2023-09-20

Perceptual Learning of German Sounds: Evidence from Functional Load (FL) and High- Variability Phonetic Training (HVPT)

Suessenbach, Lisa

UNIVERSITY OF CALGARY

Perceptual Learning of German Sounds: Evidence from Functional Load (FL) and High-Variability Phonetic Training (HVPT)

by

Lisa Süßenbach

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN LANGUAGES, LITERATURES AND CULTURES

CALGARY, ALBERTA

SEPTEMBER, 2023

Abstract

The objective of this thesis is to empirically test the practical implications of the functional load (FL) principle in German. The findings informed the selection of German phonemic contrasts for perceptual training of L2 German learners in a follow-up study. Previous research has suggested that sound contrasts carrying a high FL play a central role in conveying meaning, which closely links to the notions of intelligibility and comprehensibility of spoken utterances. Recent attention to FL in second language (L2) English pronunciation pedagogy highlights its role in selecting appropriate L2 sounds to train.

In Study 1, the FL hierarchy of German and the impact of high vs. low FL segments on intelligibility and comprehensibility is tested among L1 German listeners. Results show that high FL errors have a more detrimental effect than low FL errors, but two errors are more severe than one, regardless of FL classification. Study 2 explores two types (i.e., audio and audiovisual) of high-variability phonetic training (HVPT) for challenging German sound contrasts among beginner L2 learners. HVPT employs multiple talkers and variable phonetic environments, thereby enhancing discrimination of sound contrasts. Results showed that especially audiovisual HVPT led to reduced discrimination accuracy, suggesting a need to investigate its use for training beginner learners. These findings shed light upon FL's applicability in conjunction with word recognition models, thereby guiding future work on FL in L2 pronunciation pedagogy. They also provide insights into the theoretical implications of the HVPT technique in fostering perceptual abilities among beginner L2 learners.

Preface

Chapter 2 is entitled “Examining the functional load principle in German.” I am the sole author. The statistical computations of the experimental data have been done by Dr. Tak Fung. The computations of functional load have been provided by Dr. Christophe Coupé.

Chapter 3 is entitled “Effects of audio vs. audiovisual training on the perception of sounds by learners of German.” I am the sole author. The statistical computations of the experimental data have been done by Dr. Tak Fung.

Acknowledgements

It has been an exciting journey to turn my passion for speech sounds into this dissertation. First and foremost, I would like to thank my supervisor, Dr. Mary O'Brien, for her support, patience, wisdom, and commitment to this project: It was hard to believe sometimes that some of your days did not have 48 hours. You are also a fantastic human being.

I would also like to thank my supervisory committee: Dr. Stephen Winters for his valuable insights on the finicky and nerdy stuff, the details, such as minimal pair computations and those speech perception and stats classes I took a while ago, which proved valuable in making this thesis happen. And Dr. Angela George for her helpful comments, enthusiasm, and reliability. I loved having you on board in bringing this project to fruition.

I would also like to thank the Social Sciences and Humanities Research Council of Canada for their financial support without which I would not have been able to carry out the experimental studies for this project.

Special thanks go out to Dr. Tak Fung for statistical computations, Huteng Dai for Python tutorials, Dr. Armin Buch from the University of Tübingen for his help on quantifying phonological similarity, Dr. Christophe Coupé for providing functional load computations, Dr. John Scott for involving me in his research projects and sending me papers to advance my knowledge, Dr. John Levis for his time to discuss issues of functional load and experimental design, Dr. Lisa Hughes for her insightful questions and sparking new ways of thinking about L2 acquisition, Dr. Hildegard Farke from the University of Göttingen for a mock thesis defence, as well as Dr. Murray Munro for his insights on my project.

Last, but not least, this final paragraph goes out to a few important people in my life: My beloved grandma Irmgard and my friend Ulf for their support throughout the master's and PhD years in Canada – the hardships of being a student in this country are real, and you not only made it so much easier for me, you made it happen. I know I am extremely fortunate and privileged to earn this degree. Thank you for housing, tuition fees, and investing in my future. I would also like to thank Thomas for his support and enthusiasm about me becoming a doctor – thank you for telling me all the time how proud you are and sharing ideas on how to achieve my goals. Finally, I would like to thank the rest of my family – even though you had no clue "what this was all about", you celebrated every milestone with me (as a first-generation academic). Some of you were more patient than others, but ultimately, I know you're damn proud of me as well.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

FL =	functional load
CA =	Contrastive Analysis
SLM(-r) =	Speech Learning Model (-revisited)
PAM =	Perceptual Assimilation Model
PAM(-L2) =	Perceptual Assimilation Model(-L2)
NAM =	Neighbourhood Activation Model
ET =	Exemplar Theory
HVPT =	high-variability-phonetic training
LVPT =	low-variability phonetic training
A =	audio-only
AV =	audiovisual
ESL =	English as a second language
L1 =	first language
L2 =	second language

Chapter 1: Introduction

The utterance *"I sink I hurt my no*" presents a scenario where the intended meaning might not immediately be clear. While it is possible to infer that *sink* in this context means *think*, recognizing *no* as *nose* poses a greater challenge. Alternative interpretations, i.e., *toe* or *knee* could be considered as possible referents for the phonetic sequence *no*. All three words *toe*, *nose*, and *knee* differ from the word *no* by a single sound segment only, either by substitution or deletion thereof. However, the acoustic-phonological structure of the utterance may guide our perception toward *nose* before considering *toe* or *knee* as a plausible alternative target. This example illustrates how the substitution of individual sounds within words can significantly impact word recognition and thus impede the overall intelligibility of an utterance. Moreover, even if we understand an utterance, the degree of effort required to comprehend it can vary based on the perceived degree of 'goodness of fit'. While many lexical items resemble *no* phonologically and thus form minimal pairs with it, such as *toe* [to̞], *show* [ʃo̞], *though* [ðo̞], *bro* [bro̞], *dough* [do̞], but also *knee* [niː], *node* [no̞d], *known* [no̞n], *gnome* [no̞m], we can narrow down the options to those lexical neighbours that fit the contextual cues of the utterance (Luce, Goldinger, Auer, & Vitevitch, 2000). Likely, we will opt for a lexical item that is not only a plausible contextual fit but also an acoustic-phonologically good fit (Luce & Pisoni, 1998).

1.1 Applications of the functional load principle in L2 pronunciation pedagogy

The notion of functional load (FL) refers to the importance or significance of a particular sound contrast within the phonemic inventory of a given language. As such, this measure quantifies the degree of functionality of segmental contrasts in their ability to keep the meaning of utterances distinct. As such, a distinction like /n - t/ in English would carry a high functional load, because many word pairs hinge on the discrimination of this phonemic contrast (n= 687): e.g., ***n****ose* –

t*oes, ca****n*** *–* ***ca****t, spi****t*** *–* ***spi****n*, etc. Therefore, this distinction has a high importance and communicative value within the phonological system of the English language. Conversely, a contrast like /s - θ/ distinguishes fewer word pairs (n=96) and thus carries a lower functional load, e.g., ***s****ink –* ***th****ink*, ***s****eam –* ***th****eme, wor****se*** *–wor****th***. In line with the crucial premise of functional load is that those segments that are phonologically distinct, e.g., /n - t/ can perform this contrastive work best because they have little in common, which then ultimately prevents their confusion in real life. If we consider the example of the /n - t/ and /s - θ/ contrasts, the former will unlikely be confused by L2 speakers, whereas the latter has been reported to be commonly confused in production (Brown, 1988; Munro & Derwing, 2006). This means that we have possibly encountered this pattern in real life at some point.

(1) Target: *This thing was tricky*

Low FL substitution: *This *sing was tricky*

Conversely, we have likely not encountered a substitution like the following.

(2) Target: *He wrote a nice note.*

High FL substitution: *He wrote a nice *tote.*

The likelihood of substitution, as illustrated in (1), is higher for /s/ and /θ/ due to their substantial phonological similarity based on shared phonological features and saliency weight (Kondrak, 2000; Frisch, 1997). This means that L2 speakers are more likely to employ a close acoustic-

phonological alternative when encountering the challenging English interdental fricative /θ/[1]. Conversely, a genuine high functional load (FL) contrast, as in (2), is less prone to being substituted since /n/ and /t/ are perceptually distinct phonemes that share few phonological features and are thus less likely to be confused.

In the context of L2 pronunciation teaching, it is crucial to apply the FL principle within a framework that considers "ecologically valid" confusion patterns. Otherwise, its empirical application holds little merit in L2 pedagogy (Brown, 1988). The concept of FL has been integrated into L2 pronunciation training by focusing on those segmental contrasts that are prone to confusion and essential for effective communication and thus intelligibility (Munro & Derwing, 2006; Suzukida & Saito, 2019). As such, a FL hierarchy for L2 English is limited to contrasts commonly conflated by L2 learners, such as /r - l/ for L1 Japanese speakers, /s - θ/ for L1 German speakers, and /b - v/ for L1 Spanish speakers. Comparison of the FLs for these contrasts reveals that /r - l/ and /b - v/, despite their high phonological similarity, have a relatively high functional load based on the count of minimal pairs (Suzukida & Saito, 2019), indicating their greater impact on intelligibility and comprehensibility. In contrast, /s - θ/ only correlates with perceived accentedness (Munro & Derwing, 2006). Intelligibility is significantly compromised when substituting /r/ with /l/, as indicated by numerous lexical distinctions reliant on the /r - l/ opposition, leading to potential difficulties in understanding the utterance (3).

[1] The substitution type can also depend on the L1 of the speaker. L1 Dutch speakers commonly substitute the English interdental fricatives with /t - d/, respectively, whereas L1 German speakers will likely use /s - z/ substitutions.

(3) Target: *The ba**r** was full of customers that night*

High FL substitution: *The ba**ll** was full of customers that night*

The present study will argue that the count of minimal pairs associated with a specific phonemic contrast holds limited relevance in understanding the overall utterance. Firstly, substitutions may not always result in a real word, and even if they do, the likelihood of the erroneous word belonging to the same word class and fitting within the same syntactic position and context is low (Levis & Cortes, 2008). Consequently, one lexical item will often be more probable in the listener's interpretation. Additionally, word recognition extends beyond the discrimination of a specific phonemic contrast, such as /r - l/, and relies on the overall representation of the speech stimulus to related words stored in the human lexicon. This broader perspective encompasses the realm of word recognition.

1.2 The Neighbourhood Activation Model (NAM) in the context of the FL principle

The Neighbourhood Activation Model (NAM) (Luce & Pisoni, 1998) provides a theoretical framework for understanding word recognition in the mental lexicon. According to NAM, when a word is encountered, its representation is activated along with related words, known as phonological neighbours, that share similar phonological and semantic features with the target word. An example illustrating this is as follows.

(4) Target: *The lizard was fascinating*

Substitution: *The *rizard was fascinating*

In this example, the erroneous word **rizard* has several lexical neighbours such as *wizard*, *lizard*, and *blizzard*, with the near-minimal pair *hazard* also being a potential option. While

hazard may have the lowest semantic fit in the context of the utterance, *lizard*, *wizard*, and *blizzard* provide good acoustic-phonetic matches. Ultimately, *lizard* and *wizard* would be the closest matches to **rizard* due to the similarity between the /r/ and /l/ phonemes in English, but *blizzard* also represents a possible alternative. The activation of neighbouring words depends on the extent of overlapping sound patterns, which can either facilitate or interfere with word recognition and retrieval.

While the concept of functional load (FL) primarily represents a bottom-up processing approach in L2, where the ease or difficulty of understanding utterances is influenced by the substitution of meaningful phonemic contrasts within words, NAM represents a top-down approach to word recognition. In NAM, words are recognized as wholes rather than being processed sequentially on the phonetic level (Vitevitch, 2002). This suggests that the notion of FL can only be examined in conjunction with word recognition models, which provide a more ecologically valid approach to understanding how utterances are processed. Specifically, the number of lexical neighbours that fit the utterance context and their degree of acoustic-phonetic similarity to other words predict the intelligibility (the degree to which the utterance is understood) and comprehensibility (the effort required for understanding) of an utterance (Weber & Cutler, 2004; Sewell, 2021). While previous research on NAM has defined word similarity based on subjective similarity judgment tasks where listeners rated the similarity of speech sounds within a syllable pattern on a scale from 1 to 10 (Fallon, Groves & Tehan, 1999; Luce et al., 2000), it was deemed necessary to establish an objective measure of phonological similarity for single segments to empirically test FL, particularly when examining substitutions based on specific phonemic contrasts.

1.3 Phonological similarity

The phonological similarity of individual segments refers to the degree of similarity of shared phonological features between speech sounds within a given language. The original measure of edit distance, i.e., the minimum of operations of substitutions, insertions, and deletions required to transform one phonetic string into another (Levenshtein, 1966), marked the least sophisticated approach to employing an algorithm. This algorithm compares the phonetic characters of two strings from left to right and computes the minimum number of edits required at each position to align the strings. By considering different edit operations, the algorithm determines the optimal alignment that minimizes the total number of edits. However, one limitation to such an approach is that it only focuses on the sheer count of operations needed to transform one string into another, without considering the specific phonetic-phonological properties of the speech sounds involved (5, 6).

(5)

l	i z a r d
w	i z a r d
1	0 0 0 0 0

(6)

	l i z a r d
b	l i z z a r d
1	0 0 0 0 0 0

Both of these substitutions would require one edit operation, (5) substitution of /l/ with /w/, and (6) insertion of /b/, respectively. Based on the sheer count of phonetic alterations, both *wizard* and *blizzard* would be equally phonetically close to *lizard*. Therefore, this computation does not accurately reflect the phonetic similarities or differences between individual speech sounds

within the word. In the context of the current study, the phonetic similarity of the substitution to the underlying target sound is important to quantify, because it will have a direct impact on how well and quickly words are recognized based on their overall acoustic-phonological structure.

Kondrak's computations of phonological similarity offer a more fine-grained and improved approach to quantifying phonological similarity (2000; 2003). In his phonetic string alignment algorithm ALINE, phonetic properties and representations of speech sounds, such as phonetic transcription and feature vectors, are considered and edit distance is applied specifically to these representations. Thus, ALINE accounts for the specific phonetic characteristics and patterns of the sounds being compared by statistical analysis of a large corpus of phonetic data (Kondrak, 2000). To assign weight to the features of individual speech sounds, a saliency measure is applied to capture the degree of correlation between a specific feature and the phonemic contrast analyzed. As such, features like place of articulation, manner of articulation, voicing, vowel height, and vowel backness are commonly considered in the weighted features for saliency. This means that features that are highly distinctive for a particular contrast are assigned higher weights, indicating greater relevance in differentiating those sounds, i.e., if the place of articulation is found to be highly correlated with a particular contrast, such as /p - b/, it would receive higher weight in the calculation of phonological similarity. On the other hand, if voicing between /p/ and /b/ is not relevant in distinguishing a contrast, it may be assigned a lower weight.

While this approach acknowledges the differential importance of phonetic features in distinguishing between sounds and provides a more nuanced understanding of phonological relationships within a language, Frisch (1997) provided phonological similarity computations that capture the full range of phonetic interactions and dependencies within a given language. In

his Pairwise Segment Alignment (PSA) algorithm, phonetic strings or sequences containing the sound contrast under investigation are aligned and edit operations of phonetic units are performed to establish correspondence between the aligned segments. The weighting scheme of phonological features assigns differing degrees of relevance to features based on their discriminatory power or significance in distinguishing between the segments compared. Depending on the language being studied, and the specific contrasts under investigation, weights assigned to features can vary. As such, Frisch, Pierrehumbert, and Broe (2004) suggest that instead of considering pairwise segmental similarities by phonological features, they can be grouped into natural classes, which are sets of segments that exhibit similar phonetic properties. Thus, the computation of phonological similarity involves comparing these natural classes rather than individual segments (Frisch et al., 2004). In concrete examples, this means that rather than comparing individual segments like /p/, /t/, and /k/ by count of phonological features, they can be grouped into a natural class of voiceless stops. The similarity between voiceless stops is then computed by comparing the shared phonological features of this natural class. This analysis allows for a more abstract and linguistically motivated analysis of similarity, going beyond surface-level segmental comparisons.

Overall, these findings suggest that the FL principle and its real-world application cannot provide a comprehensive understanding of phonemic distinctions' communicative significance without considering the context of word recognition models like NAM. In conjunction with each other, the FL principle and NAM can provide a clearer understanding of the complex interplay between phonemic contrasts and lexical processes for the understanding of utterances in both L1 and L2 contexts.

1.4 How do high and low FL errors affect word recognition?

Chapter 2 of this thesis investigates the influence of functional load (FL) in German in conjunction with the Neighbourhood Activation Model (NAM) as a word recognition model. For this study, 22 native German listeners participated in a perception task, where they listened to 138 utterances containing one or two word-level substitutions resulting in nonwords. The substitutions included high FL substitutions (/iː - a/, /eː - aː/, /oː - iː/; /ʁ - n/, /v - z/, /s - n/) and low FL substitutions (/yː - aː/, /øː - oː/, /ɪ - œ/; /t͡ʃ - p/, /p͡f - t/, /j - v/), as well as segmental contrasts commonly confused by L2 German speakers (/aː - a/, /iː - eː/, /uː - yː/, /øː - uː/; /z - t͡s/, /k - x,ç/[2]). An additional set of 138 distractor utterances was included. The listeners were required to fill in the gaps with their perceived underlying target words, assessing both intelligibility and comprehensibility. A total of 552 gaps were filled, with participants indicating their recognition of words and rating the overall degree of effort required to understand each of the 276 sentences. From the collected data, intelligibility scores were computed for each gap, with correct identifications receiving a score of '1' and incorrect identifications a score of '0'. Comprehensibility scores were assigned by each listener on a Likert scale ranging from 1 (extremely easy to understand) to 9 (extremely difficult to understand). Additionally, response times were recorded for each participant on each utterance. Intelligibility, comprehensibility, and response times were computed within each error category (high FL, low FL, confusable segments, and distractors).

[2] Note that /x/ and /ç/ are not phonemic in German. They are allophones of the same phoneme in complementary distribution, with /x/ occurring when preceded by a back vowel, and /ç/ occurring when preceded by a front vowel.

The results revealed that utterances containing high FL errors were less intelligible than those with low FL errors, followed by utterances with confusable segmental errors. This pattern corresponded with the comprehensibility scores, as listeners consistently rated utterances with high FL errors as more difficult to understand compared to those with low FL errors, followed by confusable segmental errors. Regarding response times, all three error categories differed, with high FL errors yielding longer response latencies, followed by low FL errors and confusable segmental errors. However, pairwise comparisons of single or double error categories demonstrated significant differences in three comparisons only: 2 confusable segmental errors vs. 1 confusable segmental error (2CS vs. 1CS), 2 high FL errors vs. 1 low FL error (2H vs. 1L), and 2 high FL errors vs. 1 confusable segmental error (2H vs, 1CS). This showed that significant differences did not stem from the respective error categories (H, L, CS), but rather from the number of errors (2 vs. 1). Analyzing the data in terms of single vs. multiple errors revealed a clear trend, where utterances containing two segmental errors were generally more difficult to understand than those with a single error, regardless of the error category. This indicated a cumulative effect of errors, as comprehensibility significantly decreased when multiple errors were present, with only one exception (1 high FL vs. 2 confusable segmental errors; 1H vs. 2CS).

Additionally, I computed the effects of syllabic position (i.e., if there were significant differences between errors occurring at the onset or coda of words) and syllable number (i.e., if there were significant differences between errors occurring in a mono- or disyllabic word), but no significant differences were found. Previous research had demonstrated mixed findings on the status of the syllabic position and syllable number. While it has been suggested that the initial portion of a word has special status and that it is psychologically important (Cole & Jakimik,

1980), other studies have found that rhyme words that share all but the initial phoneme facilitate lexical access (Marslen-Wilson & Zwitslerlood, 1989). Similarly, research found that monosyllabic words occur in denser neighbourhoods. That is, they tend to have more lexical neighbours than multisyllabic words, which can negatively affect perception accuracy and yield longer processing times (Vitevitch & Luce, 2015). Conversely, it has been revealed that bisyllabic words affect perception to the same extent as monosyllabic words and that they can also have a high number of neighbours (Cluff & Luce, 1990).

Moreover, phonological similarity scores were computed for the segmental contrasts involved in the substitutions for each error category, based on the FL principle and the Neighbourhood Activation Model (NAM) as the word recognition model (Dai, 2021). The results indicated that both high FL and low FL contrasts had average similarity scores of 76% and 78%, respectively, while confusable segmental errors shared 93% of features by natural class comparison (Frisch, 1997; Frisch et al., 2004). Consistent with the assumption that acoustically-phonologically similar words to the target word would lead to quicker activation and provide a higher degree of 'goodness of fit', confusable segmental contrasts yielded higher comprehensibility and intelligibility scores. However, the significant difference between high and low FL errors could not be explained solely by the averaged phonological similarity scores of the segmental contrasts involved in the substitutions for each error category.

While the direct responsibility of individual segmental contrast similarity for the loss of intelligibility and comprehensibility could not be firmly established, individual listener responses suggested that if multiple lexical neighbours were available and the acoustic-phonological structure obscured the target word, utterances became more difficult to understand and gaps were incorrectly identified. This strongly indicated the importance of high FL contrasts, particularly

those likely to alter the phonological structure of the word due to dissimilarity between the substitutional and target segments, for word recognition (intelligibility) and the perceived 'goodness of fit' (comprehensibility) assigned by listeners.

Chapter 2 aimed to determine the priority of teaching segmental contrasts in German based on their relevance to communication, as proposed by the FL hierarchy. The findings indicated that typically confused contrasts do not tend to carry a high FL, while genuinely high FL contrasts are phonologically dissimilar enough to avoid confusion. Therefore, the study could not definitively inform which segmental contrasts should be prioritized in German sound instruction. However, it did reveal that cumulative errors, regardless of their specific classification, negatively affected intelligibility and comprehensibility, suggesting that individual sounds still hold value for instruction. Employing a high-variability phonetic training paradigm in two different modalities, Study 2 (Chapter 3) tests empirically those idiosyncratic sound contrasts of German that have been identified in instructional textbooks (Lado, 1957; König & Gast, 2009; O'Brien & Fagan, 2016).

1.5 Training segmental contrasts

Central to L2 speech learning is the role of a speaker's L1. L2 learners may encounter difficulties in perceiving and producing sounds accurately because their perceptual systems have become attuned to the L1 sound inventory. For example, when an L1 English listener hears the German phonemic sound /yː/ within a word, it might be perceived as a phonetically different exemplar of English /uː/ (Kuhl, Conboy, Coffey-Corina, Padden, Pivera-Gaxiola, & Nelson, 2007). In Study 2 (Chapter 3), the selection of segmental contrasts for perceptual training was guided by three prominent theories: Contrastive Analysis (CA) (Lado, 1957), the Speech Learning Model(-r)

(SLM) (Flege, 1995; Flege & Bohn, 2021), and the Perceptual Assimilation Model-(L2) (Best & Tyler, 2007). These theories provide valuable insights into the interaction between L1 and L2 differences, phonetic categorization, perceptual assimilation, and their impact on L2 perception and production.

Contrastive Analysis (CA) suggests that L2 learners' difficulties can be predicted by comparing the phonetic and phonological inventories of their L1 and L2 (Lado, 1957). It focuses on identifying potential areas of difficulty based on the differences between the two languages, suggesting that if a specific sound contrast in the L2 is absent in the L1, perceiving and producing the contrast will be challenging. For instance, considering the German /uː - yː/ contrast, which is not distinguished in English, it is implied that L1 English speakers will find this contrast difficult to perceive and produce. Conversely, the German /p – b/ contrast will pose no difficulties to L1 English speakers, because it corresponds to an existing phonemic contrast in English. While comparing the phonological inventories of English and German may seem straightforward and helps identify differences, CA has its limitations in oversimplifying L2 acquisition. It assumes that L2 difficulties solely stem from differences between L1 and L2, disregarding the complexities of phonetic categories and their formation and development in L2 learners. However, it cannot accurately predict areas of difficulty of acquisition, but rather errors that occur due to L1 interference operating outside the learner's control (Lennon, 2010).

Evolving from the CA, the Speech Learning Model(-r) (SLM, SLM-r) developed by Flege (1995; Flege & Bohn, 2021) focuses on the formation and categorization of phonetic units. Unlike CA, the SLM presents explicit hypotheses regarding the perception and production of L2 sounds, initially aiming to explain age-related limitations in achieving native-like production of

L2 vowels and consonants. According to the SLM, the earlier an individual learns an L2, the greater the likelihood of perceiving and producing L2 sounds in a native-like manner.

In the early stages of L2 learning, the SLM suggests that listeners map L2 sounds onto existing L1 phonetic categories through assimilation, perceiving and categorizing L2 sounds based on the closest L1 category. Challenging the assumptions of CA, the SLM proposes that the degree of perceived dissimilarity between an L2 sound and the closest L1 sound influences the difficulty of discrimination in the initial stages, reflecting the learners' existing phonetic categories. Thus, L2 sounds that bear a strong similarity to corresponding L1 sounds are challenging to discern and establish as a new phonetic category because they are assimilated to the closest L1 category. Consequently, L2 sounds sharing acoustic characteristics with an existing L1 sound will be categorized and produced similarly to that L1 sound. The SLM's emphasis on perceived similarity aligns with the Perceptual Assimilation Model (PAM), which was almost simultaneously published by Best (1995) and also focuses on perceptual assimilation.

Originally aimed at naïve listeners, Best and Tyler (2007) expanded the assumptions and predictions of the original Perceptual Assimilation Model (PAM) to encompass L2 learning, adopting a direct-realist perspective. Central to the theory, and aligning with the SLM, are the concepts of similarity and dissimilarity between L1 and L2 sounds. The model posits that initially, L2 sounds are perceived and assimilated into the most articulatorily-similar L1 phonological category. Unlike the SLM, which relies on acoustic-phonetic cues, PAM-L2 posits that learners perceive L2 sounds based on their articulatory gesture, leading to the identification of different assimilation patterns of L2 sounds to L1 categories.

Taken together, the SLM and PAM(-L2) models have made significant contributions to L2 perception research, providing a comprehensive understanding of the dynamic process of L2

acquisition. These models consider the learner-specific influence of L1 phonology on L2 learning and the developmental trajectory of L2 learners' perception abilities. In specific perceptual training paradigms, the aim is to train perceptual discrimination to facilitate the establishment of new phonological categories or the modification of existing L1 category boundaries, enabling accurate differentiation and production of L2 sound contrasts. Typically, this involves L2 sounds that are phonemic in the target language but not in the L1, identified through CA. For L2 German, the selection of perceptual discrimination tasks is based on contrasting the phoneme inventory of German with that of the learners' L1. Due to this, extensive research has been dedicated to the German front-rounded vowels /yː/ and /øː/ and their contrastive counterparts /uː/ and /oː/ (Strange et al., 2009). English L1 speakers have shown difficulty in distinguishing these vowel sounds because this contrast is absent from the English phonological inventory and requires differences from /uː/ and /oː/ in lip rounding and tongue position. Similarly, the vowel contrast /aː - a/ can be problematic due to quality and duration differences. Long /aː/ is produced with longer duration and is more open, while the short /a/ is pronounced with a shorter duration and more central (König & Gast, 2009). The high acoustic similarity between /aː/ and /a/ can contribute to the difficulty in perceiving and producing the contrast, which may make it more difficult for L2 learners of German to have phonological awareness of subtle differences that are not too phonetically distinct, as has been shown in L1 American English practically naïve listeners of German (Strange, Levy, & Law, 2009). Moreover, the German contrast /iː - eː/ can be challenging to L2 learners of German (Strange, Levy, & Law, 2009) with L1 English. While English has /iː/ in its vowel inventory, it lacks /eː/ (König & Gast, 2009). This is because English /iː/ has a higher F1 compared to German /iː/ and overlaps with German /eː/ in acoustic space. Because /iː/ and /eː/ are only contrastive to L1

English speakers within diphthongal /eɪ/, e.g., *beat* and *bait*, the category boundaries of the German /iː - eː/ contrast might not be clear. To my knowledge, there are no perceptual studies on the German consonantal contrasts /k - x, ç/ and /t͡s - z/. A comparison of phoneme inventories of German and English shows that these phonemic oppositions have confusability potential in English and have been tested in Study 1 as part of the confusable segmental contrasts category, even though they do not carry a high FL. Both the palatal and velar fricatives /ç, x/ are absent from the English phoneme inventory as is the affricate /t͡s/ in syllable onset position (König & Gast, 2009).

1.6 High-Variability Phonetic Training (HVPT)

Phonetic training holds significant importance in L2 acquisition and speech recognition systems. Among the various training approaches, high-variability phonetic training (HVPT) has emerged as a promising technique for improving learners' phonetic skills. As such, the HVPT technique is based on the principle that exposure to a wide range of phonetic variation enhances learners' ability to accurately perceive and produce speech sounds (Logan, Lively, and Pisoni, 1991; Lively, Logan, & Pisoni, 1993). This approach utilizes diverse phonetic exemplars that cover a broad range of variability within a specific sound category, involving multiple talkers with unique features (voice quality, accent) producing stimuli in different phonetic contexts. Through training with such varied exemplars, learners are thought to develop robust phonetic representations and become more adaptable in understanding and producing speech sounds in real-world settings (Thomson, 2018).

A seminal study by Logan et al. (1991) investigated the impact of phonetic environment and talker variability on training Japanese listeners to identify English /r/ and /l/ sounds. The

study demonstrated that exposure to a high degree of variability in phonetic context and multiple talkers over 15 sessions led to significant improvement in listeners' ability to discriminate between the /r/ and /l/ contrast, and this improvement generalized to new stimuli. Although listeners in a low-variability condition also showed improvement, it was less pronounced in distinguishing the /r - l/ contrast and had limited generalization to new stimuli. These promising findings established the HVPT technique as standard procedure in L2 perceptual training (Brekelmans et al., 2022).

Following the initial study on the /r - l/ contrast, subsequent research has further confirmed the effectiveness of HVPT for vowel contrasts (Nishi & Kewley-Port, 2007) as well as lexical tones (Sadakata & McQueen, 2014) and in different learner populations, including children (Heeren & Schouten, 2010; Hwang & Lee, 2015) and highly proficient advanced L2 learners (Bradlow, Pisoni, Akahane-Yamada & Tohkura, 1999; Cebrian & Carlet, 2014). HVPT has been applied not only to English but also to other languages, such as teaching Japanese moras to L1 English speakers (Hirata, 2004) and the French vowel contrast /e - ɛ/ to L1 Spanish speakers (Kartushina & Martin, 2019). To date, approximately 40 HVPT studies have highlighted the benefits of the technique, with a particular emphasis on the role of talker variability. The rationale behind incorporating multiple speakers with their individual speech characteristics is to provide sufficient variability that helps listeners discern which acoustic cues are relevant for discrimination (Brekelmans et al., 2022). While most HVPT studies make use of highly variable auditory stimuli, the effectiveness of training with these can be further enhanced by incorporating visual-gestural information from speakers producing the sounds (Hazan, Sennema, Iba, & Faulkner, 2005).

