Stuart, Francis)	Carolina	Māori
Modern Maori Quartet	Multiple (James, Matariki, Maaka, Francis)	Don't Fall In Love	Māori
Six60	Matiu Walters	Don't Give It Up	Māori
Sons Of Zion	Multiple (Rio, Samuel, Joel)	Hungover	Māori /Pākehā (PC)
Sons Of Zion Feat Slip-On Stereo	Multiple (Sons of Zion)	Now	Māori /Pākehā (PC)
Sons Of Zion feat. Israel Starr	Multiple (Sons of Zion)	Stuck On Stupid	Māori /Pākehā (PC)
Stan Walker		You Never Know	Māori
Stan Walker		New Takeover	Māori
Stan Walker Feat Samantha Jade	Stan Walker	Start Again	Māori
Vince Harder Feat Ryan Nz	Vince Harder	Give This A Try	Samoan-German/Pākehā
Warren Maxwell		Moments	Māori
Pākehā male pop			
Balu Brigada	Henry Beasley	Overlap	Pākehā (PC)
Drax Project	Shaan Singh	So Lost	New Zealander (PC)
L.A. Women		Hurricane Love	Pākehā
Maala		In The Air	Pākehā
Maala		Kind Of Love	Pākehā
Maala		In My Head	Pākehā
Matthew Young		Collect	Pākehā
Mitch James		No Fixed Abode	New Zealand European (PC)
Mitch James		Move On	New Zealand European (PC)
Nomad	Multiple (Will, Aasha, Cullen)	Oh My My	Kiwi (PC)
Nomad	Multiple	Love Will Call	Kiwi (PC)
Nomad	Multiple	I Won't Stop	Kiwi (PC)
Salmonella Dub	David Deakins	Searching For The Sun	Pākehā
Stevie Tonks		Give Me Love	Pākehā
Summer Thieves	Jake Bartos	Coast Roads	Kiwi (PC)
Theslacks	Mark Armstrong	Big Aroha	Ngati Pākehā (PC)
Māori/Pasifika male hip hop			
9-5ers feat. Tyra Hammond	Sabe	Talking To You	Afakasi Kiwi (PC)
Bobandii		Nazarite	NZ Māori /NZ Euro (PC)
David Dallas		Don't Rate That	Samoan And European
David Dallas		Probably	Samoan And European
Deach Feat Pt	Deach	Slow Motion	Full Samoan (PC)
Diaz Grimm feat. Raiza Biza, Iva Lankum	Diaz Grimm	Foreigners	Māori (PC)
Eno X Dirty	Manu Walters	Shampoo And Conditioner	Māori And Welsh (PC)
Lukas feat. Lunar	Lunar	Motions	Māori /Pasifika
Noah Slee feat. Melodownz	Melodownz	Lips	Samoan/English/Māori (PC)
PNC Feat Nylo	PNC	If It Wasn't For Love	Samoan/European
Ria Hall feat. Che Fu	Che Fu	Tell Me	Māori /Niuean
Savage And Tigermonkey	Savage	Zooby Doo	Samoan
Sesh feat. Pt	Sesh	Come Through	Māori
Sid Diamond feat. Donell Lewis And Mikey Dam	Sid Diamond	Problems	Cook Islands Māori /NZ Māori
Sid Diamond feat. Donell Lewis And Mikey Dam	Mikey Dam	Problems	Māori /Tongan/Australian (PC)
SWIDT	Multiple (INF, Spycc)	Little Did She Know	Māori /Pasifika
Timmy Trumpet & Savage	Savage	Freaks	Samoan
Tom Francis feat. Royce Da 5'9"; Tyler Thomas	Tom Francis	What I'm Made For	Māori /European (PC)
Ty feat. Mikey Mayz	Ty	Discovery	Niuean
Pākehā male hip hop			
@Peace	Tom Scott	Flowers	Palagi
9-5ers feat. Tyra Hammond	Edgar	Talking To You	Kiwi (PC)
Dark Tower	Jody Lloyd	Alright Now	New Zealand European (PC)
Donell Lewis feat. Rickey Okay	Rickey Okay	Put It On Me	Full Indian (PC)
Homebrew	Tom Scott	Yellow Snot Funk	Palagi
Lukas		Downfall	Pākehā
Lukas feat. Lunar	Lukas	Motions	Pākehā (PC)
Machete Clan	Multiple (Isaac, Roman, Alex)	On The Rark	NZ European (PC)
Name UL		My Side	Greek NZ (PC)
Name UL		Nice Guys Finish Thirst	Greek NZ (PC)
Pillow T.		Ride	Philippine Spanish/Kiwi of Scottish Decent (PC)
Tiki Taane feat. Ria Hall, Maitreya	Maitreya	Falling Angels	Pākehā
Times X Two	Zee	Run	Iraq New Zealander

Table A.1: NZ songs in PoPS corpus, including ethnicity information (sometimes self-reported, PC).

