THE UNIVERSITY *of* EDINBURGH

Phonemic Categorization and Phonotactic Repair as Parallel Sublexical Processes: Evidence from Coarticulation Sensitivity

Kiyoshi Ishikawa

Doctor of Philosophy
University of Edinburgh
2014

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Kiyoshi Ishikawa*)

Abstract

Phonemic perception exhibits **coarticulation sensitivity**, **phonotactic sensitivity** and **lexical sensitivity**. Three kinds of models of speech perception are found in the literature, which embody different answers to the question of how the three kinds of sensitivity are related to each other: **two-step models**, **one-step models** and **lexicalist models**.

In two-step models (Church, 1987), phonemes are first extracted, and phonotactic repairs are subsequently made on the obtained phoneme string; both phonemic categorization and phonotactic repair are sublexical, and coarticulation sensitivity should only affect initial (pre-phonotactic) phonemic categorization.

In one-step models (Dehaene-Lambertz et al., 2000; Dupoux et al., 2011; Mehler et al., 1990), phonemic categorization and phonotactic repair are sublexical and simultaneous; phonotactic repairs themselves depend on coarticulation cues. Such models can be implemented in two different versions: **suprasegmental matching**, according to which a speech signal is matched against phonotactics-respecting suprasegmental units (such as syllables), rather than phonemes, and **slot filling**, according to which a speech signal is matched against phonemes *as fillers for slots* in phonotactics-respecting suprasegmental units.

In lexicalist models (Cutler et al., 2009; McClelland & Elman, 1986), coarticulation sensitivity and/or phonotactic sensitivity reduce to lexical sensitivity. McClelland & Elman (1986) claim a lexicalist reduction of phonotactic sensitivity; Cutler et al.'s (2009) make a claim implying lexicalist reductions both of phonotactic sensitivity and of coarticulation sensitivity.

This thesis attempts to distinguish among those models. Since different perceptual processes are assumed in these three models (whether sublexical units are perceived, or how many stages are involved in perceptual processing), our understanding of how speech perception works crucially depends on the relative superiority of those three kinds of models.

Based on the results available in the past literature on the one hand, and on the results of perceptual experiments with Japanese listeners testing their coarticulation sensitivity in different settings on the other, this thesis argues for the superiority of the **slot filling** version of one-step models over the others. According to this conclusion, phonemic parsing (categorization) and phonotactic parsing (repair) are separate but parallel sublexical processes.

Acknowledgments

This is a thesis for my second doctoral degree. The first doctoral thesis was on compositionally formulated non-experimental semantics (submitted to the University of Tokyo; published, in a shortened form, by Indiana University Linguistics Club), which means that my background was in non-experimental syntax and semantics. It was after compeleting that thesis and getting a full-time teaching position at Hosei University (Tokyo, Japan) that I became interested in speech perception on the one hand, and have come to believe that linguistic evidence should be obtained through controlled experiments on the other. In order to receive training on experimental research, I decided to become a student again, utilizing a sabbatical year from Hosei.

Working under the supervision by Alice Turk (principal supervisor) and Martin Corley (second supervisor), I have learned (and probably am still learning) a lot, ranging from empirical facts already known, to designing experiments to distinguish competing hypotheses, to statistical techniques, and to thesis write-up. In fact, life under their supervision was not easy, but that was what I needed in order to become an experimental psycholinguist. The focus of this thesis would have been unbelievably (more) unclear without Alice's efforts to read through difficult-to-read drafts and give a tremendous number of critical comments; the typefaces would have been much more unreadable if Martin's remark that my LaTeX typesetting could be improved. I also thank Bobb Ladd, who decided to accept me as a PhD student (and who served as my second supervisor before Martin took the role, when Alice was on leave).

To start working in a new field does not in general have to, and in my case did not, mean

giving up old fields. I have been the head of a research team on syntax/semantics back in Tokyo, and the team's activity has been running in parallel with my work as a non-residential part-time Edinburgh student. This thesis would not have been completed without the warm understanding of my situation by the members of the team: OBA Ryo, ISII So (and SAIZEN Akira; the family names are written before the given names in Japanese).

Several Japanese and American speakers provided their utterances as stimuli for my experiments. Due to the changes of the specific goal of the thesis, unfortunately many of them failed to be employed in the experiments reported here. However, I thank the efforts of: TAHARA Mai, SUGIMOTO Yasuko, SO Suimi, ISII So (again), TERAUCHI Saori, Brian Wistner, and Aaren O'Connor, as well as KODAKA Miku, without whom the recordings of the stimuli for Experiment 8 would have been impossible.