1.7 Audiovisual HVPT and the importance of the gesture

HVPT has shown particular success in leading to improvement in the perception and production of English contrasts for Japanese listeners when auditory and visual cues are provided simultaneously, compared to audio-only training. The visually distinct /b - v/ contrast in English was significantly better discriminated by an audiovisual training group compared to an audio-only group, although both showed perceptual gains (Hazan et al., 2005). However, discrimination of the /r - l/ contrast improved for both groups, with no advantage of the audiovisual condition, likely due to the lack of visual cues (lip movement; tongue position) showing the articulatory distinctiveness between /r/ and /l/.

The theoretical framework that guides the HVPT technique is the Perceptual Assimilation Model-L2 (PAM-L2) (Best, 1995; Best & Tyler, 2007). PAM-L2 provides insights into how learners perceive and categorize new sounds and contrasts based on their existing L1 phonological categories. A key premise of this direct-realist model[3] is the significance of articulatory gestures and their influence on speech perception, which is informed by the influential model of motor theory (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Motor theory proposes that our perception of speech sounds is shaped by our own motor representations, involving mental simulation of the articulatory gestures associated with producing those sounds. In other words, when we hear speech, we internally imagine and mimic the corresponding gestures involved in producing those sounds. Within the framework of PAM-L2, the importance of articulatory gestures lies in how they impact listeners' categorization and

[3] A direct-realist model is a theoretical approach of speech perception that posits that listeners directly extract meaningful linguistic information from the acoustic signal without the need for intermediate cognitive processes or complex inferences.

assimilation of novel phonemic contrasts (Best, 1995; Best & Tyler, 2007). For example, in the case of Japanese, the distinct English gestures for /r/ and /l/ closely align with a single gestural category in Japanese, which hampers the discrimination of the /r - l/ contrast.

While PAM-L2 is largely compatible with the goals and principles of HVPT, as it provides a systematic account of how listeners assimilate and categorize novel phonemic contrasts and modify perceptual boundaries through training, the HVPT technique specifically capitalizes on the principle of talker variability, which cannot be fully captured by PAM-L2 alone. If variation in speech patterns across different speakers proves beneficial for perceptual learning compared to minimal variability (e.g., single talker), an exemplar theoretical approach can effectively account for this. Exemplar Theory (ET) focuses on capturing fine-grained details of speech perception, including the variability present in different talkers' speech signals (Goldinger, 1998; Hintzman, 1986; Johnson, 1997). In ET, each instance of a stimulus becomes a detailed memory trace that includes acoustic, phonetic, and contextual information. For speech perception, this means that specific instances or exemplars of sounds are stored and compared in memory, and perceptual judgments are made based on the similarity between the input and the exemplars. ET recognizes the importance of preserving talker-specific details in speech perception, as listeners develop exemplar-based categories of speech sounds that capture the full range of variability present in the input, such as individual speaking styles or voice quality. The study in Chapter 3 aims to integrate PAM-L2 as the overarching framework for HVPT and adopts an exemplar-based approach to account for talker variability.

1.8 The effects of audio-only vs. audiovisual HVPT

In Chapter 3, the study investigates the effects of audio-only versus audiovisual high-variability phonetic training (HVPT) on perceptual discrimination in two groups of beginner learners of German. The training technique focuses on the German sound contrasts from the confusable segments (CS) group: /aː - a/, /iː - eː/, /uː - yː/, /øː - uː/; /z - t͡s/, /k - x,ç/ (König & Gast, 2009; Strange, Levy, & Law, 2009; O'Brien & Fagan, 2016). The speech stimuli used in the training were monosyllabic CVC nonwords produced by eight L1 German speakers aged 25 to 83. While previous studies have shown the benefits of audiovisual HVPT over audio-only training due to the visibility of articulatory gestures for certain contrasts (e.g., /b - v/ for L1 Japanese listeners of L2 English) (Hazan et al., 2005), this study aimed to explore the advantages of seeing speakers' faces in conjunction with their voices, hypothesizing that visual input enhances and strengthens the training effects. Notably, studies reporting the effectiveness of HVPT have not considered the theoretical framework of exemplar theory (ET).

The training utilized recorded data of L1 German speakers producing the critical German vowel contrasts. Twelve beginner learners of L2/L3 German from a large Western Canadian university participated in the study. A pre-test was conducted where all participants listened to the critical German sound contrasts produced by a single speaker to assess their discrimination ability prior to training. The same procedure was repeated in a post-test. During the semester, six participants underwent audio-only training, while the remaining six received audiovisual training. Both groups were trained for six weeks, with three 20-minute sessions per week, using the phonemic German contrasts produced by eight different speakers. In each group, participants performed ABX discrimination tasks with corrective feedback provided for incorrect answers. The results revealed that participants achieved higher discrimination accuracy scores for each

German contrast in both the pre-test and post-test conditions with a single talker and audio-only modality. Moreover, the audio-only training group outperformed the audiovisual training group significantly on five out of the seven contrasts tested.

In summary, the study not only failed to find a beneficial effect of audiovisual HVPT but rather found that seeing multiple speakers had a detrimental effect on discrimination accuracy. These findings indicate that caution should be exercised when making use of HVPT. Firstly, the study included true beginners of L2 German in the initial stages of L2 learning when phonetic categories are still being established. Secondly, an audiovisual modality introduces additional cognitive load in that participants had to process speech stimuli produced by multiple talkers along with redundant visual information. Although HVPT has been recognized as a successful technique since the landmark study (Logan et al., 1991), few other studies have directly compared the advantages of HVPT over low-variability phonetic training (LVPT) (Dong et al., 2019; Giannakopoulou et al, 2017) in different proficiency levels, such as beginner versus advanced learners (Sadakata & McQueen, 2014; Wong, 2012).

1.9 Conclusion – Bridging Study 1 and Study 2

The functional load (FL) principle is a linguistic concept that examines the importance and communicative significance of phonemic contrasts within a given language. It has been proposed as a component of a comprehensive descriptive analysis of a language's sound system, considering the degree to which phonemes contribute to successful communication. According to this principle, certain segmental contrasts play a more crucial role in distinguishing words and perform more work in keeping utterances apart (King, 1967). In recent years, the FL principle has gained attention in L2 English pronunciation pedagogy (Sewell, 2017; 2021) as a guide for

selecting phonemic L2 sounds that might pose difficulties to distinguish among L2 learners (Brown, 1988). Studies found that high FL contrasts, which have high communicative value, are closely related to intelligibility and comprehensibility in L2 speech and should be prioritized in teaching, while low FL contrasts are mainly associated with the presence of a perceived accent (Alnafisah, Goodale, Rehman, Levis, & Kochem, 2022; Derwing & Munro, 2006; Sewell, 2017; Suzukida & Saito, 2019).

Chapter 2 of this thesis presents a study investigating the applicability of the FL hierarchy (Oh et al., 2015) that has been proposed for the German language. It investigates the functionality of phonemic oppositions in conveying meaning by examining the performance of L1 listeners of German in a word recognition task. The aim is to determine the impact of high and low functional load contrasts on the intelligibility and comprehensibility of utterances. The results of the study demonstrate that high functional load errors are more detrimental to intelligibility and comprehensibility than are low functional load errors, or confusable segmental errors. However, cumulative errors were generally found to be more grave than single errors, irrespective of classification as high, low, or confusable. Notably, high functional load contrasts of German are not among those sounds that (L2) speakers would commonly confuse, which highlights that the FL principle has theoretical merit, but lacks real-world applicability within the context of German.

Although none of the sound pairs that are typically confused rank high in the FL hierarchy of German, and Study 1 did not provide specific insights into prioritizing speech sounds crucial for comprehensibility, the results do indicate that cumulative segmental errors can still lead to a loss of intelligibility. This is why Study 2 employs a set of the typically confused sounds of German reported by existing literature, i.e., /aː - a/, /iː - eː/, /uː - yː/, /øː - uː/; /z - t͡s/, /k

- x,ç/ (König & Gast, 2009; O'Brien & Fagan, 2016), which can be trained to enhance learners' perceptual abilities.

In Chapter 3, I present the results of Study 2, which explores the HVPT technique in two modalities: audio vs. audiovisual training of German phonemic contrasts in beginner L2 learners over the course of a six-week training period. Specifically, the goal of this technique is to enhance and foster discrimination of phonemic contrasts that are vital for differentiation and that pose difficulties to L2 learners. The strengths of this technique lie in its employment of multiple talkers and variable phonetic environments, thereby exposing listeners to a large variety of stimuli. Studies have shown that the success of HVPT training, which improves perceptual discrimination and can transfer to production (Thomson, 2018), can be attributed to this very variability because the input learners receive is rich and robust and thus closely resembles an ecologically valid and naturalistic method for encountering an L2.

The results show that discrimination accuracy of L2 listeners for German sound contrasts was found to be reduced in both the audio-only and audiovisual training conditions. However, the decrease in accuracy was significantly greater in the audiovisual modality, which exhibited higher variability. In contrast, when a single talker was employed in an audio-only modality for the pre- and post-test conditions, discrimination accuracy reached a ceiling level. These findings suggest that HVPT introduces additional cognitive load to the listeners who are in the beginning stages of acquiring phonetic knowledge of a new L2 sound system, especially when dealing with two modalities (audiovisual) simultaneously. Additionally, the evidence highlights the need for reconsideration of HVPT and suggests that further investigation is required to compare the benefits of HVPT with LVPT based on learner proficiency levels.

In Chapter 4 of this thesis, I discuss the findings of both studies and suggest how they may guide future work on both the functional load principle and its applicability in ecologically valid settings of L2 pronunciation pedagogy as well as the HVPT technique and its theoretical implications that highly variable talker input will foster L2 perceptual learning to a higher degree than low-variability phonetic training.

Chapter 2: Examining the functional load principle in German

Abstract

The notion of functional load refers to the significance a linguistic element has in conveying meaning and facilitating communication within a language. Functional load has been investigated in various linguistic contexts (Brown, 1988; Catford, 1987; Hockett, 1966). For example, it has been found that in English certain consonants and vowels (speech sounds) carrying a high functional load are more crucial than others for understanding spoken utterances (King, 1967). The theoretical notion of FL and its real-world applicability have primarily been tested for English (Munro & Derwing, 2006; Suzukida & Saito, 2019). The current study empirically tests the functional load principle in German to determine the communicative value of individual sound contrasts within spoken utterances in L1 German listener perception. Twenty-two L1 German listeners judged 138 sentences of German containing either one or two high functional load, low functional load, or segmental errors of phonologically similar segments in utterances providing minimal semantic context. The results showed that high functional load errors affected comprehensibility and intelligibility to a higher extent than low functional load errors or confusable segmental errors. This could be attributed to both the phonological dissimilarity of high functional load substitutions, as well as the number of lexical neighbours available. When there were two erroneous segments, differences between high and low functional load contrasts were diminished. The results suggest that listeners perceive words based on the overall goodness of fit from acoustic-phonetic input and access lexical neighbours based on similarity.

2.1 Introduction

In recent years, understanding how phonemic contrasts as the smallest unit of speech sounds contribute to the distinction of utterances has served as a template for prioritizing certain sounds over others for phonetic pronunciation training in English. Consider the sound pairs in examples (7) and (8).

(7) /d/ <bore**d**> ['bord] and /z/ <boar**s**> ['borz]

(8) /h/ <**h**ail> ['heɪl] and /m/ <**m**ail> ['meɪl]

While /d/ and /z/ in (7) distinguish approximately 2900 minimal pairs, a sound pair like /h/ and /m/ in (8) only distinguishes around 280 minimal pairs.

It is hypothesized that the higher the functional load (i.e., the number of minimal pairs that hinge on a certain phonemic distinction), the more important the contrast to communication due to the probability of confusability of lexical items. A functional load hierarchy of sounds has been proposed for English (King, 1967), German and other languages (Oh, Coupé, Marsico, & Pellegrino, 2015), but it has to the author's knowledge only ever been empirically tested for English (Munro & Derwing, 2006; Suzukida & Saito, 2019). The current study extends the work on English and empirically tests the notion of *functional load* (FL) among German L1 listeners.

2.2 Literature Review

2.2.1 Intelligibility and comprehensibility

The notion of *intelligibility* is related to the recognition of words and utterances (Smith & Nelson, 1985). As such, it refers to a listener's actual understanding of the words of an utterance. Previous research has demonstrated that pronunciation deviations, like segmental substitutions, deletions, or epentheses, can be the source of the loss of intelligibility. Closely related to intelligibility is the notion of *comprehensibility*, which refers to the degree of effort by a listener to understand the utterance (Levis, 2020). While intelligibility is an objective measure (Is the utterance understood?), and comprehensibility is more of a subjective measure (How much effort was required to understand the utterance?), the concepts are closely related and quantifiable. Intelligibility is commonly measured by transcription of what has been heard (Kang, Thomson, & Moran, 2018), while comprehensibility is measured in scalar ratings of speech assigned by a listener to indicate the degree of effort required to understand the utterance (Isaacs & Thomson, 2013). In practice, a listener's transcription is compared to that of the actual utterance. Intelligibility is expressed as a percentage of correctly transcribed words (Gooskens, 2013). Conversely, for comprehensibility, the listener assigns a score, often on a 9-point scale, to indicate the subjective degree of effort required to understand a given utterance (Derwing & Munro, 1997).

Previous findings suggest that not all pronunciation errors are equal (Munro & Derwing, 2006; Lin, 2019). In particular, segmental errors (substitutions, deletions, epentheses) and suprasegmental errors (lexical stress, intonation, and rhythm) can affect intelligibility and comprehensibility to differing degrees. For the current study, the focus will be on segmental

substitutions only, all of which cannot be deemed equal, because some individual sounds play a larger role than others in understanding. This taps into the notion of *functional load*.

2.2.2 Functional load

The main premise of the notion of functional load (hereafter FL) is that some segments are more vital to communication, thus play a larger role in intelligibility and comprehensibility, than others. The notion involving a ranking of linguistic features of speech sounds in their priority to communication originates in the 1930s and was first quantified for English (King, 1967). As such, FL primarily concerns itself with pronunciation of phonemic oppositions and the work they do to convey information as quantified by a minimal pair count. Phonemic contrasts are assigned into either the high or low FL category based on the number of word pairs they form. For example, /d/ and /z/ in English create 2941 minimal pairs in words like in *lady* [ˈleɪdi] – *lazy* [ˈleɪzi]; *pride* [ˈpraɪd] – *prize* [ˈpraɪz] and would be categorized as high FL segments. Conversely, the contrast /z/ and /b/, which forms 87 minimal pairs, like in *bone* [ˈboʊn] – *zone* [ˈzoʊn], would be classified as a low FL contrast. However, the previous examples neglect the complexity of minimal pairs.

Catford (1987) and Brown (1991) highlight the importance of individual phoneme frequency as well as the structural distribution of phonemes (some can only occur in certain phonotactic positions). Moreover, not all minimal pairs can occur in the same syntactic positions, because they can be members of different word classes, as demonstrated by the above example *lady* (noun) – *lazy* (adjective) (Levis & Cortes, 2008). Additionally, inflectional morphology (e.g., plural or past tense forms) increases the number of minimal pairs. This is also the case for the high FL contrast /d – z/ in English, e.g., *bed – beds*; *bills – billed.* Despite these complexities, the notion of FL has been tested empirically for English in studies involving L1 speaker

comprehensibility ratings of Mandarin and Japanese-accented speech productions involving high and low FL substitutional errors (Derwing & Munro, 2006; Suzukida & Saito, 2019). The results showed that high—but not low—FL errors negatively impacted comprehensibility. This lends support to the idea that some phonemes perform more work than others in their functionality for communication.

Some high FL contrasts involve phonemic pairs that are prone to confusion by L2 speakers, e.g., /r/ and /l/ or /b/ and /v/ for L1 speakers of Japanese. Raw minimal pair counts demonstrate that these contrasts distinguish quite a few words. Those sound contrasts with the highest functional load, however, are typically those that are phonologically very distinct thereby preventing confusability (Oh et al., 2015)[4].

2.2.3 Phonological similarity

The example of the high FL contrast like /m - t/ and the low FL contrast /ɒ - ɔ/ for British English can be used to demonstrate that high FL contrasts are typically phonologically distinct, whereas low functional load contrasts often bear a substantial overlap of phonological features. The British English contrast /ɒ - ɔ/, e.g., *hock* [ˈhɔk] and *hawk* [ˈhɒk] only differs in the one phonological feature of height and distinguishes very few word pairs. In fact, in some dialects of English, the contrast has low functionality to communication, because it is no longer contrastive (see *cot – caught* merger in Canadian English). Conversely, a contrast like /m - t/ distinguishes many word pairs and is phonologically quite distinct. The two phonemes also differ in two phonological features and one perceptual property: voicing, and nasality, as well as sonority.

[4] For English, Oh et al. (2015) calculated by measures of minimal pair count and entropy that the contrasts of English that carry a high functional load include /n - t/; /m - t/; /θ,ð - m/ (APPENDIX A).

While the raw count of phonological features for the above examples is indicative of such phonological differences, this measure is insufficient to quantify phonological similarity. Conrad (1964) first determined phonological similarity by conducting a series of experiments on the effects of acoustic confusion on memory, i.e., a list of similar-sounding sounds is recalled less accurately than lists of dissimilar sounds. In the case of the recent example, a listener will perceive /m/ as more dissimilar from /t/ than from /n/, as previous studies have found in working memory serial recall tasks, which also found that rhyming words were recalled better than nonrhyming words (Fallon, Groves, & Tehan, 1999; Chow, Macnamara, & Conway, 2016). Alternatively, studies may use listeners' perceived similarity ratings of utterances (Luce, Goldinger, Auer, & Vitevitch, 2000). Even though numerous studies ascribe explanatory power to phonological similarity, they utilize it as an intuitive measure rather than a computational one. This is because there has long been no agreed-upon objective measure of phonological similarity, and 'perceived similarity' depends on the individual listener and their L1. Kondrak (2000; 2003) has quantified the degree of phonological similarity by algorithmic alignment of phonetic strings at the example of words affected by diachronic sound change. As such, he not only counted the phonological features of phones, but weighting in their salience assigning them a numerical value between '0' and '1'. Frisch (1997) also proposed a metric for measuring phonological similarity that uses features to represent sounds and that calculates similarity scores of phonetic strings on feature matches and mismatches. These features are derived from phonological patterns and constraints within a given language, their frequency and informativeness, rather than using a universal set of features. Based on these values, similarity matrices for sound contrasts within specific languages can be digitally computed (Dai, 2021).

The issue of phonological similarity between the segments involved in a phonemic contrast is crucial in the context of the present study for two reasons. Firstly, high FL substitution errors often involve dissimilar phonemic oppositions because high FL segments most effectively prevent confusability and distinguish many words. Secondly, a high FL substitution can alter the acoustic-phonological structure of a word in its entirety and thereby decrease intelligibility and comprehensibility to a higher extent than a low FL substitution. The latter suggests that FL should be examined in conjunction with word recognition processes involving lexical retrieval.

2.2.4 The Neighbourhood Activation Model (NAM)

The FL principle deems individual segmental substitutions to be the source of reduced intelligibility and treats them separately from word recognition processes. Since the FL principle assumes that it is the number of minimal pairs distinguished by a particular phonemic contrast contributing to the intelligibility of words, this suggests that the listener must guess the position in which the erroneous segment occurred within a word. This then means that the listener would have to take a unidirectional phoneme-by-phoneme approach (bottom-up) to detect the error to ultimately recognize the word. In contrast to this is the assumption that listeners perceive words in their entirety and that a heard stimulus activates competitors, acoustic-phonologically similar words from memory that resemble the structure of the input (Vitevitch, Luce, Pisoni, & Auer, 1999).

One such activation-competition model is *The Neighborhood Activation Model* (NAM) (Luce & Pisoni, 1998). This model suggests that listeners perceive words, or utterances, in their entirety within the first few hundred milliseconds of the speech signal by comparing the

acoustic-phonetic input to other possible alternatives from memory (top-down) (Luce, Pisoni, & Goldinger, 1991). Those possible lexical alternatives are called 'phonological neighbours' and often form minimal pairs or near-minimal pairs with the heard stimulus (Anderson, 2007), which directly corresponds to the notion of FL. According to NAM, recognition accuracy and latency thus depend on three factors: 1. the number of lexical alternatives in a listener's lexicon (density of the neighbourhood); 2. the context (i.e., word class and syntactic position); and 3. the phonological similarity. *Figure 1* illustrates the stimulus's **clab* phonological neighbourhood when heard in an utterance like the following.

*The *clab was not bad*

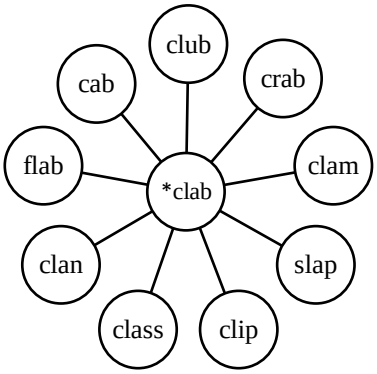


*Figure 1 Phonological neighbours of *clab*

This example shows the possible neighbours in the form of minimal or near-minimal pairs that fit the context of the utterance. Those with the fewest differences from the acoustic-phonetic structure of **clab* will be activated from memory more quickly (Vitevitch, 2015). Even if the stimulus deviates from the target word by a maximum of one segmental substitution, the phonological similarity of this substitution to the underlying word is crucial for word recognition

accuracy and speed: according to similarity computations, *crab* is acoustic-phonetically more similar to **clab* than *class* would be. This is because /l/ and /r/ share more phonological features in number and salience and have a higher confusion level according to Frisch's similarity matrix than do /b/ and /s/.

While NAM assumes that a word is recognized based on the overall degree of 'goodness of fit' to competing lexical items from memory, it acknowledges that factors like syllabic position (onset vs. coda) and number of syllables in the word (mono- vs. bisyllabic) play a role in recognition accuracy and response latency. While some studies found that segmental substitutions occurring early (i.e., at the onset of words or sentence-initially) inhibit lexical access early on in an utterance (Bent, Bradlow, & Smith, 2007), others argued that the 'goodness of fit' depends on the overall quality of the phonological match, which suggests that rhyme primes (both word and non-word) should facilitate recognition (Marslen-Wilson & Zwitserlood, 1989). In particular, monosyllabic rhyme primes sharing final segments should aid recognition (Luce, Goldinger, Auer, & Vitevitch, 2000). Additionally, mono- and polysyllabic words differ in the extent to which listeners may be able to recognize them accurately. Cole and Jakimik (1980) found that listeners detected errors in the second syllable of a word more quickly than in the first syllable because they had access to the correct information of the first syllable. This, in turn, would suggest that multisyllabic words have more redundant information as opposed to monosyllabic words, and they may therefore be recognized more quickly and with higher accuracy (Vitevitch, 2002).

As the current study empirically tests the FL principle as a part of word recognition, the notion of intelligibility corresponds to recognition accuracy (Is the utterance understood?) and

the notion of comprehensibility corresponds to the listeners' assessment of the degree of 'goodness of fit' (How much effort is required to understand the utterance?).

2.2.5 The current study

Previous research in the field of English as a second language (ESL) acquisition has tested empirically the functional load hierarchy and provided evidence that high FL segmental contrast substitutions have a negative impact on listener comprehensibility as opposed to low FL substitutions (Munro & Derwing, 2006; Suzukida & Saito, 2019). This negative effect was even more pronounced when high FL errors occurred cumulatively (i.e., when there was more than one error in an utterance), while multiple low FL errors did not negatively affect comprehensibility (Munro & Derwing, 2006). The existing studies relied on minimal pair count analyses by Catford (1987) and Brown (1991). While empirical testing of FL for other languages is scarce, a hierarchy of those phonemic contrasts that have the highest impact on communication has been proposed for several languages, including German (Oh et al., 2015). This approach combined the measures of raw minimal pair count proposed by Catford (1987), corresponding to the organization of the mental lexicon, and entropy proposed by Hockett (1966), corresponding to the frequency of token occurrence and online processing in real-life situations of communication. The goal of the current study is to test empirically the functionality of German high and low FL phonemic contrasts from the WebCELEX database (Max Planck Institute for Psycholinguistics, 2021) among L1 listeners.

An earlier version of the WebCELEX corpus served as a basis for the FL computations of German sound contrasts in the Oh et al. (2015) study, upon which the current study relies. It becomes evident that some oppositions form minimal pairs involving phonemes that are unlikely to be confused in real life, e.g., /n - ʁ/, unlike in English, where some confusable contrasts like /r/

– /l/ are part of the high FL group (Sewell, 2021). This is why in the current study an additional group of German segments exhibiting high potential for confusability (Roccamo, 2015; O'Brien & Fagan, 2016) was added for testing, irrespective of the classification as high or low FL.

In total, the following three types of errors were tested in their impact on intelligibility and comprehensibility occurring as single or two errors as part of an utterance.

(a) high FL contrasts (HFL),

(b) low FL contrasts (LFL),

(c) commonly confused segments (CS)

Three vowel and three consonantal contrasts that rank very high in their computational lists formed the high functional load category (HFL) in the current study. The second category consisted of three vowel and three consonantal contrasts that ranked lower in their proposed hierarchy (LFL). The computed functional load score 'FL' is an entropy-based complex measure taking individual phoneme frequency, shared and distinct features, minimal pair count of lemma lexemes, and merging probability into account. Oh et al. (2015) define FL_E at the level of phonemic contrasts, as a ratio ranging from 0–100%. It is also possible to focus on the level of phonemes themselves, by summing FL_E over all the contrasts in which a phoneme is involved (Oh et al., 2015)[5].

[5] The formula used confusability potential to compute FL at the level of phonemic contrasts by measures of entropy ($_E$) is the following:

$$FL_E(\varphi) = \frac{1}{2} \sum_{\psi} FL_E(\varphi, \psi)$$

More specifically, the 20,000 most frequent word-forms and lemmas were analyzed. Then, phonological entries for each language were syllabified and syllabic boundaries were considered for the computation of FL (Oh et al., 2015). Four values were determined to determine the functional load of a phonemic contrast: 1) the number of cases where confusion arises when the two segments are merged; 2) the true number of minimal pairs in the WebCELEX; 3) the number of shared phonological features; 4) the number of distinct phonological features. The higher the FL score, the higher the functional load of the contrast (APPENDIX B). The third category of commonly confused segments (CS) was added to the testing, because these phonemic contrasts did not emerge in the lists provided by Oh et al., but they constitute phonemic oppositions that are likely to be confused in ecologically valid communicative situations, i.e., they can occur as substitutions in particular L2 speaker productions whose L1 lacks the phonemic contrast. The confusability score percentage indicates the degree of phonological similarity between the two sounds that form the phonemic contrast.[6] Notably, none of the HFL contrasts bear high similarity scores, the highest resemblance occurring between /eː/ and /aː/ (86%), whereas the LFL contrasts host one highly similar contrast, /oː/ and /øː/ (93%), whereas the CS group exclusively consists of phonemic pairs whose similarity scores are above 89%. *Table 1* shows that phonemic contrasts that perform the work of keeping lexical items apart do this contrastive work best when the segments involved bear little phonological similarity.

[6] While the phonemic contrasts listed in the CS group were not among Oh et al. (2015)'s entropy-based computations, it had been assumed that their FL load must be low. In the raw data calculations obtained from Oh et al. after the decisions for the current study were made, however, the FL computations indicate a relatively high functional load for the /iː - eː/ contrast. Nonetheless, this contrast was still considered highly confusable and was therefore assigned to the CS group.

Table 1 Functional load contrasts with minimal pair count and confusability score

HFL contrasts (MP #)	Confusability score	LFL contrasts (MP#)	Confusability score	CS contrasts (MP #)	Confusability score
ʁ - n	54%	t͡ʃ - p	79%	z - t͡s	93%
(760)		(3)		(30)	
v - z	79%	p͡f – t	75%	k - x,ç	93%
(72)		(8)		(21)	
s – n	79%	j - v	57%	iː - eː	96%
(390)		(15)		(56)	
iː - a	79%	yː - aː	75%	a - aː	96%
(63)		(34)		(35)	
eː - aː	86%	oː - øː	93%	uː - yː	93%
(28)		(23)		(21)	
oː - iː	82%	ɪ - œ	89%	uː - øː	89%
(53)		(4)		(3)	

Furthermore, the current study seeks to advance our knowledge of FL for German in conjunction with word recognition and the particular makeup of words in which segmental substitutions occur, since up to now they have only been examined in isolation. Upon creating the stimuli containing the phonemic contrasts from each respective error type, it became evident that some substitutions alter the acoustic-phonological structure of a word more than others thereby obscuring the underlying word. Additionally, the listener does not know beforehand where exactly within a word a substitution occurs, lending support to the idea that words are processed in their entirety and compared against a set of lexical items from listener memory. Previous research suggests that the acoustic-phonological similarity between a stimulus and possible alternatives is decisive in word recognition accuracy and response latency (Vitevitch & Luce,

2016). These findings warrant the inclusion of word recognition processes into the empirical testing of the German functional load hierarchy.

The issue of phonological similarity is a crucial one, as substitution of an underlying segment can alter the acoustic-phonological structure of a stimulus, particularly if the substitution is dissimilar to the target one, thereby triggering activation of lexical items that form neighbours with the erroneously produced word. This suggests that words are recognized in their entirety through comparison of their acoustic-phonological structure to other lexical items (neighbours) and not in a phoneme-by-phoneme approach until detection of an erroneous segment, this warrants a closer look at the structural make-up of those words in which a segmental substitution occurs. As such, the impact of syllabic position (onset vs. coda) and syllable count (mono- vs. disyllabic) on recognition accuracy (intelligibility), the subjective listener degree of goodness of fit (comprehensibility) and speed of recognition (response time) is tested.

The research aims of the current study are to test phonemic contrasts of German in their functionality to communication by employing the FL hierarchy proposed by Oh et al. (2015). In particular, the goal is to determine whether high FL segments have a greater impact on communication (intelligibility and comprehensibility) than low FL segments and whether cumulative errors have a more detrimental impact than single errors, as has been found for English (Munro & Derwing, 2006). Drawing on the *Neighbourhood Activation Model* (NAM) (Luce & Pisoni, 1998), listeners' response accuracy and latencies are measured against the background of phonological similarity and the stimulus structure in which substitutions occur.

2.3 Methods

2.3.1 Participants

Twenty-three L1 listeners of German (11 females, 12 males) between the ages of 22 and 31 (M = 26.6 years) were tested. Seventeen participants reported speaking Standard German, one spoke East Franconian, one Hessian, one Mecklenburgian, one a Ruhr-area variety, and one Upper Saxon, but all were mainly exposed to Standard German in their daily lives. One participant had to be excluded from data collection, as they experienced technical issues, reducing the number of participants to 22 (11 females, 11 males) with a mean age of 26.3 years.