Artist	Vocalist(s) Analysed	Song	US Region
African American female pop			
Beyonce		"7/11"	South
Beyonce		Hold Up	South
Cherish feat. Sean Paul		Do It To It	Midwest
Destiny's Child		Emotion	Midwest
Eminem feat. Beyonce	Beyonce	Walk On Water	South
Fifth Harmony	Normani	He Like That	South
Fifth Harmony	Normani	That's My Girl	South
G.R.L.	Simone Battle	Ugly Heart	West
Janelle Monae		Make Me Feel	South
Janet Jackson feat. Nelly	Janet Jackson	Call On Me	Midwest
Kelly Rowland		Stole	South
Keri Hilson		Pretty Girl Rock	South
Khalid and Normani	Normani	Love Lies	South
TLC	Tionne Watkins	Girl Talk	Mixed
Willow		Whip My Hair	West
European American female pop			
Alan Walker feat. Noah Cyrus and Digital Farm Animals	Noah Cyrus	All Falls Down	South
Ariana Grande feat. Nicki Minaj	Ariana Grande	Side To Side	South
Bebe Rexha and Florida Georgia Line	Bebe Rexha	Meant To Be	East
Frenship and Emily Warren	Emily Warren	Capsize	East
Galantis and ROZES	Rozes	Girls On Boys	East
Katy Perry feat. Nicki Minaj	Katy Perry	Swish Swish	West
Kelly Clarkson		Love So Soft	South
Linkin Park feat. Kiiara	Kiiara	Heavy	Midwest
Macklemore feat. Kesha	Kesha	Good Old Days	Mixed
Macklemore feat. Skylar Grey	Skylar Grey	Glorious	Midwest
Maggie Lindemann		Pretty Girl (Cheat Codes X Cade Remix)	South
Miley Cyrus		Thinkin'	South
Pink		What About Us	East
Taylor Swift feat. Ed Sheeran and Future	Taylor Swift	End Game	East
the Chainsmokers feat. Phoebe Ryan	Phoebe Ryan	All We Know	East
African American male pop			
Calvin Harris feat. Pharrell Williams, Katy Perry and Big Sean	Pharrell Williams	Feels	South
Frank Ocean		Provider	Mixed
Frank Ocean		Ivy	Mixed
Jason Derulo		Kiss the Sky	South
Jason Derulo		Want To Want Me	South
Jason Derulo feat. French Montana	Jason Derulo	Tip Toe	South
Lloyd feat. Andre 3000	Andre 3000	Dedication To My Ex (Miss That)	South
Lunchmoney Lewis		Bills	South
Michael Jackson feat. Justin Timberlake	Michael Jackson	Love Never Felt So Good	Midwest
MKTO	Malcolm David Kelley	Thank You	West
N.E.R.D and Kendrick Lamar	Pharrell Williams	Don't Don't Do It!	South
Pharrell Williams		Happy	South
Sean Kingston feat. Chris Brown and Wiz Khalifa	Sean Kingston	Beat It	South
Will.I.Am feat. Cody Wise	Will I Am	It's My Birthday	West
Will.I.Am feat. Cody Wise	Cody Wise	It's My Birthday	East
European American male pop			
Andy Grammer		Fresh Eyes	East
Charlie Puth		How Long	East
DNCE feat. Nicki Minaj	Joe Jonas (Dnce)	Kissing Strangers	Other
Frenship and Emily Warren	Frenship	Capsize	West
Jon Bellion		All Time Low	East
Justin Timberlake		Can't Stop the Feeling!	South
Lauv		I Like Me Better	Mixed
Logan Paul		No Handlebars	Midwest
Logan Paul feat. Why Don't We	Why Don't We	Help Me Help You	Mixed
Logic feat. Ansel Elgort	Ansel Elgort	Killing Spree	East
Macklemore feat. Eric Nally	Eric Nally	Ain't Gonna Die Tonight	Midwest
Maroon 5 feat. SZA	Maroon 5	What Lovers Do	West
Nick Jonas		Find You	Mixed
Nick Jonas feat. Anne-Marie and Mike Posner	Mike Posner	Remember I Told You	Midwest
PRETTYMUCH	Austin, Nick	Would You Mind	Mixed
African American male hip hop			
Big Sean and Metro Boomin feat. Travis Scott	Big Sean	Go Legend	Midwest
Big Sean and Metro Boomin feat. Travis Scott	Travis Scott	Go Legend	South
Brockhampton	Kevin Abstract, Ameer Vann, Merlyn Wood, Dom Mclennon	Boogie	Mixed
BTS feat. Desiigner	Desiigner	MIC Drop (Steve Aoki Remix)	East
Camila Cabello feat. Young Thug	Young Thug	Havana	South
French Montana feat. Swae Lee	Swae Lee	Unforgettable	Mixed
G-Eazy feat. A$AP Rocky and Cardi B	A$Ap Rocky	No Limit	East
Hopsin		Ill Mind of Hopsin 9	West
Jaden Smith		Icon	West
Jeezy feat. J Cole and Kendrick Lamar	Jeezy	American Dream	South
Kendrick Lamar		HUMBLE.	West
Migos and Marshmello	Migos	Danger	South
Post Malone feat. Quavo	Quavo	Congratulations	South
Taylor Swift feat. Ed Sheeran and Future	Future	End Game	South
Trippie Redd feat. Travis Scott	Trippie Redd	Dark Knight Dummo	Midwest
European American male hip hop			
Action Bronson, Mark Ronson and Dan Auerbach	Action Bronson	Standing In the Rain	East
Bliss N Eso feat. Gavin James	MC Bliss Jonathan Notley	Moments	Unknown
Brockhampton	Matt Champion, JOBA	Boogie	South
Eminem feat. Ed Sheeran	Eminem	River	Midwest
G-Eazy and Halsey	G-Eazy	Him and I	West
Lil Peep		Save That Shit	East
Mac Miller feat. Ariana Grande	Mac Miller	My Favorite Part	Midwest
Machine Gun Kelly, X Ambassadors and Bebe Rexha	Machine Gun Kelly	Home	Mixed
Macklemore and Ryan Lewis feat. XP	Macklemore	Brad Pitt's Cousin	West
Macklemore feat. Kesha	Macklemore	Good Old Days	West
Marc E Bassy feat. G-Eazy	G-Eazy	You and Me	West
NF		Let You Down	Midwest
Post Malone		I Fall Apart	Mixed
Post Malone		No Option	Mixed
Yelawolf		Daylight	South

Table A.2: USA songs in PoPS corpus, including region information.

Appendix B

Materials for Phonetic Categorisation Task

Figure B.1: Participant recruitment: Advertisement physically placed around campus.