My interests in speech research were boosted by personal influences from MAKINO Takehiko, as well as interactions with my undergraduate students (particularly KAMATA Miho and TSUKADA Yasuhiro). However, the most crucial boost was by a demonstration of phoneme restoration effects by KASHINO Makio at NTT (during an annual meeting of the Japan Cognitive Science Society), although he has probably not even heard of my name; it was that demonstration that led me to perception research (rather than production research or phonology). TAJIMA Keiichi and KITAHARA Mafuyu hosted a farewell (and 'welcome to speech research') party when I was departing for Edinburgh. Keiichi, and particularly KAWASAKI Takako, have also been of help from time to time, before and after my stay in Edinburgh.

I would not have been able to apply to Edinburgh without the help by ISHIKAWA Akira and YATABE Shuichi. Peter Evans, and Miho (again), offered help for my stay in the U. K., and Shuichi (again) provided help (and goods) for my stay in Edinburgh. My life in Edinburgh was made enjoyable particularly by Jiang Liu, Wenshan Li, ARAI Manabu and UTSUGI Akira.

Finally, (the existence of) N. A. Y. has always been a source of encouragement for my research (either in speech or syntax/semantics).

Contents

List of Tables

List of Figures

Chapter 1

The Overall Goal

Phonemic perception has been assumed to be context-sensitive in several respects. First, articulatory gestures of neighboring phonemes temporally overlap, a phenomenon known as coarticulation; as a result, auditory cues for neighboring phonemes temporally overlap, and portions of auditory signals and perceived discrete phonological units do not correspond one-to-one. For example, given a /CV…/ or /…VC/ sequence, /C/ and /V/ are coarticulated, and the place of /C/ and the identity of /V/ interact and jointly determine the initial formant transitions of the physical realizations of /V/, as a result of which listeners' perception of the place of /C/ depends on the perception of the identity of /V/, for example (Liberman, 1955, among others). Such context-dependency in perceptual behavior can be described as **coarticulation sensitivity**. Second, each language has its own phonotactic constraints, which also affect phonemic perception. For example, Massaro & Cohen (1983) observed that English listeners interpret the same cues (a synthetic [s]–[ʃ] continuum) in different ways in the [_li] context vs. in the [_ri] context, not as a result of coarticulation, but rather because of the English phonotactic bans on */sri/ and */ʃli/; /ʃ/ perception is more likely in the [_ri] context and /s/ perception is more likely in the [_li] context. Such context-dependency in perceptual behaivor can be described as **phonotactic sensitivity**. Third, lexical or syntactico-semantic factors, which can

be seen as contextual factors, have also been observed to affect speech perception (e.g., Warren and Warren, 1970). Such context-dependency in perceptual behavior can be described as **lexical sensitivity**.

How are those three kinds of context sensitivity related to each other? Three kinds of models of speech perception are found in the literature, which embody different answers to that question: **two-step models**, **one-step models**, and **lexicalist models**.

Two-step models (as explicitly claimed by Church, 1987), which have often been implicitly assumed in the literature on phonotactic sensitivity, are models in which individual phonemes are first extracted from the speech signal without reference to phonotactics, and phonotactic violations within the obtained phoneme string are subsequently repaired. On the view that subphonemic details due to coarticulation could affect the initial phonemic categorization but must have been filtered out once discrete phonemes have been extracted from a continuous speech signal, two-step models imply that perception could be coarticulation-sensitive only before phonotactic repairs (phonotactic sensitivity). On this view, phonemic categorization and subsequent phonotactic repair are both assumed to be sublexical. Thus, in two-step models, the three kinds of context sensitivity are assumed to be independent of each other.

One-step models (Dehaene-Lambertz et al., 2000; Dupoux et al., 2011; Mehler et al., 1990) are similar to two-step models in that phonemic categorization and phonotactic repair are both assumed to be sublexical, but challenge the independence between coarticulation sensitivity and phonotactic sensitivity assumed in two-step models. They embody the claim that phonotactic repairs are simultaneous with phonemic perception, and hence phonotactic repairs themselves depend on coarticulation cues; for example, vowel epenthesis within a consonant cluster may result from phonotactic repair, with the identity of the vowel being affected by coarticulation cues within the consonants (Dupoux et al., 2011). In other words, in one-step models, coarticulation sensitivity and phonotactic sensitivity interact. Two versions of one-step models are conceptually possible. According to one version, which we could call **suprasegmental match-**