2.3.2 Stimuli

Utterances consisted of one sentence each and were recorded at 44.1 kHz using an integrated Apple MacBook M1 microphone. The stimuli consisted of 36 segmental pairs (consonant – consonant; vowel – vowel oppositions) grouped into the three different error types: high FL (H) (/iː - a/, /eː - aː/, /oː - iː/; /ʁ - n/, /v - z/, /s - n/), low FL (L) (/yː - aː/, /øː - oː/, /ɪ - œ/; /t͡ʃ - p/, /p͡f - t/, /j - v/), and confusable segmental contrasts (CS) (/aː - a/, /iː - eː/, /uː - yː/, /øː - uː/; /z - t͡s/, /k - x,ç/[7]). A total of 12 high functional load segmental pairs (six vowel pairs, six consonant pairs), 12 low functional load segmental pairs (six vowel pairs, six consonant pairs), as classified by Oh et al. (2015), as well as 12 potentially confusable[8] segmental pairs (eight vowel pairs, four

[7] Note that /x/ and /ç/ are not phonemic in German. They are allophones of the same phoneme in complementary distribution, with /x/ occurring when preceded by a back vowel, and /ç/ occurring when preceded by a front vowel.

[8] As this study serves as a pilot to a perceptual study for second-language speakers of German, those sounds of German that would cause potential difficulty to L1 speakers of English made up the 'confusable segments' group. These errors included phonemic contrasts of German that have been reported to cause difficulty in German learners, both in perception and production (see APPENDIX A).

consonant pairs) were included in the study. Each of the 36 contrasts appeared six times as part of a word embedded in a sentence. The sentences were open propositions that provided little semantic context for the listeners to simply guess what the underlying word was without paying attention to its phonological structure (10). A total of 276 sentences were recorded by an L1 speaker of German. These included 138 distractor items (non-erroneous sentences), 60 sentences containing one segmental error, and 78 sentences containing two segmental errors. This resulted in eight different patterns as shown in *Table 2*.

Table 2 *Error patterns tested*

error category	error category abbreviations	#
distractors	D	138
1 high functional load error	1H	18
2 high functional load errors	2H	18
1 low functional load error	1L	18
2 low functional load errors	2L	18
1 high functional load, 1 low functional load error	1H1L	18
1 confusable segmental error	1CS	24
2 confusable segmental errors	2CS	24
total utterances		276

All produced errors resulted in a phonotactically permissible German non-word. The goal was to observe which phonological neighbours that differed from the target by only one phoneme were activated or served as potential candidates for the German listeners. An example sentence from the experiment to which listeners responded was the following (9).

(9)	Item:	*Der*	**H**a**pp*	*war*	*heftig.*
	IPA:	[dɛːɐ̯	h**a**p	vaːʁ	ˈhɛftɪç]
	Target:	*Der*	*H**ie**b*	*war*	*heftig.*
	IPA:	[dɛːɐ̯	h**i**ːp	vaːʁ	ˈhɛftɪç]
	Translation:	The	hit	was	severe.

The sentences were presented to the listeners auditorily, and the task presented the written sentence with two gaps, thus requiring them to fill in the missing information with a word that was a good match to what they had heard. Irrespective of whether there were one, two, or no segmental errors in the utterance, participants always filled in two gaps so as not to disclose the substitution pattern to them (10).

(10)	Item:	*Der __1__*	*war*	*__2__.*

2.3.3 Procedure

Upon receiving ethics approval from the Conjoint Faculties Research Ethics Board at the University of Calgary, the information about the study was shared in two social media groups, and those who were interested in participating were asked to contact the researcher by e-mail upon which they received the link to the consent form and the experiment. After making initial contact, the participants completed the experiment via a *Qualtrics* survey. In the consent form, participants were informed about the procedure of the experiment. They were asked to give their consent in the online form by pressing a button and signing with their initials. The 276 utterances were divided into 4 blocks of 69 items. Each block was followed by a break. Upon completing the first two blocks, participants were allowed a longer break by filling out a language background survey to avoid listening fatigue. At the end of the experiment, each participant received a 15€ payment.

The 22 listeners completed the listening experiment in about one hour. They had been instructed to complete the experiment in a quiet setting, in front of a computer, with the use of headphones being recommended. They first completed a short training session instructing them to listen to three sentences which they saw simultaneously in an orthographic form on a screen, except for two gaps which contained up to two erroneously produced words resulting in nonwords (11). This training ensured that they would not just copy what they heard without putting it into context or assessing the degree of 'goodness of fit'. To illustrate this: if they heard a sentence like *Der *Hap war heftig*, they were advised to figure out what *Hap* meant instead of just acoustically identifying and copying what they heard.

After completing the blanks in each sentence, they were asked to provide a comprehensibility rating on a 9-point scale ('1' = very easy to understand – '9' = very difficult to understand), as is common in comprehensibility rating tasks (O'Brien, 2014; Thomson, 2018). These ratings served as a measure of the degree of effort the listeners felt they had to put into understanding the utterance.

Additionally, response times were recorded between the first and the last click participants made before progressing to the next item. Only those items taking fewer than 30 seconds were considered for analysis, simply because participants could take longer (excessive) breaks on an individual item before the official breaks after each block if they did not press the arrow at the bottom to proceed to the next item. Participants were unable to proceed, however, before they had assigned a comprehensibility score to an item, which required them to deliberately use a sliding scale after each utterance.

2.4 Data Analysis

Each of the 552 gaps from 276 utterances from each participant was computed for the variables of intelligibility, comprehensibility, and response times. For each listener, mean scores of comprehensibility they had assigned on the 9-point scale were compared for each respective error category (and distractors) and averaged. The lower the comprehensibility score assigned, the easier the utterance was to understand. For intelligibility, the two gaps within each utterance were assigned into the following possible categories: high functional load error (H), low functional load error (L), confusable segmental error (CS), and distractor item (D). Previous studies (Munro & Derwing, 1995; Derwing & Munro, 1997) have measured intelligibility in naturally-produced utterances by counting transcription errors the listeners had made and comparing them with an underlying transcription of the recorded utterance. The current study, however, utilized scripted sentences with a maximum of one segmental error per word resulting in a non-word. The measure of intelligibility was based on whether or not listeners correctly identified the underlying target word[9] (Baese-Berk, Levi, & Van Engen, 2022). The scores were averaged for each respective error group (H, L, and CS) across all 22 listeners. Finally, response times for all seven error categories (along with distractors) were computed and averaged. After the experiment, two other contributing factors were taken into account: the factors of mono- vs. disyllabicity (Is there an impact on intelligibility if the segmental substitution occurs in the first or second syllable of the word?) and syllabic position (Is there an impact on intelligibility if the segmental substitution occurs in onset or coda position?), which were examined by correlating

[9] The score for each correctly identified gap was 1. If the underlying target word was incorrectly identified or reconstructed by listeners, the score was 0.

them to the listeners' achieved intelligibility scores. Additionally, the phonological similarity between the phonemic contrasts in each error type (H, L, and CS) was computed in *Python* based on Frisch's feature weighting the German sound contrasts tested in this study (Dai, 2021). By comparing the segmental features along with proposed weightings, a similarity matrix was generated, which indicates the similarity score between two sounds, e.g., German /z/ and /v/ share 79% similarity, whereas /j/ and /v/ share 57% similarity (Frisch, 1997; Dai, 2021) (APPENDIX B).

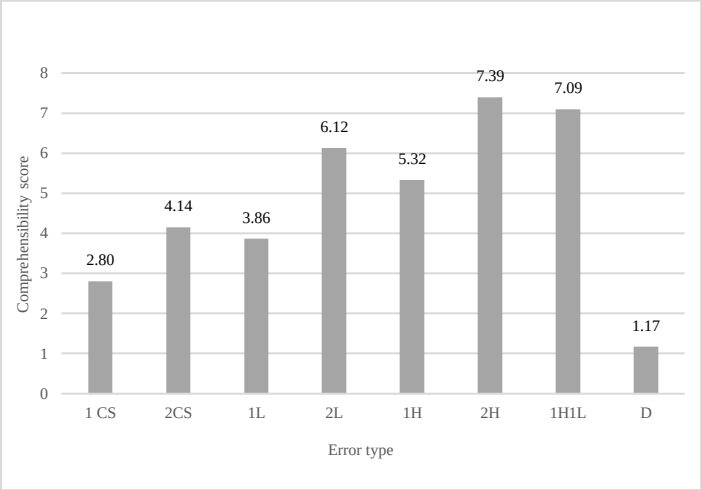
2.5 Results

2.5.1 Comprehensibility

It was hypothesized that utterances containing high functional errors (1H or 2H) would require a greater degree of effort by the listener than utterances containing low functional load errors (1L, or 2L), or those containing confusable segmental errors (1CS, or 2CS). For the 276 ratings assigned by each listener on the 9-point scale, inter-rater reliability was high and exceeded .95 (Cronbach's α).

The comprehensibility scores assigned by each listener on the 9-point-scale for each utterance were grouped into their seven respective error categories (1H, 2H, 1L, 2L, 1H1L, 1CS, 2CS) and then submitted to a one-way repeated measures ANOVA with seven levels of repeated measure. A significant effect of error condition was observed, (F (21,7) = 397.88, p < .001 with a Bonferroni-adjusted level of p <.003. Post-hoc pairwise comparison t-tests revealed that all error conditions were significantly different from each other, except the error category 1H1L compared with 2H. A comparison of the relative effects of high and low functional load errors for comprehensibility revealed that high functional load errors had a more negative impact on

comprehensibility ratings than did low functional load errors. *Figure 2* shows the mean comprehensibility scores achieved for the seven error categories and distractor items.



Note. CS = confusable segmental error; L = low functional load error, H = high functional load error, HL = high and low functional load error, D = distractors

Figure 2 *Comprehensibility scores per error category*

Mixed comparisons between the error groups containing one error only (1CS, 1L, and 1H) and those containing two errors (2CS, 2L, 2H, and 1H1L) showed that sentences with two errors yielded a higher degree of listener effort to understand the utterance than those with a single segmental error, irrespective of their classification as high or low functional load. The only exception was 2CS vs. 1H, where one high FL error was considered more detrimental to comprehensibility than two CS errors. *Figure 3* illustrates that multiple errors had a more profound impact in all comparisons. The point of convergence only occurs at comparison 9 (2CS

vs. 1H) and marks the only instance of one error being more detrimental to comprehensibility than two.

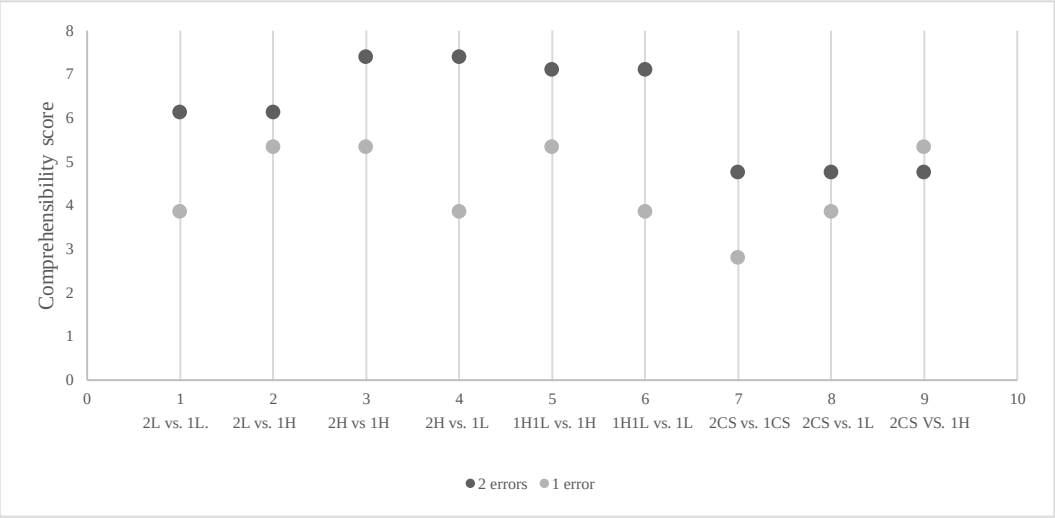


Figure 3 *Pairwise mixed comparisons comprehensibility*

Pairwise comparisons between those categories containing one error versus those containing two errors revealed significant differences (Bonferroni-adjusted) at the $p = <.005$ level, which was the criterion for statistical significance.[10] Additionally, while the post-hoc pairwise *t*-tests confirmed statistical significance, a Cohen's *d* test was conducted to measure effect size. As laid out in *Table 3* below, a medium effect (.56) was measured for the comparison of 1H vs. 2CS, the only case in which one high functional load error weighed more gravely than two errors in a single sentence. All remaining comparisons between utterances containing one vs. two segmental errors yielded large effects in that two errors were always more detrimental to comprehensibility

[10] A Bonferroni adjustment is necessary for running multiple – in this case – pairwise comparisons because it helps control the overall type I error rate by reducing the significance threshold for each individual comparison. It is thus compensating for the increased probability of finding significant results by chance.

than one. Particularly, the comparisons 2H vs. 1L (>4.41) and 1H1L vs. 1L (>4.12) had the largest effects. *Table 3* shows the effect sizes of the significant differences between error groups.

Table 3 Pairwise comparisons comprehensibility

Comparison	2 errors (mean)	1 error (mean)	Actual *p*-value	Effect size (Cohen's *d*)
2L vs. 1L	6.13	3.86	<.001	2.64
2L vs. 1H	6.13	5.33	<.001	2.30
2H vs. 1H	7.4	5.33	<.001	2.20
2H vs. 1L	7.4	3.86	<.001	4.41
1H1L vs. 1H	7.1	5.33	<.001	1.91
1H1L vs. 1L	7.1	3.86	<.001	4.12
2CS vs. 1CS	4.76	2.81	<.001	2.11
2CS vs. 1L	4.76	3.86	<.001	1.00
2CS vs. 1H	4.76	5.33	<.001	.56

These overall large effect sizes are not only indicative of listeners having little difficulty understanding sentences with only one low (L) or confusable segmental error (CS) occurring within a word in an utterance, but they also suggest that the presence of two errors consistently increased the degree of effort to understand the utterance.

2.5.2 Intelligibility

For intelligibility, the scores of '1' or '0', respectively, corresponded to the 552 gaps from the 276 utterances and were labelled as either H (high functional load error), L (low functional load error), CS (confusable segmental error) or D (distractor) in advance of data collection. Across all 22 participants, distractor items (D) received the highest intelligibility scores at 99%, followed

by confusable segmental errors (CS) at 91%, followed by low FL errors (L) at 69%, followed by high FL errors (H) at 46%. Intelligibility scores from all error categories were submitted to a one-way ANOVA with three levels of repeated measures. A significant effect of error condition on intelligibility was observed (F (21,3) = 524.04, p <.000) with a Bonferroni-adjusted level of *p* <.008. Post-hoc pairwise comparison *t*-tests revealed that all error conditions were significantly different from each other. *Figure 4* shows the mean intelligibility score achieved by all listeners of all utterances with distractors (D), confusable segmental errors (CS), low FL (L), and high FL (H) errors. It also indicates the high interrater agreement across listeners.

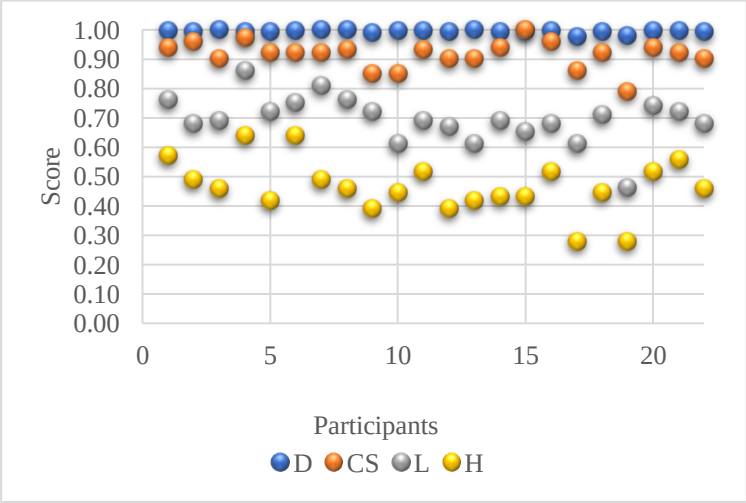


Figure 4 *Intelligibility scores per error category*

A Cohen's *d* test was conducted to measure the large effect sizes of the significant differences across all three error types and distractors. In line with the hypothesis that high FL errors would be most detrimental to intelligibility, followed by low FL errors, and finally, confusable segments that bear a high phonological resemblance, the difference between CS and H yielded

the largest effect at >6.25, followed by L and H >3.7, followed by L and CS >3.48, as laid out in *Table 4*.

Table 4 Pairwise comparisons of error categories for intelligibility

Error category	Error comparison		Actual *p*-value	Effect size (Cohen's *d*)
H vs. L	.46	.70	<.001	3.7
H vs. CS	.46	.91	<.001	6.25
L vs. CS	.70	.91	<.001	3.48

2.5.3 Response times

For the analyzed response time data from all 22 listeners, significant differences between categories high FL (H), low FL (L), confusable segments (CS), and distractors (D) could be observed, with distractor items (D) taking 10.99 seconds, confusable segments (CS) taking 12.41 seconds, low FL errors (L) taking 13.86 seconds, and high FL errors (H) taking 15.29 seconds on average. A one-way repeated measures ANOVA with a level of seven repeated measures was conducted (F (21,7) = 6.78, p <.000) with a Bonferroni-adjusted level of <.003. Post-hoc pairwise comparisons showed that the differences in response time were significantly different among three pairwise comparisons only. These involved error groups containing one versus two errors. A Cohen's *d* effect size test yielded large effects for the comparisons of 1L vs. 2H (1.16) and 2H vs. 2CS (1.04), as well as a medium effect for 2CS vs. 1CS (.7). *Table 5* lays out the pairwise comparisons along with the respective effect sizes for significant differences.

Table 5 *Pairwise comparisons response time*

Pairwise comparisons	Response times in seconds compared	Actual p-value	Effect size (Cohen's *d*)[11]
1CS vs. 1L	9.25 / 10.62	.173	
1H vs. 1L	15.13 / 10.62	.115	
1H vs. 1CS	15.13 / 9.25	.036	
1L vs. 2L	10.62 / 17.03	.036	
1H vs. 2H	15.13 / 18.76	.140	
1L vs. 2CS	10.62 / 11.75	.947	
1H vs. 2CS	15.13 / 11.75	.785	
1CS vs. 2L	9.25 / 17.03	.167	
1L vs. 2H	10.62 / 18.76	.000	1.16
1CS vs. 2H	9.25 / 18.76	.041	
1CS vs. 2CS	9.25 / 11.75	.001	.7
2L vs. 2H	17.03 / 18.76	.489	
2L vs. 2CS	17.03 / 11.75	.085	
2H vs. 2CS	18.76 / 11.76	.000	1.04

The results show that all pairwise comparisons of single errors yielded no difference in response time. Moreover, single segmental errors failed to reach significance compared with two segmental errors of the same error category (i.e., 1H vs. 2H or 1L vs. 2L). Two high functional load errors did not yield longer response times than did two low functional load errors, and neither did two low functional load errors vs. two confusable segmental errors. The only

[11] Effect sizes are only reported when differences were significant based on adjusted p-value <.003

exception showing a large effect in response time differences occurred between two high FL and two CS errors.

Three pairwise comparisons stood out in that they showed significant differences, and they involved comparisons of utterances containing two errors versus one error only: 2CS vs. 1CS yielded longer response times (p = < .001), 2H vs. 1L (p = .000), and 2H vs. 1CS (p = < .041). To underscore the finding that two errors were always more detrimental than one error only, irrespective of classification, the pairwise comparison between 2H and 2L failed to reach significance (p = .489).

Upon submitting the three error categories and distractors to a one-way ANOVA with three repeated measures, with the Bonferroni-adjusted p-value of <.008, (F (21,3) = 9.57, p = <.000), a significant effect of error condition for response time could be observed. Post-hoc pairwise comparisons reveal that CS vs. L and L vs. H failed to reach significance. The only significant difference between two error categories could be observed between the response times of categories CS and H, as well as all error categories (H, L, CS) in a pairwise comparison with the distractors which contained no errors (D). A Cohen's *d* test was conducted on each pairwise comparison that reached significance, revealing large effects between the groups H and CS (.89) and D and H (1.08) and D vs. L (.8)., and medium effects between the group D vs. CS (.64) *Table 6* shows the pairwise comparisons between the three error groups and distractor items with reported effect sizes for those that reached significance.

Table 6 Pairwise comparisons response time 2

Error category	Response time mean in seconds	Pairwise comparisons	p-value	Effect size (Cohen's d)
		CS vs. L	0.037	
H	15.29	L vs. H	0.182	
L	13.86	CS vs. H	0.000	.89
CS	12.41	D vs. CS	0.001	.64
D	10.99	D vs. H	0.000	1.08
		D vs. L	0.003	.8

While high FL errors yielded longer response times compared with confusable segments or distractor items, the response times for high FL were not significantly different from those for low FL errors. Specifically, significant differences with considerably large effects could only be detected for the comparisons of groups 2H vs. 1CS and groups 2H vs. 1L. This shows that, concerning response time, classification of errors as high functional load only mattered if they, too, appeared cumulatively.

2.5.4 Mono- vs. disyllabic stimuli

The stimuli in which a substitution occurred were either mono- or disyllabic. According to Vitevitch (2002), mispronunciations occurring in the first syllable of polysyllabic words should activate a larger set of neighbours, as disyllabic words have more redundant information in the second syllable. Therefore, monosyllabic stimuli should be harder to identify in spoken word recognition (Vitevitch & Luce, 2016). Conversely, Vitevitch (2008) also found that disyllabic words do not necessarily have sparse neighbourhoods in that some polysyllabic items will activate large numbers of neighbours depending on the phonotactic probability. Since these findings do not clearly suggest that the number of syllables correlates with recognition accuracy

or response latencies, the current study measured whether mono- vs. disyllabic stimuli were treated differently by the listeners. As such, intelligibility scores corresponding to identification accuracy achieved by the 22 listeners were computed for 105 monosyllabic stimuli and 105 disyllabic stimuli in all three error conditions (H, L, CS). A two-way ANOVA (3 by 2) design was used to determine whether there is any category effect (of 3 levels) and any group effect (of 2 levels). Results show that there is no statistical significant category by group interaction effect with $F(2,208) = .337$, $p = .714$, and that there is no statistically significant group effect with $F(1,208) = .394$, $p = .531$. This means that listeners' transcription ability was similar across all words containing erroneous segments, irrespective of syllable number. Table 7 shows the mean intelligibility score across all 22 listeners for mono- and disyllabic words containing high FL (H), low FL (L), and confusable segmental (CS) errors.

Table 7 Descriptive statistics of mono- and disyllabic words by error category

Error category	syllable condition	intelligibility mean
H	mono	0.4600
	di	0.4722
	Total	0.4662
L	mono	0.7154
	di	0.7083
	Total	0.7118
CS	mono	0.8778
	di	0.9453
	Total	0.9115
Total	mono	0.6862
	di	0.7086
	Total	0.6975

Table 8 shows the test of between-subjects effects and the results of the two-way ANOVA to which the intelligibility scores for each respective error category per group (mono- vs. disyllabic) were submitted.

Table 8 *Two-way ANOVA results for effects of syllable number and error category*

Source	Type III Sum of Squares	df	Mean Square	F	p-value
Corrected Model	7.197	5	1.439	18.086	0.000
Intercept	103.798	1	103.798	1304.219	0.000
category	7.113	2	3.557	44.691	0.000
group	0.031	1	0.031	0.394	0.531
category * group	0.054	2	0.027	0.337	0.714
Error	16.554	208	0.080		
Total	127.870	214			
Corrected Total	23.751	213			

2.5.5 Syllabic position

The effect of syllabic position of the substitution on word recognition accuracy and speed of recognition was measured, as there was no consensus on whether onset or coda substitutions would be more detrimental to intelligibility (Vitevitch, 2002). Since vowels constituted half of the stimuli and always formed nucleus substitutions, they were excluded from this analysis. The remaining 85 substitutions were considered for analysis. Of these, 56 occurred in onset position and 39 in coda position. The intelligibility score obtained from all 22 listeners corresponds to accuracy of identification and was used to calculate differences between consonantal onset and

coda substitutions. For example, if a participant listened to an item like (11) below, an intelligibility score of '1' was achieved if they correctly identified **jirklich* as *wirklich* 'really'.

(11)	Item:	****J****irklich leicht war der Test nicht gewesen.*
	IPA:	[jɪɐklɪç laɪçt vaːɐ dɛːɐ tɛst nɪçt gɛveːzn]
	Target:	*Wirklich leicht war der Test nicht gewesen*
	IPA:	[vɪɐklɪç laɪçt vaːɐ dɛːɐ tɛst nɪçt gɛveːzn]
	Translation:	The test hadn't been easy, really

An independent samples *t*-test was performed to reveal potential statistically significant group differences for intelligibility scores correlated with onset or coda substitutions. Results showed that there was no statistically significant difference between onsets and coda with $t(93) = .274$, $p = .785$. These findings do not support the hypothesis made based on previous findings on the "special status" of onsets (Vitevitch, 2002) and the co-existing claim that rhymes in which only the onset has been substituted facilitate lexical access (Marlsen-Wilson & Zwitserlood, 1989). Upon closer inspection of individual listener data, many incorrectly identified words were phonological neighbours of the target word consisting of rhymes as in example (12b

dix).

(12)	Item:	*Schon*	*fing*	*der *****E****dler an*	*zu *****w****odeln*		
	IPA:	[ʃoːn	fɪŋ	dɛːɐ eːtlɐ an	t͡suː voːdln]		
	Target:	*Schon*	*fing*	*der Adler an*	*zu jodeln*		
	IPA:	[ʃoːn	fɪŋ	dɛːɐ aːtlɐ an	t͡suː joːdln]		
	Translation:	And so		the eagle	started to yodel		

**Edler* was often identified as *Redner* 'talker', *Wedler* 'wiggler', and **wodeln* was often identified as *rodeln* 'to toboggan', or *brodeln* 'to boil'.

2.5.6 Phonological similarity

Phonological similarity for German phones was computed in *Python* employing the phonological feature weightings proposed by Frisch (1997). The feature weightings are derived from phonological patterns and constraints within the German language and include frequency and informativeness in the language based on entropy. The similarity score between two sounds can then be calculated by adding up the weights of matching feature values and subtracting the weights of their mismatching feature values. As such, a template of a German phonological inventory was downloaded from PHOIBLE and a similarity score matrix (APPENDIX C) was generated in *Python* (Dai, 2021). The program used the default weight '1' for all features, making the basic assumption that the distance between '+' and '-' values for phonological features are equal. The distance between two segments (x and y) is defined as the summed weights of their unshared features with respect to the feature set *F* and a weighted feature lattice *w*, followed by an equation adapted from Wilson and Obdeyn (2009). Weights and distances are then specified in the *dlist* in the program (APPENDIX D).

The results showed that high FL segmental contrasts bore an average similarity of .76, low FL of .78, and CS of .93. *Table 9* shows the weighted phonological similarity scores between the segmental contrasts tested in this study. The respective colours mark the error category into which the contrast belonged.

Table 9 *Weighted phonological similarity scores of phonemic contrasts tested*

	pf	k	n	z	j	y:	i:	ɪ	e:	o:	a:	u:
v				0.79	0.57							
t	0.75											
t͡s				0.93								
s			0.79									
ç												
x		0.93										
ʁ			0.54									
e:							0.96					
a							0.79				0.96	
ɑ:						0.75			0.86			
ø:						0.93				0.93		0.89
œ								0.89				
o:							0.82					

Note. red = high FL, apricot = low FL, mint green = confusable segments

More phonologically dissimilar contrasts categorized as high FL correlate with lower intelligibility scores and decreased comprehensibility, whereas confusable segments that had been classified as phonologically similar correlate with higher intelligibility and comprehensibility scores. Like the high FL segments, the low FL segments were also classified as having low phonological similarity, which suggests that both high FL contrasts and low FL contrasts should have been equally difficult to understand and yielded lower comprehensibility scores. The significant difference between high FL and low FL contrasts for intelligibility and comprehensibility can be explained, however, due to the similarity score of .93 between the low FL phonemic contrast /oː/ and /øː/, which were categorized as highly similar. Upon close inspection of those utterances with /oː/ and /øː/ substitutions, they could be determined as outliers and are likely the source of higher intelligibility and comprehensibility scores achieved

by the listeners for the in the low FL category. Because listeners were more accurate and reported more ease of understanding for /oː/ and /øː/ substitutions as opposed to the other low FL phonemic contrasts, this likely led to a correction of the overall increased intelligibility and comprehensibility scores for the low FL error category[12]. This explains why low FL errors yielded higher intelligibility scores (accuracy) and decreased effort compared with high FL errors, even though the phonological similarity averages were almost identical. Overall, this is indicative of phonological similarity playing an important role in the word recognition process, in that recognition accuracy was largely dependent on the degree to which the acoustic-phonological structure had been altered (Vitevitch & Luce, 2016). Similarly, confusable segmental contrasts were recognized with high accuracy and greater ease because the phonological similarity between the underlying and the substituted segment was high and the phonological structure well preserved.

2.6 Discussion

The results of the present study indicate that substitutional errors involving high FL phonemic contrasts of German negatively affect intelligibility and comprehensibility to a higher degree than low FL substitutions or confusable segmental errors, which is in line with the studies empirically testing FL for English (Munro & Derwing, 2006; Suzukida & Saito, 2019). This error category effect is diminished when multiple errors occur. In these instances, low and high FL substitutions do not differ significantly in listener recognition accuracy (intelligibility) and

[12] Upon excluding the /oː - øː/ contrast from the low FL error category, intelligibility computations show a decrease in intelligibility from 70% to 62%. Conversely, in the high FL group, the contrast with the lowest phonological similarity score /ʁ - n/ yielded particularly low intelligibility scores of 42%, which was below the high FL average (46%).

assessment of the goodness of fit degree (comprehensibility). Response latencies failed to reach significance when comparing high with low FL errors. It was only when two substitutions were present within an utterance that response times were significantly longer compared with one error only. These findings suggest that the presence of any two errors has a greater impact than the presence of one error, irrespective of its classification as high or low FL. Previous research on FL in English found only high FL errors had a cumulative effect on comprehensibility, whereas multiple low FL errors did not hinder comprehensibility (Munro & Derwing, 2006). These findings were thus not replicated in the current study, which can be attributed to FL being language-specific as well as the fact that the current study tested errors that rank high or low on the extremes of the FL hierarchy in German and that are typically not confused in real life.

The significant differences between high and low FL errors on intelligibility and comprehensibility observed in the current study seem to support the notion of FL, in that some phonemic contrasts form more minimal pairs than others and therefore hinder word recognition accuracy. However, we have to consider that those segmental substitutions used in the study were neither confusable in the high FL category, nor in the low FL category. Their substitutions therefore often led to a striking change in the acoustic-phonological structure of a stimulus to the target word, as shown in example (13).