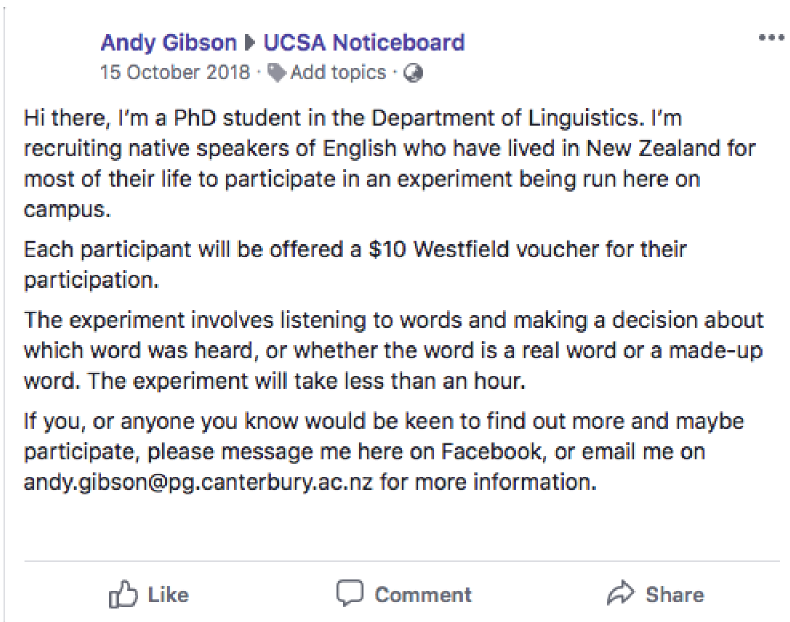


Figure B.2: Participant recruitment: Facebook post.

Perception participants

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Department: Linguistics
Telephone: 021487794
Email: andy.gibson@pg.canterbury.ac.nz

Perception of Words and Non-Words in Variable Contexts

INFORMATION SHEET

Hello,

My name is Andy Gibson and I'm a current postgraduate student in the Department of Linguistics. This study will be included as part of my doctoral thesis, and it explores the effects of context on speech perception.

Your involvement in this project will include two speech perception tasks. In one task, you will be asked to choose which of two words you hear, and press a button to respond. In the second task, you will be asked to listen to words and made-up words and decide whether the word is a real word or not. After completing the experiment, you will be asked to fill in a short survey to provide some basic information about your background. All of this should take less than an hour of your time.

I will ensure the audio is played at a comfortable volume and you may request the volume be lowered if it is too loud for you. You may take breaks in the middle of, and in between each of the tasks if you feel so inclined. Also, you are free to withdraw from the experiment at any time without penalty, and this includes removing your data from the study for any reason, so long as you make such a request prior to leaving this session. You will be offered a $10 Westfield voucher for your participation at the end of the session.

Should you be interested, you may receive a copy of the project results by contacting me at the conclusion of the project.

The results of this project may be published and presented at conferences, and the raw data from your responses will be stored permanently, but you can be assured that all information and opinions gathered during this session will be associated only with an anonymous participation number and will, therefore, be in no way tied to your identity. As mentioned above, results will be included in my doctoral thesis. To be clear, theses are public documents, and a summary of this work will therefore be made available through the UC Library.

As the principle investigator, I am undertaking this study under the supervision of Prof. Jennifer Hay, who can be contacted at jen.hay@canterbury.ac.nz. We are both available to discuss any concerns you may have about participation in the project.

This project has been reviewed and approved by the University of Canterbury Human Ethics Committee, and participants should address any complaints to The Chair, Human Ethics Committee, University of Canterbury, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz).

If you agree to participate, you are asked to complete the consent form and please return it before the taking part in the study.

Many thanks,

Andy Gibson

Figure B.3: Information sheet given to participants prior to commencing experiment.

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Department: Linguistics
Telephone: +64 3 364 2987 ext 8862
Email: andy.gibson@pg.canterbury.ac.nz

Perception of Words and Non-Words in Variable Contexts

CONSENT FORM

I have been given a full explanation of this project and have had the opportunity to ask questions. I also understand what is required of me if I agree to take part in this research.

I understand that participation is voluntary and that I may withdraw at any time without penalty. I also understand that I will be assigned a participant number prior to participation in this study, and that all information and opinions provided by me will be associated only with this participant number; my participation will therefore be anonymous and unidentifiable from the time I leave this session. Moreover, I understand that I have the right to withdraw any information and/or data provided by me so long as I do so prior to leaving this session.

I understand that this experiment will be described in a thesis, and in further publications and conference presentations, and that theses are public documents made available through the UC Library.

I understand that all data collected for this study will be kept in locked and secure facilities and/or in password protected electronic form. Further, I understand that consent forms will be destroyed after 10 years; however, the data from responses and from the survey will be kept indefinitely and may be shared with other researchers and students during future works, presentations, and/or classes. The anonymized data may also be made available online, as an appendix to a published description of the results.

I understand the risks associated with my participation and how they will be managed.

I understand that I am able to receive a report on the findings of this study by contacting the researcher at the conclusion of the project.

Should I have any questions or concerns, I understand that I can contact the researcher Andy Gibson (andy.gibson@pg.canterbury.ac.nz) or his supervisor, Prof. Jennifer Hay (jen.hay@canterbury.ac.nz), for further information. If I have any complaints, I can contact the Chair of the University of Canterbury Human Ethics Committee, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz).

By signing below, I agree to participate in this research project:

_____________________________ _____________________________ (/   /2016)

Name   Signature   Date

Figure B.4: Consent form signed by participants prior to commencing experiment.

Participant Number:

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Department: Linguistics
Telephone: +64 3 364 2987 ext 8862
Email: andy.gibson@pg.canterbury.ac.nz

Perception of Words and Non-Words in Variable Contexts

QUESTIONNAIRE

1. **Age (please circle):** under 20 20–24 25–29 30–34 35–39 40–49 50–59 60+

2. **Gender (please circle):** Female Male Other

3. **Occupation:** ________________________________

4. **Mother's occupation:** ________________________________

5. **Father's occupation:** ________________________________

6. **Have you ever been diagnosed with a hearing impairment?** Yes / No

7. **Are you:** Right-handed / Left-handed

8. **Ethnicity:** ________________________________

9. **Total years spent living <u>outside</u> New Zealand:** less than 2 2–4 4–6 6 or more

10. **List countries, other than New Zealand, that you have lived in for more than a year, giving approximate dates:**

11. **Are you fully fluent in any languages other than English?** Y / N

12. **If so, which language(s):** ____

Figure B.5: Questionnaire given to participants after having completed the experiment (page 1).