(13) Item: *Er war sehr *blann*

IPA: [ɛːɐ̯ vaːʁ zɛːʁ blan]

Target: Er war sehr blass

IPA: [ɛːɐ̯ vaːʁ zɛːʁ blas]

Translation: He was very pale

The listeners indicated that this utterance was difficult to understand and identified the underlying word as *klamm* ‘clammy’, *blank* ‘shiny’, or *bang* ‘fearful’. This illustrates that the listeners did not process the heard stimulus sequentially, phoneme by phoneme until detection of the erroneous segment and then struggled with its recognition due to the high number of minimal pairs /s / and /n/ form in German, as postulated by minimal pair counts that determine the FL of the segments. Rather, this acoustic-phonological change in the stimulus’s structure caused them to compare it to a set of minimal pairs (neighbours) that resemble **blann* and best fit into the context of the utterance. This highlights that neither for high nor for low FL contrasts did the listeners have access to a set of minimal pairs created by a particular phonemic opposition, which can then be excluded as the source of decreased intelligibility and comprehensibility for high FL errors. Therefore, it was examined whether the degree of phonological similarity of a phonemic contrast could explain the significant differences between high and low FL errors. Initially, the averaged similarity score for high and low FL phonemic contrasts tested in this study indicated almost identical similarity scores, with high FL contrasts sharing 76%, and low FL sharing 78% of the most salient phonological features. This means that the two error groups included segmental contrasts that are not commonly confused, e.g., /n - s/ for high FL, but also /j - v/ for low FL. Upon scrutinizing the phonemic contrasts in both error categories along with the obtained listener data, however, it became evident that the averages of both groups had a different composition. In particular, the low FL group exhibited one phonemic contrast (i.e., /øː - o/) with a particularly high phonological similarity score of (94%), whereas the high FL group had one phonemic contrast with a very low phonological similarity score of /n - ʁ/ (54%). The listener data for these two contrasts offer reason to believe that the significant differences in intelligibility and comprehensibility between high and low FL errors hinge on these two

particular outliers. For items with /øː - oː/ substitutions, intelligibility as well as comprehensibility scores were as high as in the CS group, thereby correcting upward the average low FL contrasts for accuracy and degree of goodness of fit. Conversely, the /n - ʁ/ distinction in the high FL group received the poorest intelligibility and comprehensibility ratings, correcting downward the overall accuracy and degree of goodness of fit. This is in line with one of the main premises of FL: the more phonologically distinct the phonemic contrast, the higher its functionality in keeping utterances apart (King, 1967). The individual listener choices of words, as illustrated in the example above (13) indicates that they process the stimuli in their entirety and compare the acoustic-phonological input to possible matches from memory, as proposed by NAM (Luce & Pisoni, 1998). The transcriptions are indicative of listeners resorting to rhymes or the next possible phonological-acoustic match (14).

(14) Item: *Das Tabu war nicht sonderlich *tachtvoll*

IPA: [das tabuː vaʁ̞ nɪçt zɔndʁ̞lɪç ta**x**tfɔl]

Target: *Das Tabu war nicht sonderlich taktvoll*

IPA: [das tabuː vaʁ̞ nɪçt zɔndʁ̞lɪç ta**k**tfɔl]

Translation: The taboo was not exactly tactful

Here, listeners transcribed and entertained the following underlying words: *achtvoll* ‘attentive’, *machtvoll* ‘powerful’, *prachtvoll* ‘glorious’. It was easier for them to arrive at a rhyming word while keeping the phonological structure of the word intact than to process the stimulus sequentially until the detection of the /k/ substitution. These alternatives are often minimal or near-minimal pairs or phonological neighbours of the encountered stimulus. The acoustic-phonological similarity of the stimulus was low when the substituted segment was phonologically dissimilar to the underlying one, which then activated a larger set of neighbours,

i.e., *saugen* ‘to suck’ had been substituted with /v/ and became **waugen*. Listeners indicated a very high degree of effort and were often inaccurate in recognizing the word, as neighbours like *laufen* ‘to run’, *wauen* ‘to bark’, *glauben* ‘to believe’, *fauchen* ‘to hiss’ were also activated. Conversely, when a phoneme had been substituted with a fairly phonologically similar substitution, possible neighbours were limited, and intelligibility and comprehensibility scores were high, i.e., the /œ/ in *Töpfer* ‘potter’ had been substituted with /oː/ and became **Topfer*. All listeners agreed on a high ‘goodness of fit’ degree (comprehensibility) and achieved high accuracy scores (intelligibility). Possible neighbours like *Opfer* ‘victim’, *Klopfer* ‘beater/knocker’, *Stopfer* ‘stuffer’ were not entertained, because the acoustic-phonological similarity guided listeners in the direction of *Töpfer*. These findings underscore the importance of phonological similarity in segmental substitutions over raw minimal pair count formed by a particular contrast. It is also worth mentioning that highly dissimilar acoustic-phonological neighbours are less likely to occur in real-world speech output, as erroneously produced words containing segmental errors usually bear some resemblance to the target word. This means that some substitutions are more likely to occur than others. For example, substitution of /œ/ in *Töpfer* with /o/ **Topfer* is common, whereas substitution of /œ / with /iː/ ****Tiepfer*** is unlikely.

Irrespective of acoustic-phonological similarity guiding the number of neighbours activated in a listener’s memory, some words have more neighbours (minimal pairs) than others, and these are entertained due to their frequency, word class membership and or syntactic position. As Levis & Cortes (2008) point out, some contrastive items can be excluded to be the underlying word, because minimal pairs are not always part of the same word class and/or do not fit in the same syntactic position. For example, if a non-word like **schassen* formed a minimal pair with *schießen* ‘to shoot’ based on the phonemic contrast of /a - iː/, the low intelligibility

scores and high degree of effort comprehensibility scores could not only be attributed to the relative dissimilarity between /a/ and /iː/, but also because *schassen* occurs in a dense neighbourhood in the context of an utterance and activated numerous (near-) minimal pairs: *schaffen* 'to achieve', *waschen* 'to wash', *schätzen* 'to appreciate/to estimate'. Conversely, an item like *lieb* 'lovely, nice' in which the substitution was phonologically very similar **leeb* still resulted in low intelligibility scores and increased degree of effort, because many neighbours were available for *leeb* in the context of the utterance: *stet* 'steady', *leer* 'empty', *lahm* 'lame'. Taken together, there is ample evidence suggesting that phonological similarity of the phonemic substitutions positively correlates with intelligibility and comprehensibility and thus the driving factor for recognition accuracy (intelligibility) and degree of goodness of fit (comprehensibility), followed by the number of lexical neighbours (minimal pairs) a stimulus has that fit the context of the utterance.

2.7 Conclusion

While the notion of FL seems appealing, because it deems responsible very concrete and specific phonemic contrasts within each word for communication, the current study suggests that intelligibility of utterances may hinge on a word in its entirety and may ultimately be linked to word recognition processes. Moreover, the findings raise doubts about the ecological validity of FL in languages other than English[13].

[13] Oh et al. (2015) examine nine languages and the functional loads carried by vowels, consonants, tones, and stress. The results showed that in German, consonants and vowels do not perform the same contrastive work as they do in English. In a language like French, for example, vowels are extremely important to perform contrastive work, whereas consonants are not. Conversely, in a language like Japanese, vowels have little importance and carry low functional loads overall.

Despite its explanatory value to make generalizations in support of prioritizing certain English segments over others, the concept carries several documented caveats (Sewell, 2017). Firstly, it has been acknowledged that genuine high FL contrasts are typically not conflated (Brown, 1991; Levis, 2020). The few exceptions reported as high FL for English, like /b - v/ and /r - l/ rely on sheer minimal pair count and involve phonemic contrasts that are prone to confusion by some L2 speakers of English. It is important to note, however, that they do not appear among the top 20 high FL contrasts in Oh et al.'s (2015) calculations. It seems contradictory that phonologically similar segmental oppositions would rank as high FL to begin with given that one of the main premises of FL is that phonologically dissimilar contrasts can distinguish utterances best[14]. Furthermore, FL research that has been done on English acknowledges that it is seldom the case that an utterance will be confused based on a particular phonemic distinction, as few real minimal pairs can function in the same syntactic position, that are members of the same word group, or that exist with the same frequency (Levis & Cortes, 2008).

These reported reservations for research on FL in English stand out clearly in German. Not only do we find no commonly confused contrasts as part of the high FL group, but those minimal pairs that hinge on a particular segmental contrast are scarce in a synthetic language with rich inflectional morphology. The following examples in (15) and (16) illustrate the difference between English and German.

[14] Note here that the /eː - iː/ contrast in German yielded high phonological similarity computations and was therefore tested within the confusable segments group, but a post-hoc glance at complete computational lists from Oh et al. show that this vowel contrast ranks relatively high in the hierarchy. Nevertheless, intelligibility and comprehensibility scores were high for this contrast in that listeners found utterances easy to understand and arrived at the target word effortlessly due to the well-preserved acoustic-phonological similarity of the speech stimulus.

(15) Target: *I saw the ball*.

Output: *I saw the bar*.

If /l/ is substituted with /r/ in English in this particular case, misunderstandings can occur, because both *ball* and *bar* fit the context of the utterance and the sentence remains grammatical. Conversely, for German, substituting /l/ with /r/ will make the sentence ungrammatical, because the feminine accusative determiner *die* [article.fem.acc] is not in agreement with the masculine noun *Ball*.

(16) Target: **Ich sah die*(f.) *Ball*(m.)

Output: *Ich sah die*(f.) *Bar*(f.)

The current study examined the interaction between word recognition processes and FL. FL suggests that words are processed sequentially in a phoneme-by-phoneme approach until detection of the erroneous segment. The evidence suggests, however, that words are processed top-down in search of the closest acoustic-phonological match that is often a (near-) minimal pair, a neighbour to the input stimulus. The more phonologically obscure the stimulus becomes, the less intelligible and comprehensible the utterance.

In summary, the ecological validity and operationalization of the notion of FL is a difficult undertaking for German. For German, phonologically dissimilar contrasts that would not be conflated in real life in the first place obscured the underlying target word to a high extent and numerous other lexical neighbours were activated with varying degrees of goodness of fit. For research on English, it is also a previously acknowledged problem that genuine high FL contrasts are not commonly confused, which is exactly **why** they carry a high FL: they can best perform the work of keeping utterances apart (Brown, 1991; Levis, 2020). However, some confusable contrasts of English, like /b - v/ or /r – l/, rank relatively high in the functional load

hierarchy compared to others, like /s – θ/ (Munro & Derwing, 2006; Suzukida & Saito, 2019). Moreover, we have to acknowledge the low probability of two lexical items hinging on a particular phonemic distinction becoming problematic because of syntactic and semantic constraints, which is seldom the case for English as an analytic language, but especially for German as a synthetic language. This is also the reason why it is difficult to predict the effect on intelligibility and comprehensibility of spoken utterances based on minimal pair counts formed by a particular phonemic distinction. Firstly, irrespective of sheer minimal pair number, there are further constraints (e.g., word class, syntactic position, or context) that will exclude a minimal pair on the basis of this contrast. This is especially the case for a synthetic language like German, which is rich in inflectional morphology and grammatical gender and case assignment. Secondly, the current study has shown that loss of intelligibility cannot be attributed to the number of minimal pairs (real words) a particular contrast forms, which is central to the notion of FL, but that intelligibility and comprehensibility are decreased even if a particular phonemic distinction results in a non-word. Especially when phonologically dissimilar segments are being substituted, the overall acoustic-phonological structure of the target word becomes obscured. Specifically, listeners compare the overall acoustic-phonological structure of a word to possible lexical neighbours, thus, a set of words that forms (near-)minimal pairs with the heard stimulus, which is not necessarily done based on minimal pair count of particular phonemic opposition (Vitevitch & Auer, 2000). These findings suggest that the FL principle has merit in theoretical linguistics in that the functionality to communication particular sound contrasts is quantifiable. Typically, such high functional load oppositions involve contrasts that are phonologically so distinct that they can easily perform this contrastive work. With respect to real-world applicability, however, i.e., measuring which sound contrasts matter in spoken word recognition based on the number of

minimal pairs that hinge on a particular phonemic distinction, functional load cannot accurately predict the loss of intelligibility and comprehensibility of utterances in real-life communication scenarios for the two reasons outlined above. Nonetheless, producing errors of individual segments in speech output can still affect comprehensibility and intelligibility, especially when multiple errors are present. Listeners take longer to process the utterance and indicate higher degrees of effort to understand the utterance because multiple erroneously produced words need to be compared to their lexical neighbours (Vitevitch & Luce, 2015).

Research on the applications of FL for English gave reason to hope that it could inform pronunciation pedagogy. That is to say, if certain sounds perform more work in keeping utterances distinct and therefore contribute to intelligibility, they should not only be prioritized, but they are more discrete, identifiable, and can be taught in isolation as opposed to more complex pronunciation features that are difficult to teach, e.g., stress or intonation. However, the current study indicates that both the concept and the application of FL need to be put on a sounder footing (Sewell, 2017). For German, prioritization of sounds purely based on FL would be futile for L1 speakers of English, because high FL contrasts are not actually confusable for them. One larger issue about notion of FL with respect to its applicability is the implication it makes regarding very particular phonemic oppositions driving the intelligibility and comprehensibility (and thus word recognition accuracy) of utterances. Firstly, this would mean that intelligibility, and conversely, the loss of intelligibility hinges on the number of real word minimal pairs formed by a particular phonemic contrast that fit in the utterance syntactically and semantically. Secondly, such an assumption would imply that listeners employ a bottom-up approach to word recognition in which they process the word sequentially until detection of the erroneous segment.

However, the results of the current study suggest that listeners employ a top-down word recognition approach, by which not the number of minimal pairs that hinge on a particular phonemic opposition is crucial, but rather, the number of (near-)minimal pairs, called lexical neighbours (intelligibility) that resemble the acoustic-phonological structure of utterance and constitute a good fit (comprehensibility) and become activated. This means that activation of minimal pairs also occurs with non-words. When a phoneme with a high phonological similarity to the target one is substituted, chances of the overall acoustic-phonological structure of the word being preserved is high, which is why substitution of /øː/ in *römisch* 'Roman' with /oː/ **romisch* would not alter the overall acoustic-phonological structure of the target word to a high extent. Firstly, the quickest 'repair' to arrive at the target word would require one operation of substituting a highly phonologically similar target sound, which would make **romisch* a real word by substitution of /øː/. Secondly, the number of lexical neighbours **romisch* generates, e.g., *komisch* 'weird', which would fit into the context of the utterance, will be a more accurate predictor of intelligibility and response times than assuming the listener will have knowledge of minimal pairs formed by a particular phonemic opposition, namely /oː - øː/ in this case. Similarly, we see that the /iː - eː/ contrast may carry a relatively high FL, even though this sound was tested as part of the confusable segment group, but that listeners had little difficulty in correctly identifying **Deeb* as *Dieb* 'thief', possibly because the number of lexical neighbours activated was low due to the acoustic-phonological structure of in the non-word substitution being well preserved by substitution /iː/ with /eː/.

These findings about FL suggest that the notion of some phonemes performing more contrastive work than others in keeping utterances apart is difficult to operationalize. While the theoretical framework of functional load clearly shows that phonemic contrasts that keep many

words apart are typically phonologically distinct and therefore unlikely to be confused, the current study was only able to empirically test the functional load hierarchy of German by simulating substitution patterns which are unlikely to occur in real life. The results suggest, however, that functional load as a quantitative measure cannot predict the loss of intelligibility and comprehensibility on the basis minimal pairs hinging on particular important phonemic oppositions. Rather, we see that spoken utterances containing one or two errors, even when resulting in non-words, can be the source of the loss of intelligibility and comprehensibility and yield longer processing times when the acoustic-phonological structure of the erroneous output stimulus activated a set of similar-sounding words that with varying degrees of goodness of contextual fit for the listeners. Analysis of the listener data strongly suggests that they employ a top-down word recognition approach to find the underlying target word instead of a bottom-up approach until detection of the erroneous segment.

In the past, the notion of functional load and its theoretical validity has been utilized to inform the prioritization of sound segments for English as second language (ESL) pronunciation instruction. The idea is that if certain segments that are prone to confusion in English carry a high functional load, e.g., /r - l/ or /b - v/, they should be prioritized in the perceptual and production training tasks. Conversely, L2 pronunciation pedagogy sought to employ the functional load principle to steer away from focusing on sound distinctions in English whose substitutions have little impact on intelligibility and comprehensibility, because they do not carry a high functional load, and only contribute to accent reduction (Munro & Derwing, 2006). While past studies have shown some promising results in that the functional load principle was able to inform such selection of particular sound contrasts of English for pronunciation teaching, the applicability to other languages seems to bear some difficulties. Moreover, there is an urgent

need to scrutinize which and when minimal pair counts (as part of simple or complex count measures) hinging on the distinction of a particular phonemic oppositions become problematic for intelligibility and comprehensibility of utterances, especially if a bottom-up processing approach is assumed. The current findings warrant a closer examination of functional load in conjunction with word recognition models and require a more solid understanding of how individual segmental substitutions can impact intelligibility and comprehensibility, even for non-words, irrespective of their functional load classification from theoretical computation.

Chapter 3: Effects of audio vs. audiovisual training on the perception of sounds by learners of German

Abstract

This study investigates whether beginning L2 learners of German can be trained to extract phonetic information from audio and audiovisual training of novel phonemic contrasts. Several phonemic consonant and vowel contrasts of German that are commonly confused by L2 speakers were used. Participants included twelve students from a large Canadian research university who were tested on their perception of German sounds. The study took place over the course of six weeks, with 18 20-minute sessions of either high variability audio-only (HVPT-A) or high-variability audiovisual training (HVPT-AV). Each modality's training sessions featured eight (4 female, 4 male) L1 speakers of German with distinct voices and, in the case of HVPT-AV, outward appearance. The results showed that the German beginner learners did not significantly benefit from the longitudinal HVPT overall. The HVPT-A group significantly outperformed the HVPT-AV group for most contrasts tested, but the pre- and post-test results indicate that both groups achieve higher discrimination accuracy when listening to a single speaker without the presence of visual information.

3.1 Introduction

The current study seeks to explore the acquisition of German phonemic contrasts among L2 learners of German. In particular, it compares the effectiveness of audio-only training with audiovisual training over the course of six weeks on learners' ability to discriminate German sound contrasts. The goals are to find out whether high-variability phonetic training (HVPT) leads to perceptual gains in both groups and to determine which kind of training modality

(HVPT-A or HVPT-AV) is more effective. Due to the increased stimulus robustness that results from idiosyncratic speaker detail, including phonetic variability from multiple speakers, the training has been found to be particularly successful in increasing learners' flexibility in developing L2 phonological categories. By adding a visual component to the training paradigm, the benefits of the technique are hypothesized to be further enhanced by providing additional rich detail to the stimuli.

3.1.1 High-variability phonetic training

Substantial work over the past decades has been dedicated to the process of L2 speech perception in an attempt to determine and better understand difficulties L2 learners may have in perceiving L2 sounds. It is likely the case that those difficulties in acquiring an L2 stem from the interference of the L1 phonological system. As such, learners, when confronted with novel sound contrasts, process them through the system of the familiar L1 (Flege, 1995; Best, 1995).

Perceptual training plays an important role in improving the perception of speech sounds. In particular, audio- and audiovisual training have been explored in a training technique called *high-variability phonetic training* (HVPT), which has found support over the last decades. In particular, it exposes L2 learners to a wide range of instances of speech sounds to enhance the perceptual flexibility and stimulus robustness in the listener (Lively et al., 1993). As such, learners are exposed to multiple tokens of minimal pairs of words containing novel sounds in different syllabic positions produced by multiple talkers. This computer-assisted training paradigm directs learners' attention to phonetic learning and the perception of sound outside the context of meaning (Thomson, 2011). Numerous studies have reported the success of HVPT over low-variability phonetic training (LVPT, that is training that involves single talker stimuli in stable phonetic environments) and found that HVPT may be more effective than other sorts of

perceptual training (e.g., Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). The majority of HVPT studies to date have been conducted with L2 English speakers, although studies have been carried out on L2 Japanese (Hirata, 2004), French (Kartushina & Martin, 2019), and Hindi (Pruitt, Jenkins, & Strange, 2006), respectively.

Generally, research employing an HVPT technique presents minimal pair stimuli (real or non-words) in forced-choice identification tasks if the participants are already familiar with the language and its orthography (Bradlow et al., 1999; Hwang & Lee, 2015) or ABX discrimination tasks, which are particularly useful when dealing with naive or beginner learners of an L2 (Iverson, Hazan, Bannister, 2005; Lim & Holt, 2011). Both procedures are deemed equally effective and promote learning gains, generalization, and retention (Flege, 1995; Carlet, 2007). Moreover, the advantage of this training technique lies in its flexibility in that it allows learners who partake in a longitudinal training approach to complete the individual sessions in a comfortable location, at their own pace, without having to sit in a laboratory. HVPT also includes corrective feedback regarding listeners' response accuracy to make them aware of incorrect responses, which helps the listener to actively re-tune their perception of whether their choice was correct or incorrect (Shinohara & Iverson, 2018; Tajima, Kato, Rothwell, Akahane-Yamada, & Munhall, 2008; Thomson, 2016). Despite the focus on variable input, the numerous studies that followed Logan et al.'s (1991) and Lively et al.'s (1993) seminal studies and that have investigated the success of HVPT are not direct replication studies and therefore relatively heterogeneous in terms of training procedure (Brekelmans et al., 2022; Thomson, 2018). As such, training duration, number of sound contrasts tested, learner proficiency levels, and number of speakers vary greatly. Moreover, very few studies directly compare HVPT to LVPT (one talker, little variety in phonetic contexts, but see Perrachione, Lee, Ha, & Wong (2011) and

Giannakopoulou et al. (2017) for examples).

While training implies that learning is achieved through repeated stimulus exposure, there is no consensus in previous studies regarding the training duration required to obtain perceptual gains. As such, training duration ranges from 5 sessions (once per day) (Leong, Price, Pitchford, & van Heuven, 2018) to 45 sessions over the course of 3-4 weeks (Bradlow et al., 1999), with varying length of 12 to 60-minute sessions. Furthermore, the number of phonemic contrasts tested in these studies differs greatly. Some test only one critical contrast, like English /i/ and /ɪ/ (Giannakopoulou et al., 2017), whereas others test up to 14 English vowels (Thomson, 2012; Shin & Iverson, 2013). Moreover, the proficiency levels of learners vary between completely naïve monolinguals who have never heard the target language (Hirata, 2004; Pruitt, Jenkins, & Strange, 2006), and those learners of a language with up to 23 years of learning experience (Rato & Rauber, 2015), with the vast majority of these published studies testing HVPT in highly advanced learners. Additionally, while most HVPT studies employ several speakers ranging from five (Lively, Logan, & Pisoni, 1993; Nishi & Kewley-Port, 2007; Shin & Iverson, 2013) to ten (Perrachione et al., 2011; Iverson, Pinet, & Evans, 2012), only three compared the differences between low variability training (one talker) and high variability training (multiple talkers) (Lively et al. 1993; Perrachione et al., 2011; Giannakopoulou et al., 2017). Some studies concluded that certain talkers may be more effective than others (Thomson, 2012) and that individual listener aptitude may play a role in the success and ultimate outcomes of HVPT (Perrachione et al., 2011; Giannakopoulou et al., 2017).

While the vast majority of HVPT studies employ audio-only, audiovisual perceptual training has also shown positive effects. For example, Hazan et al. (2005) found that the English sound contrasts /b - v/ and /r - l/ presented to high-proficiency Japanese listeners were easier to

identify in audiovisual training than in an audio-only training modality because the learners had access to information from lip movement. Adding visual information, such as seeing the articulatory gesture, the speaker's face and their mouth movements may thus enhance the robustness of a stimulus.

3.1.2 Theoretical frameworks informing the HVPT technique

Lively et al. (1993) describe the goal of HVPT training as increasing the robustness of novel phonemic contrasts through exposure to several phonetic contexts and novel talkers. Previous studies that build on this landmark study accept that HVPT is successful in improving perceptual discrimination of L2 sounds and in facilitating pronunciation. However, the known previous work on HVPT does not rely on a theoretical framework that could adequately account for the advantage of high variability (multiple talkers, varied phonetic contexts) over low variability (single talker). This section, therefore, revisits the most frequently cited speech perception models informing HVPT: the Perceptual Assimilation Model (PAM), Motor Theory and exemplar theory (ET).

In recent HVPT studies, the most frequently relied-upon theoretical framework for understanding how L2 learners perceive L2 sounds is the *Perceptual Assimilation Model L2* (PAM-L2) (Best, 1995; Best & Tyler, 2007). PAM-L2 posits that learners perceive novel sounds by mapping them onto L1 language sound categories. The degree of similarity between the L2 sounds and the learner's L1 language sound categories influence this assimilation process. As such, this process involves integrating novel sounds into existing sound categories of the L1 or creating new categories for the L2 sounds. For example, if an L2 phonological contrast exists in the L1 as well, discrimination of the two sounds is easy (e.g., German /v/ and /f/, which are also contrastive in English). Similarly, if an L2 phonological contrast does not exist in the L1 (e.g.,

German /uː/ and /yː/), discrimination will be difficult for an L1 English speaker, because there is only one phonological category for /uː/ in English. Through perceptual training, the L2 listener of German must now learn to establish a novel category for the German /uː - yː/ contrast. In the context of HVPT, PAM(-L2) guides the selection of L2 target sounds and the design of training materials to facilitate new category formation and thus prevent assimilation of target sounds to the L1 sound system.

One of PAM(-L2)'s major tenets is that this mapping process is strongly influenced by L2 learners' perception of articulatory gestures required to produce the L2 sounds, which has its foundation in the theoretical framework of Motor Theory, which claims that speech perceivers do not perceive sounds but gestures (Liberman & Mattingly, 1985). PAM unites the theories of listeners perceiving sounds along with an articulatory gesture. This means that perception of speech sounds is not solely reliant on auditory input, but that it is also influenced by the kinesthetic experiences associated with producing these sounds (Galantucci, Fowler, & Turvey, 2006). Thus, for L2 learning, as learners engage in producing L2 sounds, they also have to learn how to identify the relevant (and invariant) gestures to form an accurate and detailed perceptual representation of particular speech sounds. This theory can help account for the higher effectiveness of audiovisual training over audio-only training in Hazan et al.'s (2005) study.

Despite their strong associations with HVPT, these models cannot account for the high variability aspect in the technique that seems to come with an advantage over low variability training (i.e., one voice, few different phonetic contexts). While PAM(-L2) provides valuable insights into the perceptual categorization process, it cannot fully capture the intricate dynamics of speech perception. As such, PAM(-L2) assumes discrete categorical boundaries between sounds, which overlooks the inherent variability and context-dependent nature of speech

perception that happens in real-world communication.

In the first HVPT study, Lively et al. (1993) found that L2 English listeners benefit from the variability of stimuli, as "learners develop talker-specific, context-dependent representations for new phonetic categories" (1). These findings lend support to an alignment between HVPT and an exemplar-theoretical approach. Exemplar Theory (ET) (McClelland & Elman, 1986; Hintzman, 1986; Johnson, 1997; Goldinger, 1998) offers an alternative perspective that aligns well with the principles of HVPT and that can account for higher perceptual accuracy due to exposure to a diverse range of speech exemplars, including different talkers and stimuli variations. Although ET does not assume abstraction, it suggests that speech perception is based on the storage and retrieval of specific instances of exemplars encountered during language acquisition throughout a human's lifetime. These exemplars (i.e., a heard word) enter our memory as we perceive them along with the rich detail and leave a trace from which the listener can generate an abstract concept on the fly (Hintzman, 1986). For a speech stimulus, this means that perceptual details are stored along with the stimulus such as rich idiosyncrasies of the speaker's voice, facial expressions, and the setting in which it was perceived. The category in which an exemplar is stored along with similar exemplars expands and increases in robustness with exposure over time.

The complementary aspects of the two models of PAM(-L2) and ET can, however, be integrated to enhance HVPT. It is conceivable that listeners process L2 sounds through the system of the L1 and that perception is categorical, which means that heard speech sounds will be assigned to a phonological category through invariants in the speech signal. By systematically exposing learners to contrastive (L2) sounds and facilitating discrimination between them, HVPT promotes the development of clear perceptual categories.

ET can account for HVPT's success by emphasizing the importance of perceptual flexibility and the capacity to capture the inherent variability in speech perception. The diversity of exemplars in HVPT contexts, including talker and stimulus variation, aids learners in the process of forming rich and context-specific representations, thereby enhancing their ability to generalize learning to different speakers and ecologically valid communication situations. In summary, this means that PAM(-L2) and ET can together account best for the idea that learners establish perceptual categories for L2 sounds and benefit from variability facilitating robustness and adaptability thereof across diverse contexts.

3.1.3 The rationale for the current study

The effectiveness of HVPT has predominantly been tested in advanced language learners with years of experience in the L2 through either formal instruction or living in the L2 environment (Lively et al., 1993; Lengeris & Hazan, 2010; Lim & Holt, 2011). In line with this, benefits of HVPT are seldom reported for beginning L2 learners (Hirata, 2004). While PAM(-L2) serves as the framework for the majority of HVPT studies, the exemplar theoretical approach's focus on stimulus variability and robustness has been tested in a study comparing audio-only with audiovisual training (Hazan et al., 2005). While Hazan et al. (2005) found that audiovisual training was more successful than audio-only training, they ascribed the improvement to listeners being able to see the gesture. However, some sounds, especially vowels, are not visually or gesturally distinguishable, as Hazan et al. acknowledge. An ET approach would be able to account for the higher success rates of improved perception, not only because listeners extract information from the articulatory gesture itself, but due to the visual component in its entirety adding to the robustness and variability of the speech stimulus. If audiovisual learners outperform audio-only learners, this could be evidence in support of HVPT capitalizing on the

strength of ET.

The current study tests the HVPT technique in two L2 German beginner groups (audio vs. audiovisual) in a pre- and post-test design with an in-between training period of 6 weeks, 3 sessions per week, 20 minutes per session. An ABX discrimination task with immediate feedback was chosen, as an identification task would require a) knowledge of the language and the corresponding graphemes to match the sound; or b) training of which particular sound can be ascribed to a symbol if one were to leave out orthographic training[15].

3.1.4 Predictions

Based on previous research, the following predictions were made.

Audio-only vs. audiovisual HVPT

1) Participants in the audiovisual (AV) training group are expected to demonstrate greater perceptual gains than the audio-only (A) group as the training progresses. Input is more robust for the AV group and will generate detailed exemplars in memory.

Previous research suggests that enhanced HVPT training (i.e., an audiovisual modality) increases the robustness and variability of the stimuli and yields higher discrimination accuracy of L2 contrasts (Hazan et al., 2005). This will be tested by comparing the two groups' discrimination accuracy scores for the respective contrasts from the training sessions. The AV group is also expected to outperform the A group at post-test.

[15] Thomson (2011) used nautical flags as symbols representing an English vowel category, so that listeners were able to associate sounds to symbols rather than orthographic representations.

Individual longitudinal perceptual development

2) Participants from both groups (audio and audiovisual) are expected to make progress from pre- to post-test.

Previous studies report that HVPT training leads to increased discrimination accuracy of L2 contrasts (Bradlow et al., 1999; Nishi & Kewley- Port, 2007) This hypothesis will be tested by comparing the results at pre- and post-test and tracking training development for each individual participant.