Participant Number:

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

13. Do you consider yourself to be a musician? Y / N

14. Circle the genre(s) of music you like listening to the most:

Alternative / Blues / Classical / Country / Electronic / Hip-Hop/Rap / Jazz / Metal /

Pop / R&B/Soul / Rock / Singer/Songwriter

Other: ____

15. How much time do you spend listening to music?
Circle a number on the scale from 1 to 5, where '1' means less than ten minutes per day and '5' means more than three hours per day.

1	2	3	4	5
Less than 10 minutes per day				More than 3 hours per day

16. How much of the music you listen to is by New Zealand artists?
Answer on a scale from 1 to 5, where 1 means none and 5 means all.

1	2	3	4	5
None				All

17. How much of the music you listen to is by American artists?
Answer on a scale from 1 to 5, where 1 means none and 5 means all.

1	2	3	4	5
None				All

18. Do you think it is surprising to hear New Zealand accents in songs?
Answer on a scale from 1 to 5, where 1 means 'not at all surprising' and 5 means 'very surprising'.

1	2	3	4	5
Not at all surprising				Very surprising

19. In the parts of this experiment that had music playing, did it sound like the voices were singing?
Answer on a scale from 1 to 5, where 1 means 'not at all like singing' and 5 means 'very much like singing'.

1	2	3	4	5
Not at all like singing				Very much like singing

20. Do you have any comments about what you think this experiment was about?

Figure B.6: Questionnaire given to participants after having completed the experiment (page 2).

Phoneme Categorisation in Music, Noise and Silence: Canterbury, 2018 (#15017)

Created: 10/11/2018 08:51 PM (PT)
Shared: 10/11/2018 08:51 PM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Higher rates of 'bed' responses will be found in the music category than the noise and silence category for ambiguous stimuli (steps 3–4 of the continuum)

Noise is included as a control, since Lombard-related effects may lead to more 'bed' responses than in silence.

Higher 'bed' responses to first trial of experiment for subjects with music first than those with noise/silence first.

3) Describe the key dependent variable(s) specifying how they will be measured.
Choice of 'bed' or 'bad' on each trial is the dependent measure.

4) How many and which conditions will participants be assigned to?
The design involves 3 within-participant conditions.
Background: music vs. noise vs. silence
Note that all participants begin the experiment with the same stimulus, so that a between-subject analysis can also be tested.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Generalised linear mixed-effects models will be run predicting log-odds of responding 'bed'

Modelling procedure:
1. Initial modelling phase: intercept for participant, no slopes. With all IVs listed below
2. Pruning
IVs removed based on least significance, ratified by anova model comparison. Variables with * tested in 2-way interactions prior to removing, and kept in model until after pruning if interactions approach significance (p<.1).
3. VIF-tests
VIF tested at various points. If highest VIF is greater than 15, variables removed to reduce the multi-collinearity. They may be tested again later once correlated IVs pruned.
4. Adding slopes
After pruning, maximal slope structure added, then slopes removed based on amount of variance explained until the model converges.

• stimulus step (1–6, as a continuous variable)
• stimulus length (short vs. long). NB. The stimulus step and length will also be tested in interaction, and potentially grouped into a single variable
• previous token stimulus step
• block number
• trial number in block. The block*trial interaction will be tested early on and may be included throughout the pruning phase if significant.

Participant questionnaire IVs:
*age group
*gender
class (2 or 3-level factor based on participant and parents' occupations) *handedness
ethnicity (2 or 3 level factor, grouping ethnicities deemed similar in terms of linguistic backgrounds time spent living outside NZ
fluent in other language
*musician
favourite genres of music (grouped into a factor based on common clusters)
time spent listening to music (1-5 scales treated as continuous OR as a high/low binary split)
proportion of music listened to that is by NZ or USA artists. 5-level scales used individually, or converted to 3-level factor (US>NZ; same; NZ>US)
surprising to hear a NZ accent in a song
whether the voices sounded like singing

AsPredicted
Pre-Registration made easy

In addition to this modelling procedure, I will also run models on individual stimuli (ie. continuum steps) that show a high degree of variability in responses. These will be modelled using the same procedure.

Furthermore, a Fishers Exact test will be run on a between-subjects test of the first stimulus encountered, with 12 data points in each of the three conditions.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Participants must have grown up in NZ, defined has having spent less than 5 years total outside of NZ in their life. Participants who have been previously diagnosed with a hearing condition will be excluded.

Participants with mean RT 3SDs above or below mean of participant mean RTs will be removed.
Participants with a % 'bed' response further than 3SDs above or below the mean of participant mean 'bed' responses will be removed.

Responses where RT is <3SDs below or >3SDs above participant mean will be excluded.

Any trial in the music condition where E-Prime reports an onset delay of greater than 75ms will be removed.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
36 participants will be recruited.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
It should be noted that this experiment was previously attempted, and run with 17 participants, but an error was introduced into the E-Prime scripting which caused timing errors with the stimuli in the music condition, rendering those results unusable.

Appendix C

Materials for Lexical Decision Task

Table C.1: Words and nonwords used in lexical decision task, sorted alphabetically within lexical set.