3.2 Methods

3.2.1 L1 German speakers

8 L1 speakers of German (4 female, 4 male) aged 25, 28, 32, 34, 35, 36, 56, and 83, respectively, provided the stimulus materials. While all of them were able to speak Standard German, only three were from Lower Saxony where the local variety is considered Standard German. Three of them were from Bavarian dialect regions, one from Hesse, and one from Saxony. They had no special training in speaking.

3.2.2 Listeners

A total of 46 (28 female, 16 male, one non-binary, one "prefer not to say") participants from a large Western Canadian research university completed the perception pre-test. They were aged between 17 and 49 years. Thirty-seven were L1 speakers of English; the remaining 9 were L1 speakers of Mandarin (n=2), Cantonese (n=2) Vietnamese (n=1), Spanish (n=2), Farsi (n=1), and Serbian (n=1). They had indicated in a questionnaire that they were beginner learners of German and had enrolled in a German first-semester class for the current academic term. Some of the

participants indicated some exposure to German through immediate family or travel, but none had ever actively used or spoken it prior to enrolling in the course. Due to the 8-week commitment, attrition was high and only 12 participants (6 audio-only; 6 audiovisual) completed the experiment (1 male, 11 females; 9 English L1 speakers; 2 Cantonese L1 speakers, 1 Spanish L1 speaker). Coincidentally, the 2 Cantonese L1 speakers and the one Spanish L1 speaker were assigned to the audio-only group, while the audiovisual group only had L1 English speakers. It is important to note, however, that all listeners except one (Cantonese L1) were born in Canada and considered themselves heritage speakers of Cantonese and Spanish, respectively.

3.2.3 Pre-test and post-test materials

German non-words were chosen for the pre- and post-test to assess vowel and consonant discrimination without the influence of lexical knowledge[16] and to minimize learning effects. Moreover, the non-words allowed for a sufficient number of monosyllabic minimal pairs to be formed for each contrast tested in the study. The vowels /iː - eː/, /a - aː/ /oː - øː/, /uː - øː/, /uː - yː/ were embedded within CVC-non-words, the consonants /k - x,ç/ in VC-non-words, and /z - t͡s/ within CV-non-words. In total, a German L1 speaker produced 20 different non-words containing a critical German vowel (10) or consonant (4), resulting in 280 different non-words using the integrated Apple MacBook M2 microphone at 44.1 kHz and the integrated 1080p FaceTime HD camera. Each stimulus was produced three times, resulting in 840 tokens.

[16] Participants were beginning learners of German, but some of them had had exposure to the language, as one parent was German or they had travelled to German speaking countries before. Additionally, the training was administered throughout the academic term, so they had just started to learn German and now had exposure to the languages 3 times a week in 50-minute sessions.

Following the recording procedure, the stimuli were cut and spliced into triads, e.g. **hiem* [hiːm] - **hehm* [heːm] -**hiem* [hiːm] using the software *Filmora Wondershare 12*.

3.2.4 Training materials

The training materials consisted of the same set of monosyllabic non-words of German as the pre- and post-test materials. A total of 8 different German L1 speakers met with the researcher in Germany and produced the same 280 stimuli (20 of the 10 vowels and four consonants, 3 times) as the speaker from pre- and post-test used for the training sessions of each respective group. Speakers were recorded at 44.1 kHz using an integrated Apple MacBook M2 microphone and the integrated 1080p FaceTime HD camera. The stimuli were spliced into video and sound files using video editor *Wondershare Filmora 12*. For both the audio and audiovisual condition, each stimulus consisted of three different non-words of German as produced by three different speakers, thereby adding high intra-triad variability out of which the last word either matched Word 1 or Word 2 (ABX). For example, participants were exposed to Speaker 1 (male, 34 years) producing *guhm,* Speaker 2 (female, 26 years) producing *gühm,* followed by Speaker 3 (female, 36 years) producing *guhm*). This decision to make use of three separate speakers was based on the HVPT employed by Perrachione et al. (2011).[17]

[17] Previous studies employing discrimination tasks in HVPT training do not report whether stimuli consisted of individual triads produced by multiple talkers, thereby adding variability. The literature, suggests, however, that while the voices of multiple speakers can occur in the training stimuli throughout a session, an individual triad (stimulus) is typically produced by a single speaker only.

3.2.5 Procedure

The listeners who had been recruited from six Beginning German classes at the University of Calgary initiated contact with the researcher via e-mail. They had been informed of the nature and duration of the experiment during an information session at the beginning of the academic term. Upon signing up, they received a link to a *Qualtrics* survey with a background questionnaire and an attached consent form. Before participating in the perceptual pre-test, this survey served to assess their language background as well as proficiency level and prior knowledge of German. A total of 46 participants completed the single talker audio-only pre-test consisting of 70 items (10 per phonemic contrast), however, 34 of them did not complete the training or were later excluded because they failed to return to the post-test session. The remaining 12 were assigned to the audio-only (A) or audiovisual (AV) HVPT training group. At the beginning of each week, they received a link to three separate training sessions of approximately 20 minutes each, which they were able to complete sometime during the week, at their convenience, from a quiet location, with the use of headphones recommended. Each session consisted of 10 triads per contrast tested, with a maximum of four contrasts[18] tested per session. Each triad contained the stimulus productions of three different talkers. Vowels and consonants were tested separately. After hearing each triad, the listeners were instructed to choose whether the last stimulus of the triad was the same as the first or the second (ABX discrimination). If their response was correct, they were able to proceed to the next item. If the response was incorrect, they received immediate feedback and heard (and saw in the AV group) a repetition of

[18] Vowels and consonants were always trained separately. Due to the fact that there were five vowel contrasts but only two consonantal contrasts tested, the number of training sessions for the consonants was half of that of vowels, so the stimuli amounted to the same number throughout the six-week training period.

the triad.

At the end of the training phase, each participant completed a post-test, which was equivalent to the pre-test, with a single talker, audio-only condition, to see if the training groups had made perceptual gains over the course of the weeks and if the audio-only group performed differently from the audiovisual group. Two participants were chosen for a $150 prize draw and a book prize in exchange for their participation.

3.3 Data Analysis

One participant from the audio-only (A) training group had partaken in a related HVPT vs. LVPT experiment involving an oddity task with orthographic input before the academic term. The participant's data were included in the experiment because they did not perform differently from their peers at pre- and post-test and training. Participants' responses were coded as trial-wise accuracy (1 = correct; 0 = incorrect) at pre- and post-test as well as for the training sessions. The scores for each participant were grouped by the respective German phonemic contrast. The overall scores for each contrast were averaged for each group at pre-test, training, and post-test.

3.4 Results

3.4.1 Pre- and post-test discrimination accuracy scores

Table 10 summarizes the discrimination accuracy scores (out of 10) for the seven German sound contrasts at pre-test (prior to the 6-week training period) and at post-test (after completion of the training). Pre- and post-test were identical and the same for participants of the audio-only training group and the audiovisual training group.

Table 10 Pre- and post-test scores for the German sound contrasts tested

	/uː - øː/		/oː - øː/		/uː - yː/		/a - aː/		/iː - eː/		/k - x, ç/		/z - t͡s/	
	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post
A-group	9.3	9.83	9.33	9.33	9.66	9.83	8.33	9.16	7.5	8	10	10	8.83	9.16
AV-group	9.66	9	8.5	8.66	9.66	9.66	8.5	7.83	8.5	7.83	9.16	9.83	8.66	9.16
p-value	0.456	0.155	0.064	0.510	1	0.54	0.852	0.147	0.097	0.896	0.34	0.34	0.78	1

In the following section, the pre- and post-test results are presented for each respective contrast by comparing the A-group and AV-group discrimination accuracy scores prior and post HVPT. Then, discrimination accuracy is compared at pre- and post-test within each respective group.

/uː - øː/ contrast

The average score at pre-test ranged between 93% for the A group and 97% for the AV group. An independent samples *t*-test was performed to determine potential differences between the A and AV group. The maximum pre-test score for this contrast was 10.[19] Results showed that there was no significant difference in the scores between participants of the audio-only (*M*= 9.33, SD = 0.5) and the audiovisual condition (*M* = 9.66, SD = 0.5); *t*(10) = -1.11, *p* = .289, which was indicative of the participants in the two groups performing similarly well in the audio-only, single talker condition. At post-test, the audio-only group performed at 98%, whereas the audiovisual group achieved 90% accuracy. An independent samples *t*-test for the post-test scores showed no significant difference between the audio-only (*M* = 9.83, SD = .4) and the audiovisual group (*M* = 9, SD = 1.26); *t*(10) = 1.53, *p* = .155. A within-subjects paired *t*-test in the audio-only group from pre-test (*M* = 9.33, SD .51) to post-test (*M* = 9.83, SD = .4), *t*(5) = -2.23, *p* =

[19] The pre-test served as an indication of how well the seven German phonemic contrasts were discriminated. There were 10 items to discriminate per contrast in this initial test.

.075 yielded no significant differences. Similarly, a paired *t*-test within the audiovisual group from pre-test (M = 9.66, SD = .51) to post-test (M = 9, SD = 1.26), $t(5) = 1.08$, $p = .327$ showed no significant differences.

/oː - øː/ contrast

At pre-test, discrimination of the /oː - øː/ ranged between 85% for the AV group and 93% for the A group. Participants in the audio-only performed at 93% accuracy and 85% in the audiovisual group. An independent samples *t*-test was performed. The difference in scores between the audio-only (M = 9.33, SD = .81) and the audiovisual group (M = 8.5, SD = 1.63); $t(10) = 2.076$, $p = .064$ failed to reach significance. Participants from both groups were thus considered to perform equally well at pre-test. At post-test, the audio-only group performed at 93% again, whereas the audiovisual group performed at 87%. An independent samples *t*-test for post-test scores showed no significant differences between the audio-only (M = 9.33, SD = 1.63) and the audiovisual group (M = 8.66, SD = 1.75); $t(10) = .681$, $p = .51$. A within-subjects paired *t*-test in the audio-only group from pre-test (M = 9.33, SD = .81) to post-test (M = 9.33, SD = 1.63); $t(5) = 0$, $p = 1$ showed no significant differences at post-test. Averaged scores across participants remained identical between the pre- and post-test. A paired *t*-test within the subjects of the audiovisual group at pre-test (M = 8.5, SD = .54) and post-test (M = 8.66, SD = 1.75); $t(5) = -0.222$, $p = .832$ also yielded no significant differences.

/uː - yː/ contrast

The average score at pre-test indicated that discrimination of the German /uː - yː/ contrast was 96% for both groups. As the average scores were identical for the audio-only and audiovisual groups, no independent samples *t*-test had to be performed to determine if there were significant differences. An independent samples *t*-test for the post-test scores showed no significant difference between the audio-only (*M*= 9.83, SD = .4) and the audiovisual group (*M* = 9.66, SD = .51); *t*(10) = .62, *p* = .549. A within-subjects paired t-test in the audio-only group from pre-test (*M* = 9.66, SD = .81) to post-test (*M* = 9.83, SD = .4); *t*(5) = -1, *p* = .363 showed no significant differences. The audiovisual group performed identically (96%) at pre-test and post-test, so no paired *t*-test had to be performed.

/a - aː/ *contrast*

The average discrimination accuracy scores for participants at pre-test ranged between 83% for the A group and 85% for the AV group. An independent samples *t*-test was performed. Results showed that there was no significant difference between participants in the audio-only (*M* = 8.33, SD = 1.36) and the audiovisual group (*M* = 8.5, SD = 1.64); *t*(10) = -.191, *p* = .852. This means that there was initially no difference between the groups in the single-talker, audio-only condition. By post-test, the HVPT-A group achieved a 92% discrimination accuracy, compared to the HVPT-AV group whose performance dropped to 78%. An independent samples *t*-test revealed no significant differences between the HVPT-A group (*M* = 9.16, SD = 1.32) and the HVPT-AV group (*M*= 7.83, SD = 1.6); *t*(10) = 1.568, *p* = .147. A within-subjects paired *t*-test in the HVPT-A group from pre-test (*M* = 8.33, SD = 1.36) to post-test (*M* = 9.16, SD = 1.32), *t*(5) =

-.881, p = .418 yielded no significant difference. In the same vein, the HVPT-AV showed no significant difference from pre-test (M = 8.5, SD = 1.64) to post-test (M = 7.83, SD = 1.6); $t(5)$ = .567, p = .594.

/iː - eː/ contrast

The average discrimination score for the German /iː - eː/ contrast ranged from 75% in the A group to 85% in the AV group at pre-test. An independent samples *t*-test was performed. Results showed that there was no significant difference between the HVPT-A group (M = 7.5, SD = .54) and the HVPT-AV group (M = 8.5, SD = 1.22); $t(10)$ = -1.825, p = 0.097 at pre-test. At post-test, the HVPT-A group had improved to 80%. Accuracy in the HVPT-AV group dropped to 78% discrimination accuracy at post-test. An independent samples *t*-test revealed no significant differences between the HVPT-A (M = 8, SD = 2) and the HVPT-AV group (M = 7.83, SD = 2.31); $t(10)$ = .133, p = .896. A within-subjects paired *t*-test in the HVPT-A group from pre-test (M = 7.5, SD = .54) to post-test (M = 8, SD = 2); $t(5)$ = -.807, p = .456 revealed no significant difference. Similarly, discrimination scores at pre-test (M = 85, SD = 1.22) and post-test (M = 7.83, SD = 2.31); $t(5)$ = .79, p = .465 yielded no significant differences in the HVPT-AV group.

/k - x, ç/ contrast

The average score at pre-test for the contrast of /k - x,ç/ was at 100% in the HVPT-A group and 92% in the HVPT-AV group. An independent samples *t*-test was performed. Results showed there was no significant difference between the participants in HVPT-A (M = 10, SD = 0) and HVPT-AV (M = 9.16, SD = 2.04); $t(10)$ = 1, p = .34, indicating that performance in the groups was considered similar before training. At post-test, the HVPT-A group performed at 100%

again, and the HVPT-AV group had improved to 98%. An independent samples *t*-test for post-test scores yielded no significant difference between the HVPT-A (M = 10, SD = 0) and the HVPT-AV group (M = 9.83, SD = .4); $t(10) = 1$, $p = .34$. A within-subjects paired *t*-test in the HVPT-A group yielded no significant difference from pre-test (M = 10, SD = 0) to post-test (M = 10, SD = 0). The paired *t*-test for the HVPT-AV group also failed to reach significance from pre-test (M = 9.16, SD = 2.04) to post-test (M = 9.83, SD = .4); $t(5) = -.755$, $p = .483$.

/z - t͡s/ contrast

The averaged pre-test scores indicated that discrimination accuracy for the German contrast /z - t͡s/ was 88% for the A-group and 87% for the AV-group. An independent samples *t*-test confirmed that the HVPT-A group (M = 8.83, SD = .98) and the HVPT-AV group (M = 8.66, SD = 1.03); $t(100) = .286$, $p = .78$ were not significantly different from each other at this stage.

At post-test, both groups achieved an accuracy score of 92%, indicating slight improvement from the pre-test. The groups performed similarly and were thus not significantly different from each other. A within-subjects paired *t*-test was performed for the HVPT-A group from pre-test (M = 8.83, SD = .98) to post-test (M = 9.16, SD = 1.16); $t(5) = -1.581$, $p = .174$ and yielded no significant differences. Similarly, the paired samples *t*-test for the HVPT-AV group showed no significant difference between the pre-test (M = 8.66, SD = 1.03) and post-test condition (M = 9.16, SD = .98); $t(5) = -.888$, $p = .414$.

Summarizing the pre- and post-test results, it is evident that participants from both groups performed similarly at pre- and post-test for the respective sound contrasts. Some contrasts in particular yielded ceiling performance in terms of discrimination accuracy scores at pre-test, which were replicated by the A-group at post-test. While the A-group improved their

discrimination accuracy from pre- to post-test for five of the seven contrasts (/uː - øː/, /uː - y ː/; /a - a ː/, /i ː - e ː/, and /z - t͡s/), the AV-group's discrimination accuracy deteriorated from pre- to post-test for the contrasts /uː - øː/, /a - aː/, and /iː - eː/. However, both of these observed trends were not significant.

3.4.2 Audio-only vs. audiovisual HVPT

The hypothesis guiding the current study was that participants in both groups would experience improvements in perceptual discrimination of the seven tested contrasts within their respective HVPT modalities during the six-week training period. The subsequent presentation of the results delves into the averaged discrimination accuracy for each sound contrast, comparing the A group with the AV group across ten HVPT sessions with the p-value and effect sizes reported for each contrast. Additionally, an analysis of the individual participants' progress within the A and AV groups for each specific contrast over the six-week training period is conducted, with a focus on identifying and highlighting developmental trends.

/uː - øː/ contrast

Table 11 shows the averaged discrimination accuracy scores for the German contrast /uː - øː/ for the audio-only and the audiovisual group throughout the training sessions.

Table 11 */u: - ø:/ training scores*

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Group A	8.83	9.16	8	7.33	8	7.8	9	9	9.16	8.66
Group AV	7.83	8.33	6.4	6.2	7	7	6.83	7	8.5	7.66
p-value					0.000†					
Effect size					0.72					

Note.† = this p-value refers to the accuracy scores of this contrast achieved by the A- vs. the AV-group over the course of the training sessions.

The mean percentage of correct discrimination from each 10 training sessions was calculated and compared between subjects of the audio-only and the audiovisual group. The audio-only group achieved an average of 86% over the course of six weeks, whereas the audiovisual group achieved a score of 73%. An independent samples *t*-test yielded significant differences between the audio-only (M = 8.67; SD = 1.85) and the audiovisual group (M = 7.32. SD = 1.78); $t(112)$ = 3.85, p = <.001. A Cohen's *d* test was conducted to measure an effect size of .72 (medium effect) between the two training groups.

While the average training scores achieved by the two respective groups served as the basis of comparison, the longitudinal development for each contrast tested was tracked throughout the training phase.

Variability in longitudinal development of the /uː - øː/ contrast

Figure 5[20] shows the HVPT-A group data for the German /uː - øː/ contrast. P2A (L1 = Cantonese) and P3A (L1 = Cantonese) performed consistently between 90% and 100% discrimination accuracy for the /uː - øː/ contrast throughout the 10 training sessions. Additionally, P6A achieved between 90% and 100% discrimination accuracy, with the exception of T2 when the accuracy score was 60%. The remaining participants' scores in the HVPT-A group exhibited substantial fluctuations throughout the 6-week training period. Discrimination accuracy between /uː - øː/ ranged from 30% - 100%, with P1A, P4A, and P5A achieving lower discrimination accuracy scores mid training (T4) from which they seemed to recover by T9. Especially P4A consistently improved starting from T4, where they performed at chance for this contrast. By T9 they achieved 100% discrimination accuracy. It is important to note that the consistently high discrimination accuracy scores for P2A and P3A could be attributed to the fact that Cantonese has front rounded vowels.

Figure 6 shows the HVPT-AV group data for the German /uː - øː/ contrast. While P9AV occasionally reached ceiling discrimination accuracy scores at T1, T2, T5, and T6, participants in this group were rather consistent in discrimination accuracy throughout the training. Accuracy scores typically ranged between 40% and 90%. Reaching the end of the training period, P7AV and P11AV seemed to have made some improvement. P7AV had achieved accuracy scores between 40 % and 50% mid-training, but reached 70% discrimination accuracy at T10. Similarly, P11AV achieved between 40% and 60% discrimination accuracy throughout T3-T8,

[20] Interrupted lines indicate that a participant missed individual training sessions of the respective contrast.

but achieved 90% for this contrast at T10. P8AV stood out in that they never exceeded the 50% discrimination accuracy mark throughout the training.

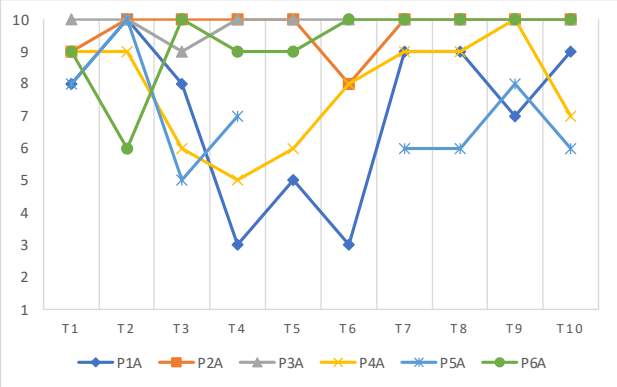


Figure 5 /u: - ø:/ contrast in the audio (A) group

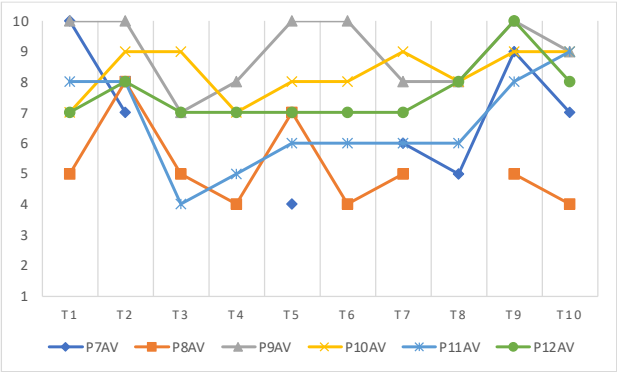


Figure 6 /u: - ø/ contrast in the audiovisual (AV) group

/oː - øː/ contrast

Table 12 summarizes the averaged discrimination accuracy of German /oː - øː/ in the audio-only and audiovisual group throughout the training sessions.

Table 12 */u: - ø:/ training scores*

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Group A	8	8.16	9	8.5	8.83	9.8	8.83	9	9.66	8.83
Group AV	7.5	7.16	8.5	7.66	6.8	7.5	6.16	8.8	7.83	8.33
p-value					0.001†					
effect size					0.59					

Note.† = this p-value refers to the accuracy scores of this contrast achieved by the A- vs. the AV-group over the course of the training sessions.

The mean percentage of correct discrimination from each of the 10 training sessions was computed and subjects in the audio-only and audiovisual groups were compared. Participants in the audio-only group achieved an accuracy of discrimination score of 88%, whereas participants in the audiovisual condition achieved an average accuracy score of 76% throughout training. An independent samples *t*-test showed significant differences between the audio-only (M = 8.84, SD = 1.51) and the audiovisual group (M = 7.62, SD = 2.48), $t(115) = 3.22$, $p = .001$. A Cohen's *d* test was conducted to measure an effect size of .59 (small effect).

Variability in longitudinal development of the /oː - øː/ contrast

Figure 7 and *Figure 8* show the HVPT-A and the HVPT-AV group data for the German /oː - øː/ contrast. P2A (L1 = Cantonese), P3A (L1 = Cantonese), and P6A's training performance stands out as their discrimination accuracy is consistently between 80-100%. Starting from T3, all three participants achieved 100% discrimination accuracy for this contrast. Like for the vowel contrast

/uː - øː/, P2A's and P3A's high discrimination accuracy for the /oː - øː/ contrast could be attributed to the fact that Cantonese has front rounded vowels. P4A and P5A reached discrimination accuracy scores between 70% to 100% throughout the ten training sessions, with P4A consistently performing at 100% starting from T6. P5A's discrimination accuracy was at chance in the initial training session and fluctuated between 50% and 70% throughout the training phase with a one-time maximum of 80% achieved at T9.

Conversely, participants in the AV group exhibited substantial fluctuation in discrimination accuracy of the /oː - øː/ contrast. While P9AV, P10AV and P12AV showed the highest consistency throughout the ten training sessions in that their discrimination accuracy ranged between 70% to 100%, the remaining participants P7AV, P8AV, and P11AV not only exhibited overall lower discrimination accuracy (between 10% and 60% for P7AV and between 20% and 80% for P8AV), but their discrimination accuracy was also subject to large fluctuations throughout the training. By T9, P11AV's discrimination accuracy seemed to stabilize and they performed at 100% in the final training sessions, whereas P7AV and P8AV were unable to improve from the beginning to the end of the training phase.

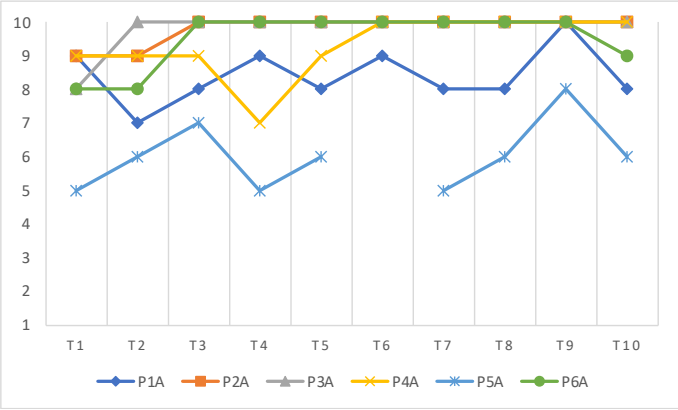


Figure 7 /o: - ø:/ contrast in the audio (A) group

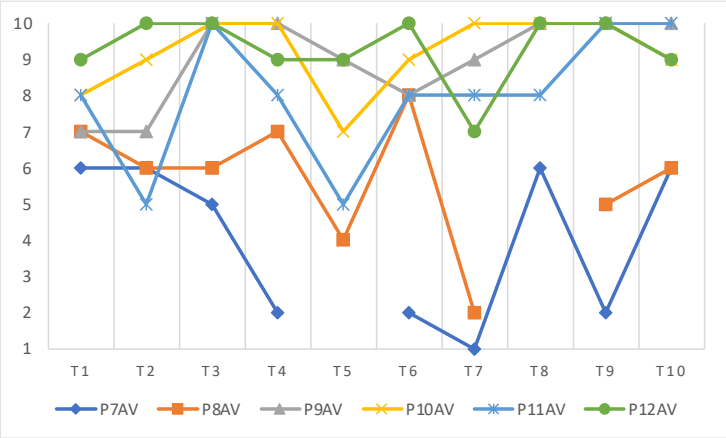


Figure 8 */o: - ø:/ contrast in the audiovisual (AV) group*

/uː - yː/ contrast

Table 10 summarizes the averaged discrimination accuracy of German /uː - yː/ in the audio-only and audiovisual group throughout the training period.

Table 13 */u: - y:/ training scores*

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Group A	8.33	8.66	9	8.83	9.16	8.6	9.16	8.83	8.83	8.83
Group AV	7.16	8.16	6.83	7.2	8.8	8.2	8	7.2	6.83	8.16
p-value					0.000†					
effect size					0.69					

Note.† = this p-value refers to the accuracy scores of this contrast achieved by the A- vs. the AV-group over the course of the training sessions.

The mean percentage of correct discrimination from each of the 10 training sessions was computed and compared between the audio-only and the audiovisual group. Participants in the audio-only group received an average discrimination accuracy of 88% in the training sessions, whereas those in the audiovisual group achieved 76% discrimination accuracy. An independent sample *t*-test showed significant differences between the audio-only (M = 8.83, SD = 1.31) and the audiovisual group (M= 7.64, SD = 2.02); $t(113) = 3.75$, $p = <.001$. A Cohen's *d* test measured an effect size of .69 (medium effect).

Variability in longitudinal development of the /uː - yː/ contrast

Figure 9 and *Figure 10* show the HVPT-A and the HVPT-AV group data for the German /uː - yː/ contrast. In line with the observations made for the two previous contrasts involving front rounded vowels, participants P2A, P3A (L1 Cantonese) and P6A achieved between 80% to 100% discrimination accuracy throughout the training phase for this contrast, with the overwhelming majority of training sessions completed by the three participants yielding discrimination accuracy scores of 100%. Participants P1A, P4A, and P5A achieved discrimination accuracy scores ranging between 60% to 100%. While P1A and P5A did not seem to gradually improve from T1 to T10, P4A exhibited the highest consistency throughout the training phase, achieving an accuracy score of 100% in the final training session (T10).

Conversely, participants in the HVPT-AV group exhibited larger fluctuation in discrimination accuracy of the German /uː - yː/ contrast. While only participants P10AV and P12AV achieved 100% discrimination accuracy for a total of four and three sessions, respectively, accuracy scores were subject to substantial fluctuations throughout the training phase. Notably, P10AV started with 100% discrimination accuracy at T1 and consistently

declined to 60% at T4 before recovering to 100% at T5 where the participant remained relatively stable throughout the remainder of the training. P9AV exhibited the highest internal consistency throughout the training phase with the lowest accuracy score of 80% occurring at T3, T8 and T10, but ceiling results in all other training sessions. Similarly, P12AV consistently performed between 90% and 100% starting from T4. In contrast, participants P7AV, P8AV, and P11AV achieved accuracy scores ranging between 30% to 90% throughout the training and thus exhibited large fluctuations throughout with P7AV and P11AV experiencing an upward trend from T9.

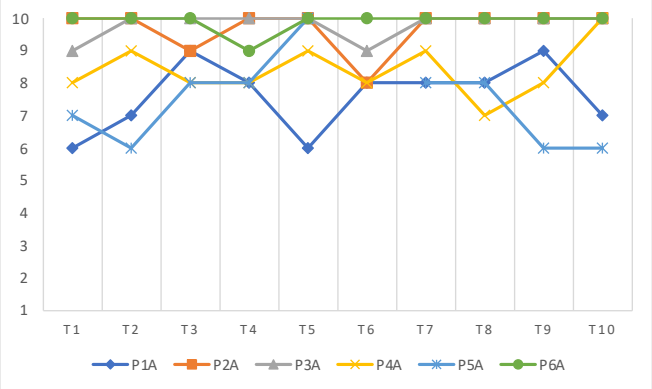


Figure 9 */u: - y:/ contrast in the audio (A) group*

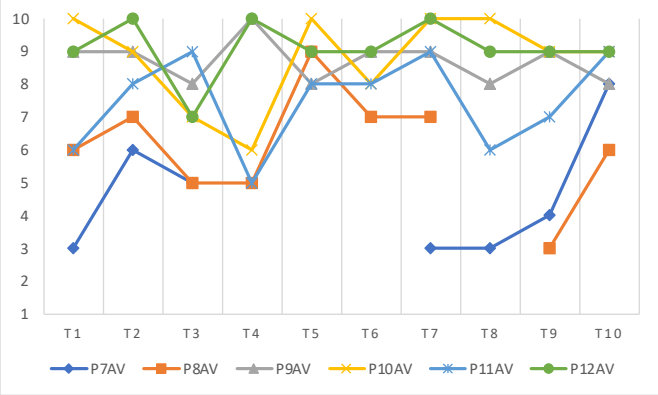


Figure 10 */u: - y:/ contrast in the audiovisual (AV) group*

/a - aː/ *contrast*

Table 11 summarizes the average discrimination accuracy of the German /a – aː/ contrast in the HVPT-A and HVPT-AV group throughout the training.

Table 14 /a - aː/ *training scores*

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Group A	8.16	8	8.16	8.5	6.66	8.6	8.8	8.66	6.66	8.83
Group AV	7.5	7	8.33	6.8	6.2	7.5	7.8	8	6	7.5
p-value					0.005†					
effect size					0.53					

Note.† = this p-value refers to the accuracy scores of this contrast achieved by the A- vs. the AV-group over the course of the training sessions.

The mean of discrimination accuracy for this contrast in the training was computed and compared between subjects of the HVPT-A and HVPT-AV groups. The HVPT-A group achieved an average of 81% throughout the training, whereas the HVPT-AV group achieved an average score of 73%. An independent samples *t*-test showed significant differences between the HVPT-A (M = 8.08, SD = 1.3) and the HVPT-AV group (M = 7.28, SD = 1.69); $t(113) = 2.857$, $p = .005$. Cohen's d tests computed an effect size of .53, which is a medium effect between the two groups.

Variability in longitudinal development of the /a - aː/ *contrast*

Figure 11 and *Figure 12* show the HVPT-A and the HVPT-AV group data for the German /a - aː/ contrast. This contrast did not involve front rounded vowels and substantial fluctuation in discrimination accuracy can be observed among all participants of the HVPT-A group. Participant P2A and P4A exhibited the largest internal consistency in that their discrimination accuracy for this contrast ranged between 70% and 100% throughout the training phase. Participants P1A and P3A exhibited the largest fluctuations throughout the training with P3A achieving accuracy scores between 50% at T5 and 100% at T7, and P1A starting out with 90% discrimination accuracy at T1 and experiencing a decline to 50% at T5 and 40% at T9. Notably, while all participants achieved discrimination accuracy scores ranging from 80% to 100% at T10, no consistent upward trend could be observed. Participant P6A started out with discrimination accuracy scores of 90% and higher from T1 to T4, but declined to 60% at T5 and was unable to achieve 90% accuracy again until the final training session (T10).

Similarly, the HVPT-AV group also experienced substantial fluctuations in discrimination accuracy for the /a - aː/ contrast throughout the ten training sessions. While 100% discrimination accuracy was only achieved by participants P10AV at T8, P9AV at T7, T8, and T10 and P12AV at T6 and T8, this contrast yielded overall lower discrimination accuracy with P8AV consistently declining from an initial 70% at T1 to 30% at T9. At no point during the training did participants P7AV, P8AV, or P11AV exceed the 80% discrimination accuracy mark. For P9AV an upward trend could be observed starting from T4. They consistently performed at 90% or higher from T6.

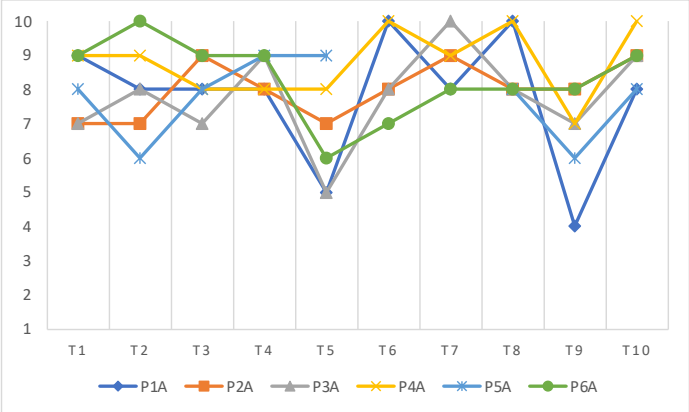


Figure 11 /a - a:/ *contrast in the audio (A) group*

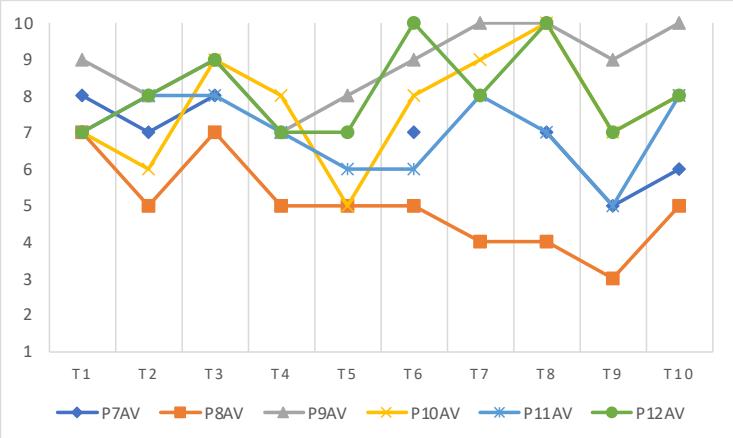


Figure 12 /a - a:/ contrast in the audiovisual (AV) group

/iː - eː/ contrast

Table 15 summarizes the average discrimination accuracy of the German /iː - eː/ contrast in the HVPT-A and HVPT-AV group throughout the training.

Table 15 */i: - e:/ training scores*

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Group A	6.16	5.5	5.5	6.66	5.66	5.6	6.4	8.5	7.83	7.16
Group AV	6.33	5.33	7.5	7.16	5.6	5.66	7.2	7	7.66	8
p-value					0.509†					
effect size										

Note.† = this p-value refers to the accuracy scores of this contrast achieved by the A- vs. the AV-group over the course of the training sessions.

The mean accuracy score throughout the training was computed and the HVPT-A group was compared to the HVPT-AV group. HVPT-A achieved 65% discrimination accuracy, and HVPT-AV 67% over the course of the six-week training period. An independent samples *t*-test yielded no significant differences between the HVPT-A (M = 6.51, SD = 1.77) and HVPT-AV (M = 6.75, SD = 2.13); $t(114) = -.662$, $p = .50$ in the training condition.

Variability in longitudinal development of the /iː - eː/ contrast

Figure 13 and *Figure 14* show the HVPT-A and the HVPT-AV group data for the German /iː - eː/ contrast. Overall, this vowel contrast proved challenging to participants in the HVPT-A group. All participants started out with discrimination accuracy scores ranging between 40% to 80% at T1. While only P2A and P3A were able to achieve 100% discrimination accuracy for this contrast at T4 and T5, and at T9, respectively, they consistently improved by T4, where P2A

achieved 100% and P3A 90% discrimination accuracy. Following a decline at T5, both seem to recover by T8 and achieve accuracy scores between 80% and 100% at T10. The data for the remaining participants suggest that all experience a substantial decline in discrimination accuracy at T3 or T4 from which they recover in the following session. P3A, P4A, P6A follow a similar pattern in that they consistently improve from T6 and reach peak performance by T8 but are unable to retain the scores for the remainder of the training phase.

Like the HVPT-A group, the HVPT-AV group experienced large internal fluctuations in discrimination accuracy for the /iː - eː/ throughout the training phase. As such, the discrepancies in accuracy were quite large for especially the participants P7AV (20% to 80%) and P11AV (40% to 100%). P9AV exhibited the highest discrimination accuracy throughout the training with discrimination accuracy being at 80% at T1 and consistent discrimination accuracy ranging between 90% and 100% between T3 and T10. Notably, while no clear overall trend towards improvement could be observed for the participants in the AV group for the /iː - eː/ contrast, participant P11AV showed a consistent upward trend after reaching a discrimination accuracy low of 40% at T6. Overall, the results suggest that this contrast was difficult to discriminate in the respective HVPT modality.

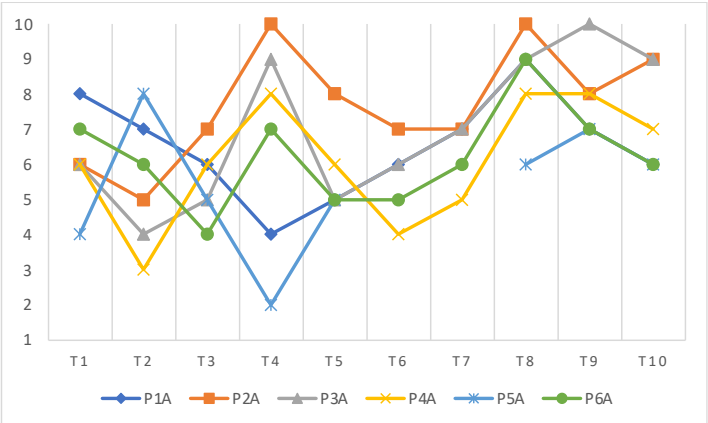


Figure 13 /i: - e:/ contrast in the audio (A) group

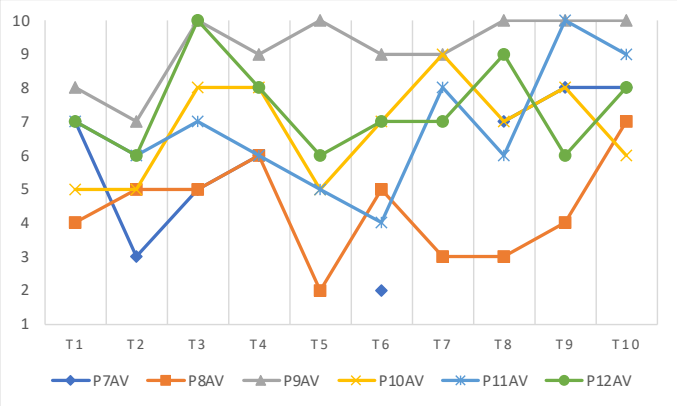


Figure 14 /i: - e:/ contrast in the audiovisual (AV) group

/k - x, ç/ contrast

Table 16 summarizes the average discrimination accuracy of the German /k - x, ç/: contrast in the HVPT-A and HVPT-AV group throughout the training.

Table 16 /k - x,ç/ training scores

	T1	T2	T3	T4	T5
Group A	8.5	8.66	9.16	9.6	9.66
Group AV	8.66	8.5	8.4	9.4	8.33
p-value			0.213†		
effect size					

Note.† = this p-value refers to the accuracy scores of this contrast achieved by the A- vs. the AV-group over the course of the training sessions. [21]

[21] Due to the fact that there were five vowel contrasts and only two consonantal contrasts trained in the current study, the consonantal contrast /k - x, ç/ only appeared in five training sessions. This means that there was an equal number of triads for each trained contrast. The other consonantal contrast, /z - t͡s/, was trained in the other five sessions.

The mean discrimination accuracy from the five training sessions was computed and compared between the HVPT-A and HVPT-AV group. The HVPT-A group achieved an average score of 91%, whereas the HVPT-AV group achieved discrimination accuracy of 86%. An independent samples *t*-test yielded no significant difference between the HVPT-A (M = 9.1, SD = .77) and the HVPT-AV group (M = 8.64, SD = 1.8); $t(55)$, = 11.257, p = .213.

Variability in longitudinal development of the /k - x, ç/ contrast

Figure 15 and *Figure 16* show the HVPT-A and the HVPT-AV group data for the German /k - x, ç/ contrast. Overall, participants in the HVPT-A group achieved high discrimination accuracy for this contrast, ranging between 80% and 100%. Participants P4A, P5A, and P6A were able to consistently improve by T3, with all three participants achieving 100% discrimination accuracy by T4. P4A and P5A were able to make the largest gains from 80% at T1 to 100% at T5.

Participants in the HVPT-AV group exhibited a similar discrimination accuracy pattern throughout the training with P9AV, P10AV, and P12AV consistently performing between 90% and 100% throughout the training phase, and P11AV improving from 70% discrimination accuracy at T1 to 100% at T4. P8AV exhibited the largest internal fluctuations with an 80% accuracy score at T1 and declining to 30% at T3. At T5, accuracy scores ranged between 90% and 100% for the P9AV, P10AV, P11AV, and P12AV, while P8AV performed at chance in the final training session.

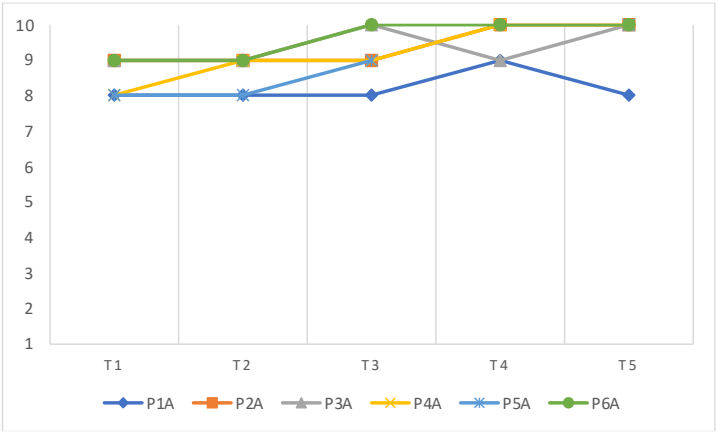


Figure 15 /k - x,ç/ contrast in the audio (A) group

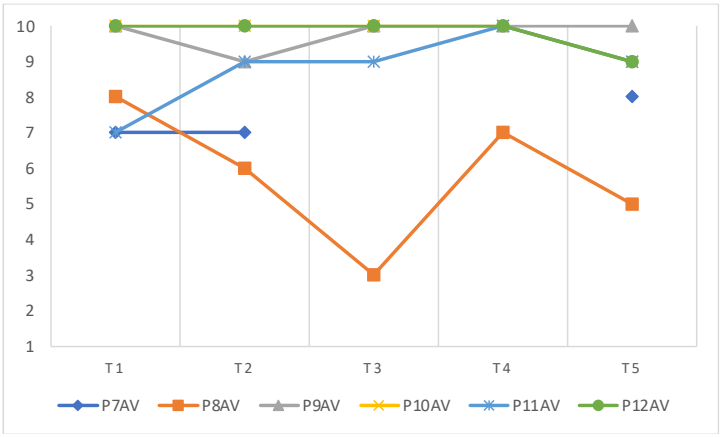


Figure 16 /k - x,ç/ contrast in the audiovisual (AV) group

/z - t͡s/ contrast

Table 17 summarizes the average discrimination accuracy of the German /z - t͡s/ contrast in the HVPT-A and HVPT-AV group throughout the training.

Table 17 */z - ts/ training scores*

	T1	T2	T3	T4	T5
Group A	8.16	8.66	8.5	8.6	9.5
Group AV	7.66	8.16	8	7.4	7.83
p-value		0.044†			
effect size		0.54			

Note.† = this p-value refers to the accuracy scores of this contrast achieved by the A- vs. the AV-group over the course of the training sessions.

The average percentage for discrimination accuracy in the five training sessions was 87% for the HVPT-A group and 78% for the HVPT-AV group. An independent samples *t*-test showed that the differences were significant between the HVPT-A (M = 8.68, SD = 1) and the HVPT-AV group (M = 7.82, SD = 2.03); $t(55) = 2.05$, $p = .044$. A Cohen's *d* test was conducted to measure an effect size of .54, which was a medium effect.

Variability in longitudinal development of the /z - t͡s/ contrast

Figure 17 and *Figure 18* show the HVPT-A and the HVPT-AV group data for the German /z - t͡s/ contrast. Generally, discrimination accuracy for this contrast was high and ranged between 60% and 100% in the HVPT-A group. In particular, participant P6A produced the most consistent accuracy scores throughout the training phase starting at 90% discrimination accuracy at T1 and reaching 100% by T5. P2A achieved accuracy scores between 80% to 100% throughout the training phase. An upward trend in discrimination accuracy for the /z - t͡s/ contrast could be observed in participant P4A who had achieved 80% discrimination accuracy in the

initial training session and who achieved 90% at T3 and T4 before reaching 100% in the final training session.

Similarly, participants P10AV, P11AV, and P12AV in the HVPT-AV group achieved discrimination accuracy scores of 80% to 100% throughout the training phase. While all three participants experienced a decline in discrimination accuracy of 10% after T3, all of them achieved 90% discrimination accuracy scores by the final training session. P7AV stood out because they achieved 60% discrimination accuracy at T1 and declined to 20% at T3 and were unable to exceed the 40% mark in the subsequent training sessions.

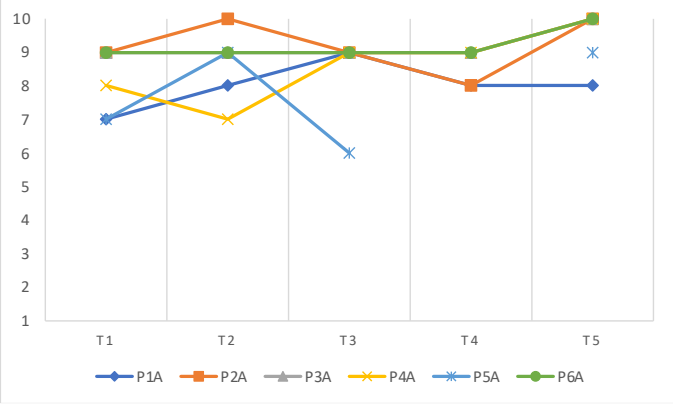


Figure 17 /z - ts/ contrast in the audio (A) group

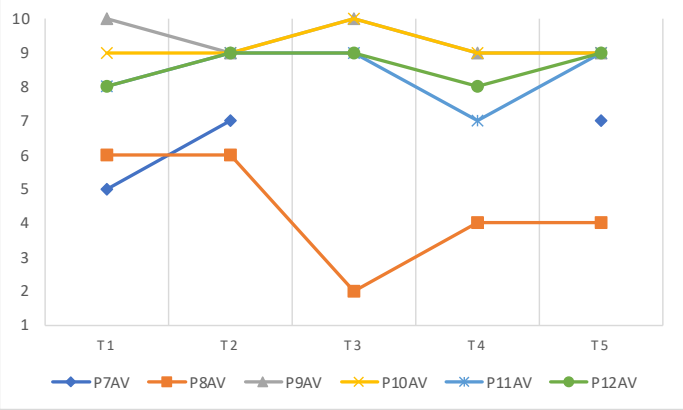


Figure 18 /z - ts/ contrast in the audiovisual (AV) group

3.5 Discussion

3.5.1 Pre- and post-test

The findings indicate no statistically significant improvement of the trained German vowels and consonants from pre-test to post-test. On the one hand, participants in both groups performed at ceiling for the contrasts: /uː - øː/ (93% A vs. 96% AV), /uː - yː/ (96% for both groups), /k - x,ç/ (100% A vs. 91% AV) at pre-test. In contrast, the /iː - eː/ contrast had the lowest pre-test accuracy for both the A and AV group (75% A vs. 85% AV).

At post-test, the HVPT-A group replicated the pre-test scores for the /oː - øː/ (93%) and /k - x,ç/ (100%) contrast. Additionally, discrimination for the A group improved from pre- to post-test for /uː - øː/ (93% vs. 98% [+4%]), /uː - yː/ (96% vs. 98% [+2%]), /a - aː/ (83% vs. 91% [+8%]), /iː - eː/ (75% vs. 80% [+5%]), and /z - t͡s/ (88% vs. 91% [+3%]). Notably, for the contrasts /a - aː/ and /iː - eː/, for which discrimination accuracy was the lowest out of all contrasts at pre-test, the post-test results indicated the highest gains in discrimination accuracy (8% and 4%, respectively).

Conversely, the HVPT-AV group improved numerically from pre- to post-test for the contrast /oː - øː/ (85% vs. 86% [+1%]), /k - x, ç/ (91% vs. 98% (+7%), and /z - t͡s/ (86% vs. 91% [+5%]) only. Notably, discrimination accuracy in the AV group declined from pre-test to post-test for the contrasts /uː - øː/ (97% vs. 90% [-7%]), /a - aː/ (85% vs. 78% [-7%]), and /iː - eː/ (85% vs.78% [-7%]). Especially the contrasts /a - aː/ and /iː - eː/, which showed the lowest discrimination accuracy of all contrasts at pre-test and therefore left room for improvement, did not improve at post-test.

All observed within- and between-group differences from pre- to post-test were statistically insignificant. While it is impossible to determine whether HVPT and its respective A- or AV-modality had any impact on discrimination accuracy in the low variability post-test, the results indicate a clear difference between the HVPT-A and the HVPT-AV group for the contrasts /a - aː/ and /iː - eː/. The HVPT-A group was able to make the largest gains from pre- to post-test, whereas the HVPT-AV group achieved lower discrimination accuracy scores for both contrasts at post-test (-7%).

The results obtained from pre- and post-test served as an independent test to measure discrimination accuracy prior to the training and at the end of the training. However, training discrimination accuracy scores from the respective HVPT modality could not be linked to pre- and post-test scores, as these constituted a different testing approach with the pre- and post-test stimuli coming from a single speaker in an audio-only condition.

3.5.2 Training

Comparing the discrimination accuracy scores between the A- and AV-group for each respective contrast showed that the A-group consistently outperformed the AV-group for all seven tested contrasts. The averaged training scores yielded significant differences for five contrasts: /uː - øː/, /oː - øː/, /uː - yː/, /a - aː/, and /z - t͡s/. The higher discrimination accuracy scores in the A group for the /iː - eː/ contrast and the /k - x,ç/ contrast were insignificant. The training scores obtained from participants of the A and AV group throughout the training sessions suggest that discrimination accuracy fluctuated throughout the training sessions. As such, all contrasts showed substantial fluctuations in both the A and AV group throughout the training, with no clear upward trend emerging between T1 and T10. For example, ceiling

accuracy scores for some contrasts were often achieved in earlier training sessions only to decline again in later training sessions. Notably, however, the /iː - eː/ contrast, which had yielded particularly low accuracy in the initial training sessions for participants in both groups, exhibited a consistent upward trend towards T10. Additionally, participants from each group also demonstrated substantial internal inconsistency throughout the six-week training period as well. This fluctuation between T1 and T10 in individual participants and the lack of a clear upward trend towards the end of the training could point to HVPT being overwhelming to beginner-level L2 learners.

This general finding is in line with Perrachione et al. (2011), where it has been suggested that individuals with weaker perceptual abilities may benefit from reduced variability in speech stimuli. This is attributed to beginner learners taking advantage of processes that adapt to the consistent, predictable features of a specific speaker's phonetics in the initial stages of L2 learning (Mullennix, 1997). This consistency of the acoustic-phonetic speech signal is no longer present in highly variable speech stimuli. Additional cognitive resources are required, which can reduce discrimination accuracy by increasing processing costs (Ben-Artzi & Marks, 1995; Mullennix & Pisoni, 1990). For L2 sound perception, a contrast that is phonemic in the L2 but not in the L1 (e.g., /uː - yː/) will exhibit acoustic-phonetic ambiguity leading to potential conflation of the two German phonemes. This ambiguity is further enhanced when the speech signal is highly variable (i.e., coming from different speakers who produce sounds with considerable variation). Therefore, it is conceivable that especially new learners will have difficulty with high variability of L2 speech sounds.

This is why it is perhaps unsurprising that the current study found significant differences in discrimination accuracy between the two HVPT modalities (A and AV) from the training for

five of the contrasts tested for which the HVPT-A group consistently achieved numerically higher accuracy scores than the HVPT-AV group. The only contrasts for which the difference was insignificant were the contrasts /iː - eː/, /k - x, ç/, and /z - /t͡s/. While it is important to consider that the Cantonese L1 speakers in the HVPT-A group might have corrected upward the overall discrimination accuracy average for the front rounded vowels, thereby possibly inducing significant training differences between the two modalities, the HVPT-AV group seemed to have an additional disadvantage when compared to the HVPT-A group. Adding to the argument of HVPT adding cognitive load, the AV modality had even more ambiguity added to the speech stimuli through visual footage than the HVPT-A group. As such, it required additional multimodal processing resources to recognize the speech sounds, thereby leading to reduced accuracy (Acha, 2009; Boers, Warren, and Deconinck, 2017).

The longitudinal training scores from each of the 12 participants are also indicative of perceptual abilities being highly individual, as listeners can vary in their perceptual orientation to acoustic details during phoneme perception (Yu & Zellou, 2019). L2 learning is susceptible to a great deal of individual variability, like motivation (Jiao, Jin, You, & Wang, 2022) affect (Gkonou, Daubney, & Deaele, 2017), and aptitude (Granena & Long, 2013), all of which could not be studied within the scope of this study. Since the L1 phonological makeup also impacts the degree of phonemic discrimination in L2 (Kartushina & Frauenfelder, 2014), data from the Cantonese L1 speakers was treated separately, as they performed at ceiling for the contrasts involving front rounded vowels /uː - yː/, /oː - øː/, and /uː - øː/ throughout the training. Additionally, individual participants have different perceptual abilities. Findings regarding the longitudinal development for each participant from both groups yield insight into the state of the perceptual development of their L2 German phonological categories and point to listener-

specific individual differences. For example, P8AV performed consistently worse than their peers in discrimination accuracy across all contrasts, whereas P12AV achieved high discrimination accuracy for all contrasts throughout the training. A few participants showed fewer internal fluctuations than others (e.g., P4A, P6A) and produced more stable discrimination accuracy scores throughout the training, while some exhibited a slight upward trend emerging towards the end of the training (e.g., P9AV). Additionally, some participants had higher discrimination accuracy for some contrasts than others, e.g., P7AV generally showed lower discrimination performance for the consonantal contrasts, but achieved higher discrimination accuracy than some of their peers for the vowel contrast /a - aː/.

According to *Prediction 1,* it was hypothesized that the HVPT-AV group would outperform the HVPT-A group because input was enhanced through gestural information and additional visual cues to form robust exemplars (Hazan et al., 2005). While it had been assumed for the current study that seeing speakers' faces would create more robust exemplars in listener memory and yield higher perceptual discrimination accuracy, this prediction was not borne out. In particular, distinctive visible articulatory features, such as lip movement, were absent from the speakers' productions of these contrasts. This is unlike the Hazan et al. (2005), where listeners showed significantly better discrimination accuracy when the articulatory gesture was visually distinct, such as in /b/ and /v/. In the current study, none of the contrasts tested was visually distinct to the listener so that they would have been able to obtain additional contrastive information from lip movement or tongue position. The opposite was the case in that adding visual information by showing the speakers' faces seemed to have an adverse effect because listeners had to process the stimuli bimodally, which added cognitive load and decreased discrimination accuracy. The HVPT-A group, which had less variability because there was no

access to facial or gestural information, consistently outperformed the HVPT-AV group throughout the training. This finding is in line with that of Hazan et al. (2005), who found that HVPT improved discrimination accuracy in Japanese listeners of the English consonantal contrasts /b - v/ and /r - l/, where the articulatory gestures are distinctive, but that there was no advantage of the AV modality when gestural distinctiveness was not visible to the listener (i.e., the /r - l/ contrast). As such, *Prediction 1* could not be confirmed.

The findings only partially confirmed *Prediction 2* in that differences from pre- to post-test for both groups were insignificant in all cases. This could largely be attributed to two factors: 1. The pre-test scores obtained were at 90% or higher for four of the tested contrasts, so improvement from pre- to post-test was hardly possible. 2. The pre- and post-test condition was fundamentally different from the training condition. Specifically, participants from both groups did not seem to experience much difficulty with the discrimination of ABX items produced by a single speaker in an audio-only condition, and ceiling effect could be observed. Due to this inconsistency between the pre- and post-test stimuli (single talker, audio-only) and the HVPT training stimuli (multiple speakers, audiovisual), it was impossible to attribute any insignificant improvement in the A-group and insignificant decrease in the AV-group from pre- to post-test to the training.

While discrimination accuracy results cannot be directly compared to those results obtained from pre- and post-test, the overall high/ceiling discrimination accuracy achieved by true beginner learners in their first week of learning German and the substantial fluctuations and lower accuracy scores in the training (especially in the HVPT-AV group) are indicative of listeners working to process the extreme variability and ambiguity of speech stimuli in the beginning stages of L2 learning. The high pre- and post-test scores may point to beginner

learners benefitting from low-variability stimuli, which allows for a more focused and consistent exploitation of phonetic information from a speech signal while L2 phonological categories are still in the making. Evidence from previous studies on low-aptitude learners (Perrachione et al., 2011) and child learners (Giannakopoulou et al., 2017) suggests HVPT is not by default the superior technique when it comes to training L2 perceptual discrimination. As such, Perrachione et al. (2011) found that only some individuals with strong perceptual abilities might benefit from HVPT. In line with this, it has been found that listeners with lower perceptual abilities may be hindered by the effects of high variability (Sadakata & McQueen, 2014). While these studies discuss the benefits of high vs. low variability in the context of learner aptitude, Giannakopoulou et al. (2017) found that child learners benefit from a single talker condition in discriminating the English /iː - ɪ/ contrast and exhibited a reverse effect of HVPT hindering perceptual learning. While the current study did not measure language aptitude or employ child learners, true beginner learners might constitute a group that might benefit from low-variability input.

These findings warrant a need to not only compare and contrast the benefits of both LVPT and HVPT to firmly establish whether HVPT is truly superior to the LVPT, but to focus on the respective technique within different proficiency groups. The current literature on HVPT suggests that the majority of beneficial training effects were found in highly advanced learners who already have experience with the respective L2 sound contrasts (Nishi & Kewley-Port, 2007; Lengeris & Hazan, 2010; Lim & Holt, 2011), and that those studies that included inexperienced learners (Bradlow et al., 1999; Pruitt, Jenkins, & Strange, 2006; Iverson, Pinet & Evans, 2012) compared them to experienced learners without comparing the effects of HVPT and LVPT.

3.6 Conclusion

The present study aimed to investigate the effect of HVPT on the discrimination of novel German sound contrasts in beginner learners of German. The 12 participants were divided into an audio-only (HVPT-A) and an audiovisual training (HVPT-AV) modality group, respectively. The findings revealed that the participants in the A-group significantly outperformed those in the AV group throughout six weeks of training for five out of the seven German sound contrasts tested: /uː - ø/; /oː - øː/; /uː - yː/; /a - aː/, and /z - t͡s/.

These training differences could be attributed to several factors. As pointed out in the discussion, some individuals in the HVPT-A group achieved particularly high training discrimination accuracy scores for the contrasts of /uː - ø/; /oː - øː/; /uː - yː/. One possible explanation may be L1 background. P3A and P6A, both of whom speak L1 Cantonese, a language that has rounded front vowels, demonstrated excellent discrimination of these segments. High perceptual aptitude may provide further insights. P2A, whose L1 is English, outperformed many other participants. Taken together, these three participants may have had an impact on the overall discrimination scores for the A-group. Another possibility is that the presence of visual information in the HVPT-AV group might have added too much variability by requiring listeners to divide their attention between recognizing differences in speech sounds while performing another task of processing visual input (Mattys & Wiget, 2011). Thus, the integration of sight and sound into a single percept in this multimodal condition might require additional cognitive resources for the HVPT-AV group as opposed to the HVPT-A group.

However, while the HVPT-A group outperformed the HVPT-AV group throughout the training phase, discrimination accuracy scores achieved at pre- and post-test from participants in both groups indicate that discrimination accuracy is higher with less variability overall (single

voice, no visual modality). This finding aligns with the idea that reducing variability and focusing on a single modality can be beneficial in some listener groups, i.e., child learners (Giannakopoulou et al., 2017) or learners with lower perceptual abilities (Sadakata & McQueen, 2014). This might also apply to beginner learners who benefit from a more stable speech signal with less variability. As such, L2 phonemic contrasts that are not contrastive in the L1 carry a great deal of acoustic ambiguity in the initial stages of learning. Thus, mapping novel L2 sounds to specific phonetic categories if the L2 phonological system is still developing seems to be more difficult when variability is high. While previous studies have shown ample evidence that HVPT may lead to significant improvements in learners' perception accuracy of difficult contrasts due to increased robustness of the incoming stimuli from variable talkers, this may primarily be the case for advanced learners who have had exposure to and experience with the L2 (Pisoni et al, 1993; Iverson & Evans, 2009; Cebrian & Carlet, 2014). The two seminal studies reporting the success of HVPT over LVPT (Logan et al., 1991; Lively et al., 1993) are indicative of this technique underlying an exemplar theoretical approach, in that learners store and categorize speech sounds as individual instances (exemplars) through the clustering of similar exemplars. The diversity of these instances present in natural speech makes learners more adept at discriminating subtle acoustic differences between sounds thereby allowing them to develop detailed and specific representations of target sounds (Johnson, 1997; Goldinger, 1998). However, PAM(-L2) is the commonly referenced framework (Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005; Hazan et al., 2005), because it is assumed that experience and training will lead to robust and flexible representations of target sounds, thereby enhancing the ability to generalize phonetic contrasts. While this model does not convincingly capture the benefit of multiple talkers over a single talker, there is evidence that this perceptual learning and

the makeup of new phonological L2 categories works better with less variability in the initial stages, when the focus can solely be on the phonetic differences, as produced by a single speaker (Brekelmans et al., 2022). Thus, there may be a trade-off between the benefits of high variability for generalization and the cost of processing multiple talker input.