Non-rhotics						Rhotics					
BATH	ZATH	GOAT	ZOAT	LOT	ZOT	START	ZART	NORTH	ZORTH	NURSE	ZURSE
asking	banches	bloated	boapy	boxes	bloppy	arming	barvet	cornet	chorget	burden	blerking
basket	blanting	boating	bozes	brothel	chodded	arty	carching	corpus	chormel	burning	eardest
blasting	blassing	broken	broaded	clotted	chodges	carving	charvist	formal	dorty	certain	eargy
branches	brancing	clothing	broaming	coffee	dronchel	charming	clarny	forty	florcen	curtain	furden
casket	dasping	coping	broselet	coffin	dwotten	farming	garking	hoarding	korty	dirty	furning
casting	dassing	cosy	choathy	costly	flonger	garden	garning	hornet	lorcet	early	gurky
chanting	dastness	doting	croating	coughing	fodding	garnet	glarshest	morning	morbing	earnest	kearthly
clasping	fample	floating	donely	crosses	frodgen	harden	hargy	mortal	morgeous	earthly	kirty
classes	fasking	frozen	foaping	foster	frozzy	harming	narshing	normal	ormet	flirty	mernel
crafting	gasking	hoping	fosted	gotten	gloffle	harpist	parshing	orbit	ortal	journal	purnel
dancing	hasket	hosting	foting	knowledge	gorrow	hearty	plarsy	organ	plorset	journey	purney
drafting	lasket	loaded	fozen	longing	groledge	karma	sarnel	porpoise	shorkel	kernel	sersect
fasten	masty	lonely	frobing	losses	gronest	largest	snarfen	portal	shormet	perfect	snurchen
fasting	pasking	noble	groating	notches	kosken	market	tarma	portrait	skorking	person	verlip
gasping	plancis	noted	hoaken	offer	modging	pardon	tharving	shorten	sorpoise	thirty	vertain
glasses	plasses	oaky	koble	office	obbin	partial	varling	sporty	vorden	worsen	zearning
granted	praffing	ocean	mozing	often	ploxes	parting	varpon	thorny	vording	yearning	zirty
grasping	pranted	phoning	ploguing	prodded	prothy	party	varsen	warden	yormal		
lancet	prasses	poaching	poven	profit	rozzing	sharpen	zarchet	warning	zornet		
lasted	saffing	poky	roaning	quarry	soffice	sharpest	zarden				
laughing	safting	posted	shoated	rotten	sposses	starving	zarmest				
masking	sasten	roading	smoated	sausage	sprollen						
passing	spasking	roses	smoby	slotting	swonches						
planting	splaffing	smoky	smophy	soften	thoffit						
prancing	stanches	soaking	soating	sorrow	thwabble						
rafting	stasking	soapy	spoaming	sorry	twokker						
sample	talving	spoken	swoded	squabble	twoppy						
shafting	tanches	token	thoving	stopping	wothing						
slanted	wancet	trophy	troaky	stronger	yoster						
staffing	zancing	woken	woney	waffle	zosses						
stances	zasking	woven	zoting	washing	zotted						

Appendix C

CONFIDENTIAL - FOR PEER-REVIEW ONLY

AsPredicted
Pre-Registration made easy

Lexical Decision in Music, Noise and Silence: Canterbury, 2018 (#15016)

Created: 10/11/2018 08:42 PM (PT)
Shared: 10/11/2018 08:52 PM (PT)

This pre-registration is not yet public. This anonymized copy (without author names) was created by the author(s) to use during peer-review. A non-anonymized version (containing author names) will become publicly available only if an author makes it public. Until that happens the contents of this pre-registration are confidential.

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
The hypothesis for this lexical decision task is that lexical access will be faster/more accurate to hearing an American English (AmE) accent, relative to a New Zealand English (NZE) accent, in a musical context than in noise or silence. Participants will be NZ natives, and will likely therefore be faster overall to NZE than AmE, and faster in the quiet listening condition than noise or music. But an interaction is hypothesised between condition (music/noise/silence) and voice (AmE/NZE) such that AmE is facilitated or NZE is inhibited in the music condition. Noise is included as a control, in case there are Lombard-related effects.

3) Describe the key dependent variable(s) specifying how they will be measured.
Reaction time for correct responses to words (incorrect responses and responses to non-words will be discarded from the main analysis). Two measures of RT will be tested and reported on: RT from stimulus start, and RT from stimulus end.
These reaction times will be logged, centred and scaled since they are likely to be right-skewed.

Accuracy: accuracy will provide a third DV. This is a binary variable (correct/false) measured for all real-word stimuli (non-words excluded).

4) How many and which conditions will participants be assigned to?
The design involves 6 within-participant conditions based on 2 voices occurring in the context of 3 background noise conditions:
Voice: NZE vs. AmE
Background: music vs. noise vs. silence

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Linear mixed-effects models on RT. Generalised linear mixed effects models on accuracy.

Modelling procedure:
1. Initial modelling phase: intercepts for participant and word, no slopes.
Models begin with all of the main effects listed below, and voice*condition. This interaction may be withheld until other non-significant variables have been pruned from the model.

2. Pruning
IVs removed based on least significance, ratified by anova model comparison. Variables with * tested in 2-way interactions and in 3-way interaction with voice*background prior to removing, and kept in model until after pruning if interactions approach significance (p<.1).

3. VIF-tests
VIF tested at various points. If highest VIF is greater than 15, variables removed to reduce the multi-collinearity. They may be tested again later once correlated IVs pruned.

4. Adding slopes
After pruning, maximal slope structure added, then slopes removed based on amount of variance explained until the model converges.

Main IV:
voice*background

Participant IVs:
*age group
*gender
class (2 or 3-level factor based on participant and parents' occupations)
*handedness
ethnicity (2 or 3 level factor, grouping ethnicities deemed similar in terms of linguistic backgrounds
time spent living outside NZ

Verify authenticity: http://aspredicted.org/blind.php?x=hw4t4k

Version of AsPredicted Questions: 2.00

fluent in other language
*musician
favourite genres of music (grouped into a factor based on common clusters)
time spent listening to music (1-5 scales treated as continuous OR as a high/low binary split)
proportion of music listened to that is by NZ or USA artists. 5-level scales used individually, or converted to 3-level factor (US>NZ; same; NZ>US)
surprising to hear a NZ accent in a song
whether the voices sounded like singing

Word IVs:
length of stimulus
*dialect difference type: rhoticity, BATH, GOAT. Separate models may be run for these dialect difference types

Lexical Frequency IVs:
spoken frequency OR sung frequency
*ratio of sung:spoken frequency (or binary factor based on ratio)

Experimental controls
*block (6-level factor OR continuous OR break into two variables: 1st/2nd voice; block 1,2,3 for voice)
*trial number within block. block*trial*condition will be tested during the pruning phase

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Participants must have grown up in NZ, defined as having spent less than 5 years total outside of NZ in their life. Participants who have been previously diagnosed with a hearing condition will be excluded.

Responses where RT is shorter than 400ms or >3 SDs above participant mean will be excluded.

Any participant with overall accuracy of less than 3 SDs below mean of participant mean accuracies (across both words and non-words) will be removed.

Any trial in the music conditions where E-Prime reports an onset delay of greater than 75ms will be removed.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
36 participants will be recruited.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
It should be noted that this experiment was previously attempted, and run with 17 participants, but an error was introduced into the E-Prime scripting which caused timing errors with the stimuli in the music condition, rendering those results unusable.

Stimulus speakers

Department: Linguistics
Telephone: +64 3 364 2987 ext 8862
Email: andy.gibson@pg.canterbury.ac.nz

Perception of Words and Non-Words in Variable Contexts

INFORMATION SHEET

Hello,

My name is Andy Gibson and I'm a current postgraduate student in the Department of Linguistics. This study will be included as part of my doctoral thesis, and it explores the effects of context on speech perception.

I would like to invite you to participate in this project, by recording some of the stimuli that will be used in the experiments. Your participation would involve recording you reading a list of approximately 300 words and non-words. The researcher will be on hand to help with any pronunciation queries. The recording session will take less than an hour of your time.

The task involves very few risks. Some general information about your background will be given in publications and at conferences, such as your age group, gender and place of origin. It is also possible that your voice may be recognised by some participants. You may choose (on the consent form) whether you are happy for samples of the recordings to be played at conferences or included as supplementary material in publications, or whether you would prefer the recordings to be used solely in the experiment. The recordings of your voice may also be used in follow-up experiments of a similar nature. You are welcome to take breaks at any time during the recording session. Also, you are free to withdraw from the experiment at any time without penalty, and this includes removing your data from the study for any reason, so long as you make such a request prior to leaving this session.

Should you be interested, you may receive a copy of the project results by contacting me at the conclusion of the project.

As the principle investigator, I am undertaking this study under the supervision of Dr. Jennifer Hay, who can be contacted at jen.hay@canterbury.ac.nz. We are both available to discuss any concerns you may have about participation in the project.

This project has been reviewed and approved by the University of Canterbury Human Ethics Committee, and participants should address any complaints to The Chair, Human Ethics Committee, University of Canterbury, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz).

If you agree to participate, you are asked to complete the consent form and please return it before the taking part in the study.

Many thanks,

Andy Gibson

Stimulus speakers

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Department: Linguistics
Telephone: +64 3 364 2987 ext 8862
Email: andy.gibson@pg.canterbury.ac.nz

Perception of Words and Non-Words in Variable Contexts

CONSENT FORM

I have been given a full explanation of this project and have had the opportunity to ask questions. I also understand what is required of me if I agree to take part in this research.

I understand that participation is voluntary and that I may withdraw at any time without penalty. I also understand that my voice will be used in experiments, and that it is possible that some participants will recognize my voice. Moreover, I understand that I have the right to withdraw any information and/or data provided by me so long as I do so prior to leaving this session.

I understand that this experiment will be described in a thesis, and in further publications and conference presentations, and that theses are public documents made available through the UC Library.

I **do / do not** give permission for samples of the recordings to be played at conferences, and/or supplied as supplementary information in publications such as the researcher's doctoral thesis.

I understand that all data collected for this study will be kept in locked and secure facilities and/or in password protected electronic form. Further, I understand that consent forms will be destroyed after 10 years; however, the recordings of my voice will be kept indefinitely and may be shared with other researchers and students during future works, presentations, and/or classes. I also understand that these recordings may be used in follow-up experiments of a similar nature.

I understand the risks associated with my participation and how they will be managed.

I understand that I am able to receive a report on the findings of this study by contacting the researcher at the conclusion of the project.

Should I have any questions or concerns, I understand that I can contact the researcher Andy Gibson (andy.gibson@pg.canterbury.ac.nz) or his supervisor, Prof. Jennifer Hay (jen.hay@canterbury.ac.nz), for further information. If I have any complaints, I can contact the Chair of the University of Canterbury Human Ethics Committee, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz).

By signing below, I agree to participate in this research project:

____________________________ ____________________________ (/ /2016)

Name Signature Date

Appendix D

Detailed Description of LDT Model 3

This additional analysis looks at the reaction time data from the lexical decision task (LDT) from a different perspective, by measuring from the offset of stimuli rather than from the start. This analysis was included in the preregistration since it proved to be a more sensitive measure to the congruence results presented by Walker and Hay (2011) than the RTstart measure. In LDT Model 2 (in Section 4.3.2.2), we saw that participants take longer to react to a longer word, but in this model, the length of the word is bundled into the dependent variable itself by subtracting length from RT.

The variable measuring reaction time from the end of the stimulus (RTend) was created by subtracting from a given reaction time the length of the relevant soundfile. In order to then log, scale and centre the resulting RTend data, a further step of processing needed to be undertaken. Since participants sometimes make their decision before the end of the word, subtracting the length of the audio file from the RT sometimes resulted in negative values. Negative values cannot be logged, and so a constant needed to be added to RTend. It was decided that the average length of the soundfiles was a reasonable choice of constant (747ms) so RTend would be on a comparable scale to RT. Once this constant was added, the modified version of RTend was logged, scaled and centred.