Interpreting the training results of the current study, we do see an upward trend in discrimination accuracy towards the end of the training (T9/T10) in some of the participants. This indicates that learning to distinguish L2 contrasts in an HVPT condition might pay off even in beginner learners after prolonged training exceeding the period of 6 weeks (ten 20-minute sessions). Previous research has produced ambiguous results regarding the optimal length of the training phase and duration of individual sessions, with ten 30-minute sessions over the course of 2-3 weeks showing perceptual gains in highly experienced Japanese learners of English on the consonantal contrast /r - l/, which could not be attributed to the HVPT condition specifically (Iverson, Hazan, & Bannister, 2005) Likewise, Lambacher et al. (2005) found that perception of only a few English vowels improved for Japanese L2 learners of English with six 20-minute sessions over the course of six weeks. The ideal duration of HVPT for perceptual gains is thus not known, but trends suggest that perceptual discrimination of L2 sounds in the early stages of development may take more time than low-variability training. The strengths of the HVPT technique, however, lie in its representation of a real-life L2 learning scenario, as it reflects the natural variability to which we are exposed in real-world speech communication. Learners typically encounter different speakers with idiosyncratic characteristics and can adapt to high variability over time, thus, training in an HVPT condition might still be beneficial with prolonged exposure and experience.

3.6.1 Summary, outlook, future research

The findings of the current study are specific to beginner learners, and further research is needed to explore the effectiveness of HVPT-A as compared to HVPT-AV in learners of different proficiency levels. In this context, it is noteworthy that the overwhelming majority of research investigating the effects of HVPT includes advanced or highly proficient learners of the L2s instead of true beginners (Lively et al., 1993; Wang & Munro, 2004; Thomson, 2012; Rato & Rauber, 2015). The previous work suggests a consensus that HVPT is more beneficial and automatically bears an advantage over LVPT (Logan et al., 1991; Lively et al., 1993; Hazan et al., 2005), which is why it has been integrated as standard methodology into the field of pronunciation training (Brekelmans et al., 2022).

The evidence from previous studies suggests that highly variable stimuli may be beneficial to those learners who have already had significant exposure to the L2 and who have been able to set up distinct categories for L2 sounds (Bradlow et al., 1997; Iverson & Evans, 2009). As such, highly variable input from talkers (i.e., stimuli that vary in voice quality and facial information) seems to be beneficial for proficient learners or learners with high language aptitude (Perrachione et al., 2011). Adding an audiovisual modality to HVPT can further increase the cognitive load in that seeing a talker's face does not alleviate the added cost of processing multiple talkers (Brekelmans et al., 2022). Since the current study featured relatively few participants, effects of HVPT vs. LVPT in addition to modality effects could not be directly compared. The pre- and post-test results, for which participants achieved consistently higher discrimination accuracy scores, lend support to the assumption that low variability may be beneficial in early L2 learning stages, as listeners can focus on the specific phonetics of a single talker. This then also raises questions about the design of HVPT stimuli in discrimination tasks.

Previous studies employing triads in ABX discrimination tasks do not mention how the voices of multiple talkers are presented to listeners within stimulus trials (Hazan et al., 2005). This is problematic for two reasons.

1) If the training features several speakers, but each triad only features a single speaker who produces three stimuli, there is a trial-by-trial consistency and the listener can adapt to the speaker-specific acoustic-phonetic details.
2) Within-stimulus talker variability substantially differs in the demands they make on listeners' speech perception, because the listeners have to deploy additional resources to process the stimulus. There seems to be a fundamental difference between stimulus triads produced by a single talker (Giannakopoulou et al., 2017), in which there is internal consistency, thereby allowing predictability of the stimulus's phonetic features, and stimulus triads that feature the voices of different talkers, thereby dramatically increasing intra-stimulus variability (Perrachione et al., 2011).

The current study employed an extreme form of HVPT in that all ABX triads featured the voices of three different talkers, requiring the listener to tune in to the acoustic-phonetic differences of the stimulus rather than talker-specific phonetic differences. Future studies employing the HVPT technique discrimination tasks should reassess whether training featuring multiple talkers can be considered true HVPT when triads are internally speaker-consistent.

In the future, studies with beginner-level L2 learners should be carried out and the benefits of LVPT vs. HVPT directly compared. While HVPT constitutes a more realistic approach to encountering the L2, the robustness of the stimuli along with the high degree of variability may cause difficulties in the initial stages of L2 learning when categories have not yet been established. Similarly, there is no agreed-upon training duration after which benefits of

HVPT emerge. Additionally, there is no consensus about how multiple talkers' voices should be presented in discrimination tasks stimuli in HVPT ABX discrimination tasks, i.e., should a single trial feature a single talker or three different talkers? Previous studies suggest that small amounts of HVPT already pay off; however, these studies predominantly tested advanced learners who have had extensive experience and exposure to the language (Brekelmans et al., 2022; Saito, Hanzawa, Petrova, Kachlika, Suzukida, & Tierney, 2022). Furthermore, even though low variability perceptual discrimination may be good (accuracy scores were high at pre-test in the current study), high variability might still be beneficial after prolonged exposure given that discrimination seemed to improve towards the last training sessions in the current study. Thus, the optimal number and distribution of training sessions for different learner proficiencies have yet to be determined. The current findings suggest that additional research is needed to determine the circumstances under which HVPT can support and enhance the efficacy of phonetic training and that learner proficiency levels play an important role when opting for high variability over lower variability

Chapter 4: Conclusion

Previous research has demonstrated that certain segmental contrasts within a language play a more significant role in communication than others. These contrasts, which are characterized as carrying a high functional load, typically involve phonemic oppositions that are acoustically and phonologically distinct. They are crucial for maintaining perceptual distinctions and avoiding confusion. The concept of functional load has been extended to inform studies on the intelligibility and comprehensibility of ESL (e.g., Munro & Derwing, 2006; Suzukida & Saito, 2019). With English serving as a lingua franca and having more L2 than L1 speakers, the goal of L2 pronunciation pedagogy has shifted from accent reduction to emphasizing intelligibility and comprehensibility. In this context, functional load has gained considerable appeal, as some high FL contrasts align with intelligibility and comprehensibility rather than accentedness (Munro & Derwing, 2006).

The theoretical implications of functional load suggest that phonemic distinctions crucial for effective communication are those that generate numerous minimal pairs. Conversely, distinctions that result in only a limited number of minimal pairs are deemed less important for successful communication. Empirical studies on L2 English pronunciation attainment focus on confusable segments that carry a relatively high functional load compared to other segments. For instance, Munro and Derwing (2006) and Suzukida and Saito (2019) conducted experimental studies in which L1 English listeners rated Cantonese-accented and Japanese-accented speech productions containing vowel and consonantal errors with low and high functional load. Their findings indicated that the high functional load English contrasts /r/ - /l/ and /b/ - /v/ had a negative impact in intelligibility and comprehensibility, in contrast to the salient but low functional load contrast /s - θ/.

The empirical findings concerning FL in L2 English with the emphasis placed on intelligible and comprehensible speech output, rather than accent reduction, have provided insights that can be applied to L2 German pronunciation pedagogy. While Oh et al. (2015) proposed an overall FL ranking based on more complex measures beyond minimal pair count, this principle had not been empirically tested among L1 German listeners. Although the phonemic contrasts that typically pose challenges for L2 German learners, such as the phonemic contrasts tested in Study 2 (e.g., the vowels /uː - yː/ and /oː - øː/ and the consonants /k - x,ç/ and /z - t͡s/), did not rank highly in the FL hierarchy, Study 1 in Chapter 2 of this thesis focused on segmental substitutions involving both high and low FL contrasts within German utterances, namely /n - ʁ/ and /v - z/ (high FL), and /p͡f - t/ and /j - v/ (low FL), irrespective of their potential for confusion as observed in non-native speakers. To assess the impact of single or cumulative substitution errors that are common in L2 German on intelligibility and comprehensibility, a third category of confusable segments (CS) was examined in the study. The results demonstrated that errors in high FL contrasts significantly impaired intelligibility and comprehensibility compared to confusable segmental errors. However, L2 learners of German are unlikely to produce both the high and low FL contrast substitutions that were the focus of the study. Nevertheless, significant effects on intelligibility and comprehensibility when errors occurred cumulatively supported the notion that individual segmental errors, regardless of classification as low or high FL, can affect word recognition and understanding of utterances. Although Study 1 did not allow for the prioritization of specific sound contrasts in the instruction of L2 German based on their functionality in communication, it provided insights into the theoretical concept of functional load and the processes involved in word recognition during utterance comprehension. Previous research on L2 English utilizing FL assumed that intelligibility and comprehensibility

rely on the number of minimal pairs formed by a particular sound contrast. However, it remains unclear how a large number of distinctions, such as English /r - l/ distinguishing approximately 650 minimal pairs (e.g., *a**l**ive - a**rr**ive*, *c**l**oud - c**r**owd*, ***l**aw - **r**aw*), alone can account for decreased communication. On the one hand, this would imply that listeners possess implicit knowledge of the specific location of a substitution within a word. On the other hand, it would suggest that only substitutions resulting in real words can lead to misunderstandings and contribute to reduced intelligibility. However, high FL substitutions rarely result in real words with similar frequency in the lexicon, belonging to the same word class, and fitting the same syntactic position within the same semantic context (Levis & Cortes, 2008). Study 1 provided evidence that listeners process utterances and words in a top-down manner by comparing the input to a set of lexical neighbours that align with the overall acoustic-phonological structure of the stimulus (18).

(18) Output: *I heard that this action was *irregal.*

IPA: [aɪ hɝd ðæt ðɪs ækʃn wʌz ɪriːgɫ]

Target: 'I heard that this action was illegal.'

In this example, the substitution of /l/ with /r/ results in a non-word. Depending on the density of the neighbourhood of **irregal* and the activation of competitors involving (near-)minimal pairs, several underlying target words can be entertained, such as: *legal, irreal, illegal*. Due to the limited number of neighbours available from memory as well as the relatively high phonological similarity of English /r/ and /l/ (Dai, 2021), the underlying target word might be recognized by the listener with relative ease. Even though /r/ and /l/ carry a high FL among confusable segments of English by minimal pair count, this alone cannot account for a potential loss of intelligibility of an utterance. By comparison, an utterance involving a low functional load

substitution, such as English /s - θ/, can still cause intelligibility and comprehensibility issues (19).

(19) Output: *She did the mass.*
IPA: [ʃiː dɪd ðə mæs]
Target: 'She did the math.'

In this example, a substitution classified as low FL can still result in misunderstanding and increased difficulty of comprehensibility. Even though the /θ - s/ substitution involves a relatively similar sound contrast which does not alter the acoustic-phonological structure of the word to a high extent, a listener might still entertain possible lexical neighbours that resemble the erroneous output word, such as: *mass, mess, math, best.*

In general, Study 1 has revealed certain limitations and identified several considerations regarding the functional load (FL) principle and its implications for L2 pedagogy. Firstly, it should be noted that German, as a language with a substantial number of L1 speakers and a relatively homogeneous language community, differs from English, which holds the status of a lingua franca the world over with a high prevalence of L2 speakers. In the context of English, the emphasis on intelligible and comprehensible speech output, rather than unaccented speech production, is crucial due to the varieties of speakers and listeners who need to establish common ground for communication (Sewell, 2017). This diversity in the repertoire is largely absent in German. Moreover, Study 1 found that prioritizing speech sounds in their communicative value to intelligibility and comprehensibility in German may yield limited results, as high FL contrasts are not typically confused by L2 speakers of German. On the other hand, FL as a theoretical framework for ranking segmental contrasts based on their role in distinguishing utterances holds promise for L2 studies examining the impact of segmental errors on intelligibility and

comprehensibility. Therefore – while not all high FL contrasts are confusable – English L2 pedagogy has identified those contrastive segments with confusability potential that carry relatively high FLs based on minimal pair count (Catford,1987; Brown, 1991). More sophisticated FL measures, such as those based on information theory, entropy, or a global approach (Oh et al., 2015), suggest that using minimal pair count and restricting the analysis to confusable contrasts only provides indirect reflections of the actual 'information value' associated with a phonemic contrast. Although Study 1 did not provide guidance regarding which sounds should be targeted in perceptual training for L2 German learners, Study 2 was designed to offer valuable insights into the perceptual training of German sound contrasts. This, in turn, can contribute to future pedagogical approaches.

Study 2 in Chapter 3 of this thesis investigated the effects of audio-only versus audiovisual HVPT on discrimination accuracy in beginner L2 learners of German. The HVPT technique has shown promise in previous studies (Aliaga-García and Mora, 2009; Barriuso and Hayes-Harb, 2018). Its strengths lie in its real-world applicability, as it provides focused exposure to multiple voices and can lead to quantifiable changes in the L2 perceptual systems and comprehensible L2 productions. In contrast, other perceptual training techniques and minimal pair training on identification and discrimination of speech sounds (Grenon, Sheppard, & Archibald, 2018; Pisoni, Aslin, Percy, & Hennessy, 1982) often involve laboratory studies that present target sounds in less ecologically valid settings.

Despite research demonstrating the value of HVPT, Study 2 revealed an unexpected outcome. The limitation of using a single talker who produced monosyllabic stimuli containing seven critical sound contrasts in the pre- and post-test modality resulted in higher discrimination accuracy among beginner learners of German compared to audio-only HVPT and audiovisual

HVPT. Surprisingly, throughout the six-week training period, the audiovisual condition resulted in significantly worse discrimination accuracy for five out of the seven German sound contrasts when compared to the audio-only condition. This finding contradicts previous work by Hazan et al. (2005) who suggested that adding a visual modality could enhance the stimuli and lead to greater improvement than audio-only training. However, Hazan et al. (2005) only found improvement for the /v - b/ contrast in English by Japanese L2 learners, where lip movement provided salient information. In the case of sounds that are articulatorily distinct, but not visibly so, such as /r - l/, the audiovisual training group did not outperform the audio-only group.

Previous studies on high-variability phonetic training (HVPT) have predominantly relied on the Perceptual Assimilation Model (PAM) or PAM-L2 as a guiding framework, emphasizing the perception of articulatory gestures by listeners. However, this framework fails to explain the advantages of exposure to multiple talkers compared to a single talker. The seminal study by Lively et al. (1991) suggests that HVPT is primarily based on an exemplar theoretical model, where exposure to multiple voices and the rich details of a speaker enhance the robustness of stored stimuli in the listener's database, along with other instances of the speech sound (Goldinger, 1998). HVPT is assumed to lead to higher success rates and generalizability to novel talkers due to the variability of speech stimuli. Based on this rationale, the current study included an audiovisual modality, predicting that observing a speaker's face and idiosyncratic features would enhance stimulus robustness and facilitate perceptual learning. However, the findings of the current study indicate the opposite effect in the AV group, with significant fluctuations in discrimination accuracy across all trained contrasts and overall accuracy lower than in the pre- and post-test condition, as well as significantly worse than in the A-only group.

These results suggest that at this stage, reducing the number of talkers and using audio-only stimuli might be preferable for listeners. Although laboratory-based research employing single talkers may lack ecological validity in some respects, the current study's findings highlight the potential benefits of lower variability training. This could mean being trained by a single talker, which reduces phonetic variability, would allow listeners to tune in to the speaker-specific phonetics to determine acoustic-phonetic differences between phonemic L2 sounds during the early stages of L2 learning. Furthermore, having ABX discrimination triads that are internally consistent (i.e., produced by a single speaker) may be beneficial because listeners are exposed to less within-stimulus variability, even if multiple speakers were to be featured in the the overall training (Perrachione et al., 2011). A comprehensive understanding of L2 phonological acquisition requires consideration of proficiency levels and experience. While advanced learners seem to adapt to diverse accents and talkers in real-world settings, beginner learners may be overwhelmed by multimodal processing, which demands additional cognitive resources (Mattys and Wiget, 2011). This means that in the early stages of L2 acquisition, phonemic differences of L2 sounds are not distinct to listeners and they therefore have to rely heavily on auditory information for phonetic perception. A multimodal approach (i.e., seeing a speaker along with auditory information) will create ambiguous and inconsistent auditory input when speech sounds are still relatively new, due to lack of phonemic knowledge, especially for listeners with weaker language ability, i.e., beginner/inexperienced learners (Giovannone & Theodore, 2021). With more experience with a language, this effect will likely become reduced.

In line with recent criticism of HVPT (Brekelmans et al., 2022), the current study highlights the necessity for a more comprehensive understanding of the effectiveness of HVPT in different learner groups (beginning vs. advanced). Beginner learners are rarely tested in HVPT

studies. Furthermore, the effectiveness of HVPT and LVPT are not directly contrasted to find an advantage of talker variability (Brekelmans et al., 2022).

4.1 Implications, limitations, and future research

The present thesis investigates the functional load principle in German. The theoretical implications of FL lie in the ranking of phonemic contrasts in their functionality to communication in a given language through corpora analysis. The principle has gained popularity in ESL pronunciation teaching, due to its real-world application potential to inform language instruction and the prioritization of speech sounds that matter to the intelligibility and comprehensibility of spoken utterances. As such, the FL principle has been tested empirically for the English language in an L2 context in an effort to determine which typically confused sound contrasts of English matter for comprehensible communication (Munro & Derwing, 2006; Suzukida & Saito, 2019). This prioritization is often assessed using minimal pair counts. For example, in English, contrasts like /n - t/ or /m - t/ rank high in distinguishing word pairs (Oh et al., 2015) due to their low phonological similarity, making them less prone to confusion in a real-world setting.

On the other hand, English contrasts like /r - l/ and /b - v/ (Munro & Derwing, 2006; Suzukida & Saito, 2019), which are commonly confused by certain L2 English speaker groups, were selected in previous studies based on their relatively high functional load ranking. Conversely, salient contrasts like /s - θ/ (Brown, 1991; Suzukida & Saito, 2019), which are also prone to confusion, ranked lower in the FL hierarchy. This distinction between high and low FL contrasts is crucial in ESL pronunciation instruction, as high FL errors negatively affect intelligibility and comprehensibility and increase listener effort, whereas low FL errors mainly

impact perceived accentedness (Munro & Derwing, 1995; Munro & Derwing, 2006). It is important to highlight that genuine high FL contrasts, such as /n - t/, which are not typically confused, have not been tested and trained in English. Yet, the FL principle has served to guide the prioritization of English sound contrasts that matter for intelligibility and comprehensibility within the group of confusable segmental contrasts (Munro & Derwing, 2006; Suzukida & Saito, 2019).

The current study empirically tested the applicability of FL computations for German (Oh et al., 2015) and found that high FL contrasts do indeed have a more significant impact on the intelligibility and comprehensibility of speech compared to low FL errors. However, genuine high FL contrasts in German consisted of phonemic oppositions with little confusability potential due to their phonological dissimilarity, such as /ʁ - n/ or /ʁ - d/ (Oh et al., 2015), whereas none of the typically confused segmental contrasts, i.e., /uː - yː/ and /z - t͡s/, would be classified as carrying a relatively high FL, as has been found for English. Instead, even lower-ranked contrasts like /p͡f - t/ and /j - v/ (Oh et al., 2015) were unlikely to cause confusion due to relative phonological dissimilarity.

To further explore ecologically valid confusion patterns, an additional group of substitutional errors (i.e., those that L2 learners are likely to confuse) was tested, including contrasts like /oː - øː/ and /k - x/. The significant differences between the three error groups (high FL (H), low FL (L), confusable (CS)) were largely dependent on two factors: 1) the phonological similarity of the segments forming the contrast; and 2) the number of lexical neighbors activated when comparing the acoustic-phonological similarity of the stimulus.

An in-depth analysis of the listener data obtained from participants in Study 1 revealed that phonologically similar substitutions led to rapid and accurate activation of the underlying target word, e.g., (20).

(20)	Item:	Das Verböt war romisch
	IPA:	[das fɐbøːt vɐ ʁoːmɪʃ]
	Target:	Das Verbot war römisch
	IPA:	[das fɐboːt vɐ ʁøːmɪʃ]
	Translation:	This prohibition was Roman.

High intelligibility and comprehensibility scores could be ascribed to the high phonological similarity of /oː - øː/ and the low number of lexical neighbours of **Verböt* and **romisch* activated, as the acoustic-phonological structure of the underlying target words *Verbot* and *römisch* were well preserved, so the utterance with the erroneously produced words was still intelligible and comprehensible. Conversely, the listener data showed that highly dissimilar substitutions (as part of the high FL group) altered the overall acoustic-phonological structure of the word. Additionally, if the output had a high number of lexical neighbours, intelligibility and comprehensibility was decreased because many competitors were activated, as the example (21) below shows.

(21)	Item:	Den *Hust gab es *ihne Band
	IPA:	[deːn hʊst gaːp ɛs iːnə bant]
	Target:	Den Hund gab es ohne Band
	IPA:	[deːn hʊnt gaːp ɛs oːnə bant]
	Translation:	The dog came without a leash

Substitution of /n/ with /s/ and /o/ with /iː/ altered the acoustic-phonological structure of the underlying target words. The number of lexical neighbours was high for **Hust* (*Hut* 'hat'; *Husten* 'cough'; *Frust* 'frustration'), but also near-lexical neighbours: *Gast* 'guest' and *Kuss* 'kiss') and low for **ihne* (*ohne* 'without'; *eine* 'a[indef.art.fem.]'), thereby decreasing recognition accuracy and thus intelligibility with listeners indicating higher effort to comprehend the utterance.

Furthermore, this study revealed that the occurrence of two errors in the same sentence had a detrimental impact on both intelligibility and comprehensibility. This was further attributed to the inhibition of contextual information access due to open-ended propositions that contained little context and semantic priming being utilized, thereby reducing the likelihood of understanding the utterance. Notably, this negative effect of double errors on intelligibility and comprehensibility was consistent across error categories (H, L, CS).

The findings shed light on the limitations of assessing functional load solely based on minimal pair counts of individual phonemes. As observed in English, a phonemic opposition's minimal pair count might not always be the primary factor influencing intelligibility loss, as substitutions may not even form real words, occur in the same syntactic position, or result in word of the same lexical category, which automatically reduces confusability (Levis & Cortes, 2008). Rather, it has been found that substitutions leading to non-words can also impact listener effort based on the overall number of lexical neighbours activated, which is not solely substitution-specific but depends on the alteration of the entire sound structure (Luce & Pisoni, 1998; Luce et al., 2000).

In the case of German, a language with rich inflectional morphology, case, and gender assignment, the likelihood of a phonemic substitution resulting in a contextually appropriate real

word is highly unlikely. Nevertheless, the current study demonstrated that substitution errors, especially when they accumulate, can negatively impact a speaker's comprehensibility regardless of their classification as high or low functional load. These findings underscore that the theoretical implications made from corpora, i.e., in predicting sound change, do not transfer well to real-life communicative situations. As such, while it seems intuitive that conflation of segmental contrasts that distinguish many word pairs must in practice have a negative impact on intelligibility and comprehensibility, the current study has revealed limitations of FL and its applicability to ecologically valid settings.

In particular, the impact of phonemic distinctions on intelligibility and comprehensibility based on minimal pair count remains unclear. Firstly, it has been found that phonemic contrasts that distinguish many word pairs have low confusability potential. Secondly, the evidence from the current study suggests that, depending on the degree of phonological similarity between the underlying sound and the substitution, the acoustic-phonological structure of the word can be altered to such an extent that lexical neighbours resembling the overall acoustic make-up of the word will be activated (Luce & Pisoni, 1998). This suggests that utterances are not processed sequentially, in a phoneme-by-phoneme approach until detection of the erroneous segment, which implies explicit listener knowledge of the location in which the substitution occurred. Rather, the findings suggest that FL cannot be studied independently of word recognition models proposing top-down processing of words. Therefore, the explanatory power of functional load in prioritizing certain phonemic contrasts over others has its limitations.

In conclusion, the theoretical complexities of functional load, initially used to predict sound change within a language (King, 1967), suggest that it can serve as a heuristic rather than a quantitative measure for making L2 pedagogical recommendations. More research is needed to

explore functional load in diverse languages and contexts, with an understanding that a single (potentially high functional load) phoneme can only offer a partial explanation for the varying difficulty of erroneously produced words in context. As Study 1 could not determine the segmentals of German for L2 pronunciation training based on functional load computations, Study 2 empirically tested the perception of commonly confused German contrasts using a high-variability phonetic training paradigm.

Study 2 investigated the use of high-variability phonetic training (HVPT) as a standard methodology in L2 perceptual training for beginner learners of German. While HVPT has been widely adopted due to its advantages over low-variability phonetic training (LVPT), the results of Study 2 suggested the need for cautious consideration when employing this technique.

Previous research has shown that training listeners with highly variable speech input, involving multiple talkers and varying phonetic contexts, can lead to enhanced perceptual learning that generalizes to new words more effectively than LVPT (Logan et al., 1991; Lively et al., 1993). This ecologically valid approach has been commonly integrated into L2 perceptual training due to the assumption that it is superior to training with a single talker.

In this context, the present study compared the effects of audio-only versus audiovisual HVPT in a beginner group of L2 German learners over the course of six weeks. Initially surprising, both modality groups, especially the HVPT-AV group, experienced difficulties in discriminating German phonemic contrasts compared to a low-variability pre- and post-test condition, where listeners performed with high accuracy. The HVPT-A group outperformed the HVPT-AV group throughout the training period, suggesting that multimodal processing added cognitive load.

This study highlights the limitations of the widely held belief that HVPT is by default advantageous over LVPT (Brekelmans et al., 2022). While the current study was unable to compare effects of HVPT with LVPT, there is reason to believe that increased variability can in fact be overwhelming for true beginner learners. This calls for future research directly contrasting HVPT with LVPT. In fact, some studies have reported no advantages of HVPT, with a few even considering low-variability phonetic training (LVPT) (being trained by a single talker) to be more effective, especially for learners with low language aptitude (Perrachione et al., 2011). While recent HVPT studies suggest compelling results and an advantage over LVPT there are especially too few studies directly comparing HVPT to LVPT across L2 speakers with varying proficiency and experience. The three studies that directly compare HVPT to LVPT produced contradictory results. The influential work by Lively et al. (1991) compared the effects of HVPT and LVPT among Japanese learners of English and has since served as a framework for other HVPT studies. However, a more recent study on Dutch speakers learning L2 Japanese examined the effects of HVPT with a restricted set of tokens produced by a single speaker and more variable tokens produced by multiple speakers (Sadakata & McQueen, 2014). This study found that both types of training significantly improved generalization to untrained Japanese fricatives and novel speakers. In contrast, a recent training study by Giannakopoulou et al. (2017) found no advantage of HVPT over LVPT in adult and child L1 Greek speakers. Moreover, the impact of HVPT on true beginner L2 learners with developing phonological categories has not been thoroughly investigated.

The findings suggest that extremely variable input coming from multiple talkers (within a single triad) along with visual input can add substantial cognitive load to listeners in the initial stages of L2 learning. Adding to the findings from previous studies contrasting HVPT to LVPT,

this could point to a low-variability approach being more helpful, as ambiguity in the speech signal is reduced. HVPT produces ambiguous information through talker-variability, especially in the audiovisual modality, which may pose challenges for learners with little exposure to crucial phonetic information. However, the findings of Study 2 suggest that with prolonged training exposure, an upward trend in discrimination accuracy with HVPT might emerge as listeners adjust to the ambiguity. Especially in the HVPT-A group, discrimination accuracy would increase towards the final sessions of training for some of the German contrasts.

Despite the abovementioned limitations, it is important to highlight that HVPT constitutes an ecologically valid approach to L2 perceptual training, since real-world listening experiences involve exposure to multiple speakers in various settings. Nonetheless, the present study and previous research raise the need for further investigation of HVPT in beginner groups, as most significant effects have been reported in advanced learners who may have developed strategies to manage ambiguity (Bradlow et al., 1997; Cebrian & Carlet, 2014). HVPT may not universally be the optimal technique, especially for certain L2 learner groups like those with low aptitude (Perrachione et al., 2011), child learners (Giannakopoulou et al., 2017), and beginners. The complexity and richness of multi-talker input may hinder accurate perception in these contexts. As such, future studies should carefully examine the factors that contribute to the success of HVPT in specific learner groups and include a range of comparison groups to reinvestigate the believed superiority of training with highly-variable speech input, especially in the beginning stages of L2 learning.

References

Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. *British Journal of Educational Technology, 40*, 23–31. DOI: 10.1111/j.1467-8535.2007.00800.x

Aliaga-García, C. and Mora, J. C. (2009). Assessing the effects of phonetic training on L2 sound perception and production. In M. A. Watkings, A. S. Rauber and B. O. Baptista (Eds.), *Recent research in second language phonetics/phonology: Perception and production* (pp. 2–31). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Alnafisah, M., Goodale, E., Rehman, I., Levis, J., & Kochem, T. (2022). The impact of functional load and cumulative errors on listeners' judgments of comprehensibility and accentedness. *System, 110* (102906). DOI: https://doi.org/10.1016/j.system.2022.102906

Baese-Berk, M.M., Levi, S.V., & Van Engen, K.J. (2022). Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations. *The Journal of the Acoustical Society of America*, *153*(1), 68–76. DOI: 10.1121/10.0016806

Barriuso, T.A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *The CATESOL Journal, 30*, 177–194. DOI: http://www.catesoljournal.org/wp-content/uploads/2018/03/CJ30.1_barriuso.pdf

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*(8), 1151–1162. DOI: https://doi.org/10.3758/BF03208371

Bent, T., Bradlow, A. R., & Smith, B. L. (2007). Segmental errors in different word positions and their effects on intelligibility of non-native speech. All's well that begins well. In O-S. Bohn, & M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege* (pp. 331–347). (Language Learning and Language Teaching; Vol. 17). John Benjamins Publishing Company. DOI: https://doi.org/10.1075/lllt.17.28ben

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., & Tyler, M. (2007). Nonnative and second-language speech perception: Commonalities and complementaries. In O. S. Bohn, & M. Munro (Eds.), *Second-language Speech Learning: The Role of Language Experience in Speech Perception and Production*. A Festschrift in Honour of James E. Flege (pp. 13–34). Amsterdam: John Benjamins.