The same strategy was used for deciding which of the preregistered options for each IV would be used. This process resulted in the use of exactly the same versions of all IVs as were used in LDT Model 2. The base model was slightly different, as it was found to explain more variance without the inclusion of slxVar: lmer(RTend.lsc ∼ Condition * Voice + length.sc + (1|Subject) + (1|Stimulus)). The outcome of the comparison process was very similar to that presented for LDT Model 2, but is included here, with the relevant p-values, for the sake of completeness.

Decision-making processes for choosing amongst preregistered versions of IVs for LDT Model 3:

- NZSEI: A two-level factor was better in comparison to the base model (log-likelihood p=0.19) than a tertile split (p=0.43).
- MusicListening: This did not improve the base model significantly either as a continuous variable (p=0.83) or as a factor (p=0.75), but the binary split was the better of the two.
- NZsurpris: This improved fit better when modeled as a continuous variable (p=0.56) than as a binary factor (p=0.99)

- StimSingy: Whether people thought the stimuli in the music conditions sounded like they were sung improved fit better when modeled as a binary factor (p=0.57) than as a continuous predictor (p=0.95)

- NZ vs. US music listening: Amount of US music listening improved the model more (p=0.21) than amount of NZ music listening (p=0.99). The three-way factor USNZMusic (US-dom, Equal, NZ-dom), however, did better than either of the scales in isolation (p=0.16).

- Lexical frequency: Speech frequency (p=1.7e-05) and song frequency (p=1.6e-09) of the word are both highly predictive of reaction times, with faster responses to high frequency words in both cases. Song frequency was used in model fitting since it had the lower of the two p-values.

- Songiness: The ratio of sung frequency to spoken frequency did not improve the base model when added as a continuous predictor (p=0.72), or as a median split of this ratio (p=0.32). The two-level factor was used.

- Block: The decision making process for Block was the same as that described above for LDT Model 2. Comparing three-way interactions with Trial and Condition to all component two-way interactions, Block(continuous) did better than Block(6-level factor, p=0.08 vs. p=0.12). The factor was much better than the continuous version, however, when comparing the Block*Trial interaction to the component main effects (p=0.001 for factor, p=0.39 for continuous). Similarly, treating Block as a factor was better for the Block*Condition interaction (p=0.003 for factor, p=0.096 for continuous). Through the same rationale as described for LDT Model 2, the three-level factor, SubBlock, was used.

Using these versions of the IVs, along with all the other IVs listed in the preregistration, the model fitting procedure was conducted, leading to a model with the following structure:

LDTmod3 = lmer(RTend.lsc ∼ Condition * Voice + Gender + songFreq.lsc + slxVar * Voice + Musician * Condition + AgeBinary * SubBlock + SonginessBin * USNZMusic + SonginessBin * length.sc xxcheck + SonginessBin * Trial50 + SubBlock * Trial50 + (1 | Subject) + (1 | Word))

Table D.1 shows the final preregistered model for RT (logged, centred and scaled) as measured from the end of stimuli. The maximum VIF in this model was 7.5, this time for the SubBlock3 term (due to the SubBlock*Trial50 interaction). Several main effects are similar to those found in LDT Model 2, while some of the significant interactions are difficult to interpret and may be due to the modelling procedure favouring the addition of interactions for the least significant variables first. Problems relating to the preregistered model fitting procedure will be discussed in the next section, but first, each of the significant terms in LDT Model 3 will be described.

- Condition by Voice interaction: There is a significant facilitation for the US voice in the Music condition when comparing Music and Noise (p=0.031).

Table D.1: Preregistered LMER model for reaction time from end of stimulus (LDT model 3).

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.222	0.186	52.798	-1.194	0.238
ConditionNoise	0.095	0.040	4575.632	2.390	0.017
ConditionSilence	-0.044	0.040	4578.749	-1.113	0.266
voiceUS	-0.070	0.057	4609.586	-1.223	0.221
GenderM	0.383	0.136	29.976	2.826	0.008
sungFreqs	-0.156	0.031	130.531	-5.056	0.000
slxVarGOAT	-0.135	0.098	171.075	-1.374	0.171
slxVarLOT	-0.019	0.098	174.351	-0.191	0.849
slxVarrhoticity	-0.049	0.087	176.510	-0.570	0.569
Musiciany	-0.302	0.173	33.163	-1.745	0.090
AgeBinaryyounger	0.082	0.156	32.688	0.529	0.600
SubBlock2	-0.077	0.062	4571.547	-1.239	0.216
SubBlock3	-0.270	0.062	4580.502	-4.379	0.000
ratioBinaryspeechy	-0.065	0.087	491.962	-0.752	0.452
USNZMusicNZdom	0.443	0.306	31.831	1.447	0.158
USNZMusicUSdom	-0.131	0.179	31.536	-0.735	0.468
lengths	-0.342	0.025	3028.050	-13.802	0.000
Trial50	0.000	0.001	4573.813	0.147	0.883
ConditionNoise:voiceUS	**0.113**	**0.053**	**4569.422**	**2.159**	**0.031**
ConditionSilence:voiceUS	**0.097**	**0.052**	**4572.612**	**1.882**	**0.060**
voiceUS:slxVarGOAT	0.109	0.073	4694.564	1.504	0.133
voiceUS:slxVarLOT	0.150	0.068	4624.015	2.209	0.027
voiceUS:slxVarrhoticity	0.223	0.065	4692.383	3.447	0.001
ConditionNoise:Musiciany	0.157	0.065	4571.052	2.407	0.016
ConditionSilence:Musiciany	0.076	0.069	4571.367	1.109	0.268
AgeBinaryyounger:SubBlock2	0.126	0.055	4573.649	2.291	0.022
AgeBinaryyounger:SubBlock3	0.108	0.055	4576.027	1.958	0.050
ratioBinaryspeechy:USNZMusicNZdom	-0.000	0.104	4542.980	-0.004	0.997
ratioBinaryspeechy:USNZMusicUSdom	0.115	0.057	4545.253	2.023	0.043
ratioBinaryspeechy:lengths	0.058	0.029	3696.577	1.995	0.046
ratioBinaryspeechy:Trial50	-0.003	0.001	4575.668	-2.213	0.027
SubBlock2:Trial50	-0.001	0.002	4575.953	-0.766	0.443
SubBlock3:Trial50	0.006	0.002	4578.528	3.093	0.002

This effect approaches significance for the difference between Music and Silence ($p=0.06$). When Condition is relevelled in an otherwise identical model, we find that Noise and Silence are not significantly different to one another ($p=0.76$). Figure D.1 plots this interaction, showing that when modelling the data on the basis of RT minus stimulus length, the interaction of Condition by Voice still holds, and looks very similar to that shown for LDTmod2.

- Gender: Males responded more slowly than females ($p=0.008$).
- Frequency: Responses were faster to frequent words ($p<0.001$ xx get actual?).
- Sociolinguistic Variable by Voice interaction: for the reference level of Condition (which is Music), responses are faster to the US voice than the NZ voice for the reference level of slxVar (which is BATH). This facilitation for the US voice is less for GOAT, LOT and rhoticity, respectively. This does not interact with Condition, that is, responses to the US voice are faster than to the NZ

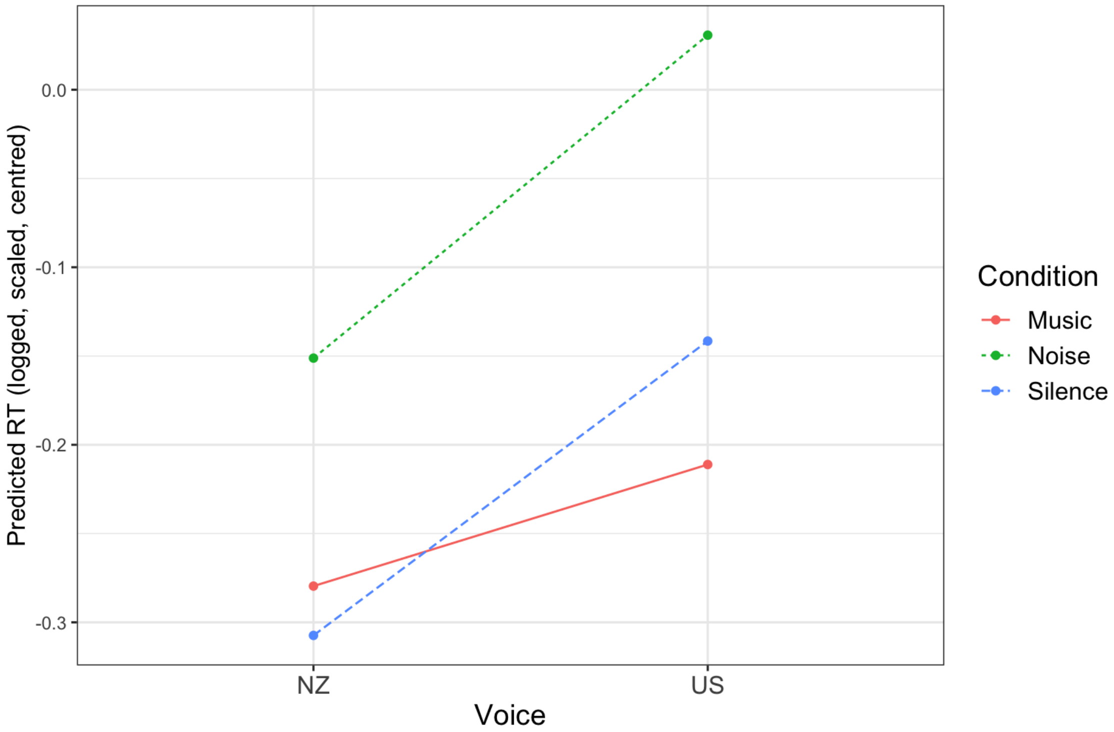


Figure D.1: Interaction of Condition and Voice in the preregistered model with RT measured from the end of stimuli (LDT model 3).

voice to the greatest extent for BATH and to the least extent for rhoticity. It is possible that this relates to differences between the length of the words in the different dialects. For example, it could be that the rhotic variants realised in US English are consistently longer than in NZ English.

- Interaction between Condition and Musician: Musicians have an RT facilitation in the Music condition as compared to non-musicians.

- Age by SubBlock interaction: In this interaction, younger participants are slower than older participants in SubBlock 1 (the reference level), with this difference being even greater in SubBlocks 3 and 2, respectively. The fact that younger participants are predicted to be slower than older participants in this model and faster than older participants in LDT Model 2 is a sign that these interactions are the result of over-fitting and likely examples of Type I errors. It could also be, however, that there is a genuine interaction between age group and whether the RT is measured from the start and end of the stimuli. This could be explored through checking interactions of Age with Length more carefully, but since this is not related to the research questions, nor is Age normally distributed, it will not be further investigated.

- Interaction of Songiness with USNZMusic: 'Equal' is the reference level for USNZMusic, and for that level, responses are faster to speechy words. This effect is the same for those who listen to more NZ music than US music, but for US-dominant listeners, speechy words are responded to more slowly than songy words.

- Interaction of Songiness with Length: The main effect of length in this model is opposite to that shown in LDT Model 2, and this is a predictable outcome of modelling the reaction time from the end of the stimuli. In this model, longer words have faster RTs, because the point from which RT was measured was actually further into the trial for longer words and earlier in the trial for shorter words. This main effect of longer words having faster RTs is slightly mitigated for speechy words.
- Interaction of Songiness with Trial50: In this model, as in LDT Model 2, responses to speechy words (relative to songy words) get faster in the later trials of a given SubBlock.
- Block by Trial interaction: As was seen in LDT Model 2, participants get faster in SubBlocks 2 and 3, but slow down as SubBlock 3 goes on.