Boers, F., Warren, P., He, L., and Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System*, *66*, 113–29. DOI: 10.1016/j.system.2017.03.011

Bradlow, A.R., Akahane-Yamada, R., Pisoni, D.B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics 61* (5). 977–985. DOI:10.3758/BF03206911

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language, 126*, 1–4. DOI: https://doi.org/10.1016/j.jml.2022.10435

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly, 22*(4), 593–606. DOI: https://doi.org/10.2307/3587258

Brown, A. (1991). *Pronunciation Models*. Singapore: Singapore University Press.

Carlet, A. (2007). Different high variability procedures for training L2 vowels and consonants. *International Congress of Phonetic Sciences*, 944–948.

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systematic description of English phonology. In J. Morley (Ed.), *Current Perspectives on Pronunciation: Practices Anchored in Theory* (pp. 87–100). Washington, DC: TESOL

Cebrian, J. & Carlet, A. (2014). Second-language learners' identification of target-language phonemes: A short-term phonetic training study. *Canadian Modern Language Review*, *70*(4), 474–499. DOI: https://doi.org/10.3138/cmlr.2318

Celce-Murcia, M., Brinton, D., & Goodwin, J. (2010). *Teaching Pronunciation: A Course Book and Reference Guide*. Cambridge University Press, New York.

Chow, M., Macnamara, B.N. & Conway, A.R.A (2016). Phonological similarity in working memory span tasks. *Memory & Cognition*, *44*, 937–949. DOI: 10.3758/s13421-016-0609-8

Cole R.A, & Jakimik, J. (1980). How are syllables used to recognize words? *Journal of the Acoustical Society of America*, *67*(3), 965–970. DOI: 10.1121/1.383939

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*, 75–84. doi:10.1111/j.2044–8295.1964.tb00899.x

Dai, H. (2021). Gradient similarity in Lezgian laryngeal harmony: Representation and computation, *38th West Coast Conference on Formal Linguistics*, 147–157. Cascadilla Proceedings Project.

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ, 7*, Article e7191. DOI: https://doi.org/10.7717/peerj.7191

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence for four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. DOI: https://doi.org/10.1017/S0272263197001010

Derwing, T., & Munro, M. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476–490. DOI:10.1017/S026144480800551X

Derwing, T. M. (2010). L2 speakers' impressions of the role of pronunciation after 7 years in an ESL environment. Paper presented at the *2nd Annual Conference on Pronunciation in Second Language Learning and Teaching.* Ames, Iowa.

Fallon, A.B., Groves, K. & Tehan, G. (1999). Phonological similarity and trace degradation in the serial recall task: When CAT helps RAT, but not MAN. *International Journal of Psychology*, *34*(5/6), 301–307. DOI: https://doi.org/10.1080/002075999399602

Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. In: W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). Timonium, MD: York Press.

Flege, J., & Bohn, O. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress* (pp. 3–83). Cambridge: Cambridge University Press. DOI:10.1017/9781108886901.002

Frisch, S.A. (1997). *Similarity and frequency in phonology*. Northwestern University, Evanston: Illinois.

Frisch, S.A., Pierrehumbert, J.B., & Broe, M.B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory, 22*(1), 179–228. DOI: https://doi.org/10.1023/B:NALA.0000005557.78535.3c

Galantucci, B., Fowler, C.A., & Turvey, M.T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review 13*(3), 361–377. DOI: 10.3758/bf03193857

Giannakopoulou, A., Brown, H., Clayards, M. & Wonnacott, E. (2017). High or low? Comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ, 5*(5), E3209. DOI: 10.7717/peerj.3209

Giovannone, N., & Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research, 64*(3), 707–724. DOI: https://doi.org/10.1044/2020_JSLHR-20-00283

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. DOI: 10.1037/0033-295x.105.2.251

Gooskens, C. (2013). Methods for measuring intelligibility of closely related language varieties. In R. Bayley, R. Cameron, & C. Lucas (Eds.), *The Oxford Handbook of Sociolinguistics* (pp. 195–213). Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780199744084.013.0010

Grenon, I., Sheppard, C., & Archibald, J. (2018). Discrimination training for learning sound contrasts. *Proc. International Symposium on Applied Phonetics* (ISAPh 2018), 51–56, DOI: 10.21437/ISAPh.2018-9

Hahn, L.D. (2012). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201–223. DOI: https://doi.org/10.2307/3588378

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication, 47* (3). 360–378. DOI: https://doi.org/10.1016/j.specom.2005.04.007

Heeren, W. F. L., & Schouten, M. E. H. (2010). Perceptual development of the Finnish /t- tː/ distinction in Dutch 12-year-old children: A training study. *Journal of Phonetics, 38*(4), 594–603. DOI: https://doi.org/10.1016/j.wocn.2010.08.005

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review 93*, 328–338. DOI: 10.1037/0033-295X.93.4.411

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *The Journal of the Acoustical Society of America 116.* 2384–2394. DOI: https://doi.org/10.1121/1.1783351

Hockett, C. (1966). *The quantification of functional load: A linguistic problem Report Number RM-5168-PR.* Santa Monica: Rand Corp.

Hwang, H. & Lee, H.-J. (2015). The effect of high variability phonetic training on the production of English vowels and consonants. *International Congress of Phonetic Sciences*.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*, 135–159. DOI: https://doi.org/10.1080/15434303.2013.769545

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267–3278. DOI: https://doi.org/10.1121/1.2062307

Iverson, P. & Evans, B.G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, *126*(2), 866–877. DOI: https://doi.org/10.1121/1.3148196

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145–160. DOI: https://doi.org/10.1017/S0142716411000300

Jiao, S., Jin, H., You, Z., & Wang, J. (2022). Motivation and its effect on language achievement: Sustainable development of Chinese middle school students' second language learning. *Sustainability, 14*(16), 9919. DOI: https://doi.org/10.3390/su14169918

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson & Mullennix (Eds.) *Talker Variability in Speech Processing.*(pp.145–165) San Diego: Academic Press..

Kang, O., Thomson, R. I., & Moran, M. M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*(1), 115–146. DOI: https://doi.org/10.1111/lang.12270

Kartushina, N. & Martin, C. D. (2019). Talker and acoustic variability in learning to produce nonnative sounds: Evidence from articulatory training. *Language Learning 69*, 71–105. doi: 10.1111/lang.12315

King, R.D. (1967). Functional load and sound change. *Language*, *43*(4). 831–852. DOI: https://doi.org/10.2307/411969

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of NAACL 2000*: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 288–295.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, *37*(3), 273–291.

König, E. & Gast, V. (2009). *Understanding English-German Contrasts. Grundlagen der Anglistik und Amerikanistik.* Erich Schmidt Verlag.

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2007). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society, 12*, 979–1000. DOI: https://doi.org/10.1098/rstb.2007.2154

Lado, R. (1957). *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press, Ann Arbor, MI, US.

Lambacher, S.G., Martens, W.L., Kakehi, K., Marasinghe, C.A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics, 26*, 227–247. DOI: 10.1017/S0142716405050150

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Physics Doklady*, *10*, 707–710.

Lengeris, A. & Hazan,V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *The Journal of the Acoustical Society of America 128*. 3757–3768. DOI:10.1121/1.3506351

Lennon, P. (2010). Contrastive Analysis, Error Analysis, Interlanguage. In: S. Gramley & V. Gramley (Eds.): *Bielefeld Introduction to Applied Linguistics.* (pp. 51–60). Bielefeld: Aisthesis.

Leong, C. X. R., Price, J. M., Pitchford, N. J., & van Heuven, W. J. B. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLoS ONE, 13*(10), Article e0204888. https://doi.org/10.1371/journal.pone.0204888

Levis, J. M., and Cortes, V. (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. In C. A. Chapelle, Y.-R.Chung, & J. Xu (Eds.). *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment* (pp. 197–208). Ames, IA: Iowa State University.

Levis, J.M. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation, 6*(3). 310–328. DOI: https://doi.org/10.1075/jslp.20050.lev

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431–461. DOI: https://doi.org/10.1037/h0020279

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36. DOI: https://doi.org/10.1016/0010-0277(85)90021-6

Lim, S. & Holt, L.L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science 35*(7). 1390–1405. DOI:10.1111/j.1551-6709.2011.01192.x.

Lin, I. (2019) Functional load, perception, and the learning of phonological alternations. *University of California*, Los Angeles: California

Lively, S.E., Logan, J.S., & Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America 94*(3), 1242–1255. DOI: 10.1121/1.408177

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America, 89* (2), 874–886. DOI: https://doi.org/10.1121/1.1894649

Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62*(3), 615–625. DOI: https://doi.org/10.3758/BF03212113

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1991). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 122–147). The MIT Press.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1–36. DOI: 10.1097/00003446-199802000-00001

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576–585. DOI: 10.1037/0096-1523.15.3.576

Martinet, A. (1952). Function, structure, and sound change. *Word* 8, 1–32. DOI:10. 1080/00437956.1952.11659416

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language, 65*(2), 145–160. DOI: https://doi.org/10.1016/j.jml.2011.04.004

Max Planck Institute for Psycholinguistics. (2021). WebCelex Retrieved on October 5th, 2021!http://celex.mpi.nl.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86. DOI: https://doi.org/10.1016/0010-0285(86)90015-0

Mullennix, J.W. & Pisoni, D.B. (1990) Stimulus variability and processing dependencies in speech perction. *Perception & Psychophysics, 47*, 379–390. DOI: 10.3758/bf03210878

Mullennix, J.W. (1997). On the Nature of Perceptual Adjustments to Voice. In K. Johnson & J. Mullennix (Eds.), *Talker variability in speech processing* (pp.67–84). San Diego: Academic Press.

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. DOI: https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., and Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*(4), 520–531. DOI:10.1016/j.system.2006.09.004

Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research, 50*(6), 1496–1509. DOI: https://doi.org/10.1044/1092-4388(2007/103)

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, *64*(4), 715–748. DOI: https://doi.org/10.1111/lang.12082

O'Brien, M. G., & Fagan, S. M. B. (2016). *German Phonetics and Phonology: Theory and Practice* (1st ed.). Yale University Press. DOI: https://doi.org/10.12987/9780300225181

Oh, Y. M., Coupé, C., Marsico, E., and Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics*, *53*, 153–176. DOI:10.1016/j.wocn.2015.08.003

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America, 130*(1), 461–472. DOI:https://doi.org/10.1121/1.3593366

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 297–314. DOI: https://doi.org/10.1037/0096-1523.8.2.297

Pruitt, J.S., Jenkins, J.J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America 19*. 1684–1696. DOI: https://doi.org/10.1121/1.2161427

Rato, A., & Rauber, A. (2015). The effects of perceptual training on the production of English vowel contrasts by Portuguese learners. *International Congress of Phonetic Sciences*.

Roccamo, A. (2015). Teaching pronunciation in just ten minutes a day: A method for pronunciation instruction in first-semester German language classrooms. *Unterrichtspraxis/Teaching German*, *48*(1), 59–83. DOI: https://doi.org/10.1111/tger.10181

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology, 5*(November), 1–15. DOI: https://doi.org/10.3389/fpsyg.2014.01318

Saito, K., Hanzawa, K., Petrova, K., Kachlika, M., Suzukida, Y., & Tierney, A. (2022). Incidental and multimodal high variability phonetic training: Potential, limits, and future directions. *Language Learning*, *72*(4), 1049–1091. DOI: 10.1111/lang.12503

Sewell, A. (2017). Functional Load Revisited. *Journal of Second Language Pronunciation, 3,* 57–79. DOI:10.1075/jslp.3.1. 03sew

Sewell, A. (2021). Functional load and the teaching-learning relationship in L2 pronunciation. *Frontiers in Communication 6*, 1–6. DOI: 10.3389/fcomm.2021.627378

Shin, D.-J., & Iverson, P. (2013). Training Korean second language speakers on English vowels and prosody. *The Journal of the Acoustical Society of America 133*(5). 3333. DOI: https://doi.org/10.1121/1.4805598

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics, 66*(C), 242–251. DOI: 10.1016/j.wocn.2017.11.002

Smith, L.E., & Nelson, C. L. (1985). International intelligibility of English: directions and resources. *World Englishes*, *4*(3), 333–342. DOI: https://doi.org/10.1111/j.1467 971X.1985.tb00423.x

Strange, W., Levy, E.S., & Law, F.F. (2009). Cross-language categorization of French and German vowels by naive American listeners. *The Journal of the Acoustical Society of America, 126*(3), 1461–76. DOI: 10.7916/d8-gbwq-kg81

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research, 25*(3), 431–450. DOI: 10.1177/1362168819858246

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K.G. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *The Journal of the Acoustical Society of America*, *123*(1). 397–413. DOI: https://doi.org/10.1121/1.2804942

Thomson, R.I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal 28* (3). 744–765. DOI: 10.11139/cj.28.3.744-765

Thomson, R.I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning, 62*(4).1231–1258. DOI: https://doi.org/10.1111/j.1467-9922.2012.00724.x

Thomson, R.I. & Derwing, T.M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344. DOI: 10.1093/applin/amu076

Thomson, R.I. (2016). Does training to perceive L2 English vowels in one phonetic context transfer to other phonetic contexts? *Canadian Acoustics - Acoustique Canadienne 44* (3). 198-199.

Thomson, R.I. (2018). High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation, 4*(2), 208–231. DOI: 10.1075/jslp.17038.tho

Thomson, R. I. (2018). Measurement of accentedness, intelligibility and comprehensibility. In Okim Kang & April Ginther (Eds.), *Assessment in Second Language Pronunciation, 11*(29) (pp.11–29). York: Routledge. 10.4324/9781315170756-2

Thomson, R.I. (2018). High variability [pronunciation] training (HVPT). A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation, 4*(2), 208–231. DOI: https://doi.org/10.1075/jslp.17038.tho

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*(1-2), 306–331. DOI: 10.1006/brln.1999.2116

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 735–747. DOI: 10.1037//0278-7393.28.4.735

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research, 51*(2), 408–422. DOI: 10.1044/1092-4388(2008/030)

Vitevitch, M.S. & Luce, P.A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, *2*(1), 75–94. DOI: https://doi.org/10.1146/annurev-linguistics-030514-124832

Wang, X. & Munro, M.J. (2004). Computer-based training for learning English vowel contrasts. *System, 32* (4), 539–552. DOI: 10.1016/j.system.2004.09.011

Weber, A., and Cutler, A. (2004). Lexical competition in non-native spoken-word Recognition. *Journal of Memory and Language, 50*(1), 1–15. DOI:10.1016/S0749-596X(03)00105-0

Wedel, A., Kaplan, A., and Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128*(2), 179–186. DOI: https://doi.org/10.1016/j.cognition.2013.03.002

Wilson, C. & Obdeyn, M. (2009). Simplifying subsidiary theory: Statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms, 1-48.

Wong, J. W. S. (2012). Training the perception and production of English /e/ and /ae/ of Cantonese ESL learners: A comparison of low vs. high variability phonetic training. *14th Australasian International Conference on Speech Science and Technology*. Retrieved from https://assta.org/proceedings/sst/SST-12/SST2012/PDF/AUTHOR/ST120021.PDF.

Yu, A. & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, *5*(1), 131–150. DOI: doi/10.1146/annurev-linguistics-011516-03381

APPENDIX A

Calculations from Oh et al. (2015) of high functional load contrasts of German and English as per the WebCELEX database:

Table 18 High functional load consonant and vowel contrasts in German and English

German (WebCELEX)		English (WebCELEX)	
[ʁ, r] <**R**at>	[n] <**N**aht>	[n] <**n**ose>	[t] <**t**oes>
[m] <Rah**m**>	[ʁ, r] <ra**r**>	[m] <ra**m**>	[t] <ra**t**>
[d] <Fa**d**en>	[ʁ, r] <fah**r**en>	[ð] <**th**ough>	[m] <**m**ow>
[z] <**S**and>	[ʁ, r] <**R**and>	[s] <**s**ea>	[t] <**t**ea>
[v] <**W**ein>	[z] <**s**ein>	[d] <le**d**>	[t] <le**t**>
German (WebCELEX)		English (WebCELEX)	
[a] <R**a**tte>	[ɛ] <r**e**tte>	[aɪ] <f**i**le>	[eɪ] <f**ai**l>
[ɪ] <S**i**nn>	[aɪ] <s**ei**n>	[ɪ] <b**i**t>	[æ] <b**a**t>
[a] <K**a**ste>	[ɪ] <K**i**ste>	[eɪ] <v**ei**l>	[iː] <v**ea**l>
[aː] <z**a**hlen>	[iː] <z**ie**len>	[aɪ] <l**i**ke>	[iː] <l**ea**k>
[a] <B**a**nn>	[aɪ] <B**ei**n>	[ɪ] <b**i**t>	[ɒ] <b**ou**ght>

APPENDIX B

Table 19 Functional load contrasts and computations as per Oh et al. (2015)

HFL contrasts (MP #)	FL	LFL contrasts (MP#)	FL	CS contrasts (MP #)	FL
ʁ - n	0.110642	t͡ʃ - p	0.0000267901	z - t͡s	0.00370753
v - z	0.0258207	p͡f – t	0.0001649	k - x,ç	0.000757359
s – n	0.0458408	j - v	0.00623411	iː - eː	0.0378528
iː - a	0.0481865	yː - aː	0.0173188	a - aː	0.00422798
eː - aː	0.018917	oː - øː	0.00340562	uː - yː	0.00346689
oː - iː	0.0393373	ɪ - œ	0.000218672	uː - øː	0.0000779718

APPENDIX C

similarity-matrix[3587]

	b	ç	d	eː	f	h	iː	j	k	x	n	ʁ	l	m	œ	oː	p	p͡f	s	t	t͡s	uː	v	yː	z	øː	ŋ	ɑ	ɑː	ɔ	ɛ	g	ɪ	ʃ	ʊ	ʏ
b	1.0	0.54	0.82	0.5	0.86	0.79	0.5	0.57	0.82	0.75	0.71	0.71	0.64	0.89	0.54	0.5	0.96	0.89	0.71	0.79	0.75	0.5	0.89	0.5	0.75	0.5	0.75	0.57	0.54	0.54	0.54	0.86	0.54	0.71	0.54	0.54
ç	0.54	1.0	0.64	0.54	0.61	0.61	0.57	0.64	0.71	0.79	0.57	0.71	0.57	0.46	0.5	0.39	0.57	0.57	0.71	0.68	0.68	0.43	0.57	0.5	0.68	0.46	0.61	0.5	0.46	0.43	0.57	0.68	0.61	0.79	0.46	0.54
d	0.82	0.64	1.0	0.39	0.68	0.68	0.39	0.46	0.71	0.64	0.89	0.61	0.82	0.71	0.36	0.32	0.79	0.71	0.86	0.96	0.89	0.32	0.71	0.32	0.89	0.32	0.64	0.46	0.43	0.36	0.43	0.75	0.43	0.79	0.36	0.36
eː	0.5	0.54	0.39	1.0	0.46	0.54	0.96	0.89	0.57	0.61	0.43	0.68	0.5	0.54	0.86	0.86	0.46	0.43	0.39	0.36	0.36	0.82	0.5	0.89	0.43	0.93	0.64	0.82	0.86	0.79	0.93	0.61	0.89	0.39	0.75	0.82
f	0.86	0.61	0.68	0.46	1.0	0.86	0.46	0.54	0.75	0.82	0.61	0.71	0.61	0.79	0.5	0.46	0.89	0.96	0.79	0.71	0.75	0.46	0.96	0.46	0.75	0.46	0.64	0.54	0.5	0.5	0.5	0.71	0.5	0.79	0.5	0.5
h	0.79	0.61	0.68	0.54	0.86	1.0	0.54	0.61	0.75	0.82	0.61	0.71	0.61	0.71	0.5	0.46	0.82	0.82	0.79	0.71	0.75	0.46	0.82	0.46	0.75	0.46	0.64	0.61	0.57	0.5	0.57	0.71	0.57	0.79	0.5	0.5
iː	0.5	0.57	0.39	0.96	0.46	0.54	1.0	0.93	0.61	0.64	0.43	0.64	0.5	0.54	0.82	0.82	0.46	0.43	0.39	0.36	0.36	0.86	0.5	0.93	0.43	0.89	0.68	0.79	0.82	0.75	0.89	0.64	0.93	0.39	0.79	0.86
j	0.57	0.64	0.46	0.89	0.54	0.61	0.93	1.0	0.68	0.71	0.5	0.71	0.57	0.61	0.82	0.75	0.54	0.5	0.46	0.43	0.43	0.79	0.57	0.86	0.5	0.82	0.75	0.79	0.75	0.75	0.89	0.71	0.93	0.46	0.79	0.86
k	0.82	0.71	0.71	0.57	0.75	0.75	0.61	0.68	1.0	0.93	0.61	0.79	0.54	0.71	0.54	0.5	0.86	0.79	0.68	0.75	0.71	0.54	0.71	0.54	0.64	0.5	0.86	0.61	0.57	0.54	0.61	0.96	0.64	0.68	0.57	0.57
x	0.75	0.79	0.64	0.61	0.82	0.82	0.64	0.71	0.93	1.0	0.57	0.86	0.57	0.68	0.57	0.54	0.79	0.79	0.75	0.68	0.71	0.57	0.79	0.57	0.71	0.54	0.82	0.64	0.61	0.57	0.64	0.89	0.68	0.75	0.61	0.61
n	0.71	0.57	0.89	0.43	0.61	0.61	0.43	0.5	0.61	0.57	1.0	0.54	0.86	0.82	0.39	0.36	0.68	0.64	0.79	0.86	0.82	0.36	0.64	0.36	0.82	0.36	0.75	0.5	0.46	0.39	0.46	0.64	0.46	0.71	0.39	0.39
ʁ	0.71	0.71	0.61	0.68	0.71	0.71	0.64	0.71	0.79	0.86	0.54	1.0	0.54	0.64	0.64	0.68	0.68	0.68	0.64	0.57	0.61	0.64	0.75	0.57	0.68	0.61	0.75	0.79	0.75	0.71	0.71	0.82	0.68	0.64	0.68	0.61
l	0.64	0.57	0.82	0.5	0.61	0.61	0.5	0.57	0.54	0.57	0.86	0.54	1.0	0.68	0.46	0.43	0.61	0.57	0.79	0.79	0.75	0.43	0.64	0.43	0.82	0.43	0.61	0.57	0.54	0.46	0.54	0.57	0.54	0.71	0.46	0.46
m	0.89	0.46	0.71	0.54	0.79	0.71	0.54	0.61	0.71	0.68	0.82	0.64	0.68	1.0	0.57	0.54	0.86	0.82	0.64	0.68	0.68	0.54	0.82	0.54	0.68	0.54	0.86	0.61	0.57	0.57	0.57	0.75	0.57	0.64	0.57	0.57
œ	0.54	0.5	0.36	0.86	0.5	0.5	0.82	0.82	0.54	0.57	0.39	0.64	0.46	0.57	1.0	0.86	0.5	0.46	0.36	0.32	0.32	0.82	0.54	0.89	0.39	0.93	0.61	0.79	0.75	0.93	0.93	0.57	0.89	0.36	0.89	0.96
oː	0.5	0.39	0.32	0.86	0.46	0.46	0.82	0.75	0.5	0.54	0.36	0.68	0.43	0.54	0.86	1.0	0.46	0.43	0.32	0.29	0.29	0.96	0.5	0.89	0.36	0.93	0.57	0.82	0.86	0.93	0.79	0.54	0.75	0.32	0.89	0.82
p	0.96	0.57	0.79	0.46	0.89	0.82	0.46	0.54	0.86	0.79	0.68	0.68	0.61	0.86	0.5	0.46	1.0	0.93	0.75	0.82	0.79	0.46	0.86	0.46	0.71	0.46	0.71	0.54	0.5	0.5	0.5	0.82	0.5	0.75	0.5	0.5
p͡f	0.89	0.57	0.71	0.43	0.96	0.82	0.43	0.5	0.79	0.79	0.64	0.68	0.57	0.82	0.46	0.43	0.93	1.0	0.75	0.75	0.79	0.43	0.93	0.43	0.71	0.43	0.68	0.5	0.46	0.46	0.46	0.75	0.46	0.75	0.46	0.46
s	0.71	0.71	0.86	0.39	0.79	0.79	0.39	0.46	0.68	0.75	0.79	0.64	0.79	0.64	0.36	0.32	0.75	0.75	1.0	0.89	0.96	0.32	0.75	0.32	0.96	0.32	0.57	0.46	0.43	0.36	0.43	0.64	0.43	0.93	0.36	0.36
t	0.79	0.68	0.96	0.36	0.71	0.71	0.36	0.43	0.75	0.68	0.86	0.57	0.79	0.68	0.32	0.29	0.82	0.75	0.89	1.0	0.93	0.29	0.68	0.29	0.86	0.29	0.61	0.43	0.39	0.32	0.39	0.71	0.39	0.82	0.32	0.32
t͡s	0.75	0.68	0.89	0.36	0.75	0.75	0.36	0.43	0.71	0.71	0.82	0.61	0.75	0.68	0.32	0.29	0.79	0.79	0.96	0.93	1.0	0.29	0.71	0.29	0.93	0.29	0.61	0.43	0.39	0.32	0.39	0.68	0.39	0.89	0.32	0.32
uː	0.5	0.43	0.32	0.82	0.46	0.46	0.86	0.79	0.54	0.57	0.36	0.64	0.43	0.54	0.82	0.96	0.46	0.43	0.32	0.29	0.29	1.0	0.5	0.93	0.36	0.89	0.61	0.79	0.82	0.89	0.75	0.57	0.79	0.32	0.93	0.86
v	0.89	0.57	0.71	0.5	0.96	0.82	0.5	0.57	0.71	0.79	0.64	0.75	0.64	0.82	0.54	0.5	0.86	0.93	0.75	0.68	0.71	0.5	1.0	0.5	0.79	0.5	0.68	0.57	0.54	0.54	0.54	0.75	0.54	0.75	0.54	0.54
yː	0.5	0.5	0.32	0.89	0.46	0.46	0.93	0.86	0.54	0.57	0.36	0.57	0.43	0.54	0.89	0.89	0.46	0.43	0.32	0.29	0.29	0.93	0.5	1.0	0.36	0.96	0.61	0.71	0.75	0.82	0.82	0.57	0.86	0.32	0.86	0.93
z	0.75	0.68	0.89	0.43	0.75	0.75	0.43	0.5	0.64	0.71	0.82	0.68	0.82	0.68	0.39	0.36	0.71	0.71	0.96	0.86	0.93	0.36	0.79	0.36	1.0	0.36	0.61	0.5	0.46	0.39	0.46	0.68	0.46	0.89	0.39	0.39
øː	0.5	0.46	0.32	0.93	0.46	0.46	0.89	0.82	0.5	0.54	0.36	0.61	0.43	0.54	0.93	0.93	0.46	0.43	0.32	0.29	0.29	0.89	0.5	0.96	0.36	1.0	0.57	0.75	0.79	0.86	0.86	0.54	0.82	0.32	0.82	0.89
ŋ	0.75	0.61	0.64	0.64	0.64	0.64	0.68	0.75	0.86	0.82	0.75	0.75	0.61	0.86	0.61	0.57	0.71	0.68	0.57	0.61	0.61	0.61	0.68	0.61	0.61	0.57	1.0	0.68	0.64	0.61	0.68	0.89	0.71	0.57	0.64	0.64
ɑ	0.57	0.5	0.46	0.82	0.54	0.61	0.79	0.79	0.61	0.64	0.5	0.79	0.57	0.61	0.79	0.82	0.54	0.5	0.46	0.43	0.43	0.79	0.57	0.71	0.5	0.75	0.68	1.0	0.96	0.86	0.86	0.64	0.82	0.46	0.82	0.75
ɑː	0.54	0.46	0.43	0.86	0.5	0.57	0.82	0.75	0.57	0.61	0.46	0.75	0.54	0.57	0.75	0.86	0.5	0.46	0.43	0.39	0.39	0.82	0.54	0.75	0.46	0.79	0.64	0.96	1.0	0.82	0.82	0.61	0.79	0.43	0.79	0.71
ɔ	0.54	0.43	0.36	0.79	0.5	0.5	0.75	0.75	0.54	0.57	0.39	0.71	0.46	0.57	0.93	0.93	0.5	0.46	0.36	0.32	0.32	0.89	0.54	0.82	0.39	0.86	0.61	0.86	0.82	1.0	0.86	0.57	0.82	0.36	0.96	0.89
ɛ	0.54	0.57	0.43	0.93	0.5	0.57	0.89	0.89	0.61	0.64	0.46	0.71	0.54	0.57	0.93	0.79	0.5	0.46	0.43	0.39	0.39	0.75	0.54	0.82	0.46	0.86	0.68	0.86	0.82	0.86	1.0	0.64	0.96	0.43	0.82	0.89
g	0.86	0.68	0.75	0.61	0.71	0.71	0.64	0.71	0.96	0.89	0.64	0.82	0.57	0.75	0.57	0.54	0.82	0.75	0.64	0.71	0.68	0.57	0.75	0.57	0.68	0.54	0.89	0.64	0.61	0.57	0.64	1.0	0.68	0.64	0.61	0.61
ɪ	0.54	0.61	0.43	0.89	0.5	0.57	0.93	0.93	0.64	0.68	0.46	0.68	0.54	0.57	0.89	0.75	0.5	0.46	0.43	0.39	0.39	0.79	0.54	0.86	0.46	0.82	0.71	0.82	0.79	0.82	0.96	0.68	1.0	0.43	0.86	0.93
ʃ	0.71	0.79	0.79	0.39	0.79	0.79	0.39	0.46	0.68	0.75	0.71	0.64	0.71	0.64	0.36	0.32	0.75	0.75	0.93	0.82	0.89	0.32	0.75	0.32	0.89	0.32	0.57	0.46	0.43	0.36	0.43	0.64	0.43	1.0	0.36	0.36
ʊ	0.54	0.46	0.36	0.75	0.5	0.5	0.79	0.79	0.57	0.61	0.39	0.68	0.46	0.57	0.89	0.89	0.5	0.46	0.36	0.32	0.32	0.93	0.54	0.86	0.39	0.82	0.64	0.82	0.79	0.96	0.82	0.61	0.86	0.36	1.0	0.93
ʏ	0.54	0.54	0.36	0.82	0.5	0.5	0.86	0.86	0.57	0.61	0.39	0.61	0.46	0.57	0.96	0.82	0.5	0.46	0.36	0.32	0.32	0.86	0.54	0.93	0.39	0.89	0.64	0.75	0.71	0.89	0.89	0.61	0.93	0.36	0.93	1.0

Figure 19 German phonological similarity feature matrix as generated in Python

APPENDIX D

```
dlist = {
    'syll':1,
    'stress':1,
    'long':1,
    'consonant':1,
    'sonorant':1,
    'continuant':1,
    'delay':1,
    'approximant':1,
    'tap':1,
    'trill':1,
    'nasal':1,
    'voice':1,
    'sg':1,
    'cg':1,
    'labial':1,
    'round':1,
    'labiodental':1,
    'coronal':1,
    'anterior':1,
    'distributed':1,
    'strident':1,
    'lateral':1,
    'dorsal':1,
    'high':1,
    'low':1,
    'front':1,
    'back':1,
    'tense':1
    }      You, 6 minutes ago
```

Figure 20 Dlist in Python with assigned feature weights of German phones