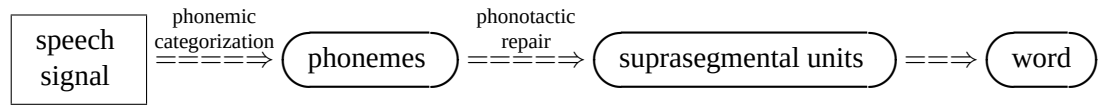
ing (Dehaene-Lambertz et al., 2000; Dupoux et al., 2011; Mehler et al., 1990), a speech signal is matched against phonotactics-respecting suprasegmental units (such as syllables), rather than phonemes, while according to the other version, which we could call **slot filling**, a speech signal is matched against phonemes *as fillers for slots* in phonotactics-respecting suprasegmental units. For example, suppose a Japanese listener hears an English speaker's utterance of *crisp*. According to the **suprasegmental matching** version, a Japanese listener is equipped with a repertoire of, say, syllables including /ku/, /ri/, /su/, and /pu/, but not syllables with onset clusters or codas; the speech signal is matched against such a repertoire, and the closest match would be /ku.ri.su.pu/, with the underlined portions 'epenthesized'; according to this version of one-step models, the perception of individual phonemes is not perception but rather a result of post-perceptual meta-analysis of the perceived suprasegmental units. In contrast, according to the **slot filling** version, phonemic perception is a real perceptual process, but phonemes are perceived only as fillers for slots within the structural frames of phonotactically relevant suprasegmental units. For example, when Japanese listeners hear the same *crisp* stimulus, they hear four syllables, where /k/ is perceived only as a filler for the onset /C/ slot of a /CV/ syllable, and hence /u/ is epenthesized as a filler for the nucleus /V/ slot; similarly for /s/ and /p/, with the resulting percept of /kurisupu/. Thus, while the **suprasegmental matching** version employs one common mechanism (matching against the acquired repertoire of suprasegmental units) to account for behavior previously described as coarticulation sensitivity and phonotactic sensitivity, the **slot filling** version assumes that coarticulation-sensitive phonemic categorization and phonotactics-repairing suprasegmental unit construction are separate processes but crucially depend on each other. Although I am aware of no proposal of the **slot filling** version (to be elaborated in Chapter 3), not only the **suprasegmental matching** but also the **slot filling** version are possible as specific implementations of the core idea of one-step models.

Lexicalist models (Cutler et al., 2011; McClelland & Elman, 1986) claim that coarticulation sensitivity and/or phonotactic sensitivity reduce to lexical sensitivity. McClelland &

Elman (1986) assume that phonemic units spread activations to lexical representations on the one hand, and receives feedbacks from activated lexical representations on the other; alleged phonotactic effects arise because so-called phonotactically licit strings of phonemic units receive feedback activations from a larger number of words than so-called phonotactically illicit strings, resulting in the perceptual bias toward so-called phonotactically licit strings. Thus alleged phonotactic effects is reduced to lexical sensitivity. Cutler et al. (2009) claim that (what they assume to be) the [C] portions excised from /…CV…/ utterances could give rise to the perception of /V/ only as a result of lexical activations. Cutler et al.'s (2009) claim has stronger implications than McClelland & Elman's (1986). Cutler et al.'s stimuli should constitute a phonotactic violation in the native language of the participants (Japanese) without the perception of some /V/ after (what they assumed to be) [C]. Thus such /V/ perception could arise from phonemic categorization of (what they assumed to be) [C] as /CV/ based on the coarticulation traces of /V/ (coarticulation sensitivity) on the one hand, and from phonotactic repair epenthesizing /V/ after /C/ (phonotactic sensitivity). Therefore, Cutler et al.'s (2009) attribution of such /V/ perception to lexical activations implies that both phonotactic sensitivity and coarticulation sensitivity (if real) should be seen as a result of lexical activation. (See Chapter 3 for a more detailed exposition of Cutler et al.'s 2009 claim.) According to such models, coarticulation/phonotactics-sensitive perception of non-word stimuli will be guided by the recognition of existing words resembling the non-word stimuli, just as coarticulation/phonotactics-sensitive perception of word stimuli will be guided by activated multiple candidate lexical entries; hence coarticulation/phonotactic sensitivity is seen as a result of perceptual assimilation of the speech signal to existing words and hence is reduced to lexical sensitivity.

The goal of this thesis is to distinguish among the three kinds of models (two-step models, one-step models, and lexicalist models; including a comparison between the two versions of one-step models). Since different perceptual processes are assumed in these models (whether

(a) two-step models



(b) one-step models

(c) lexicalist models

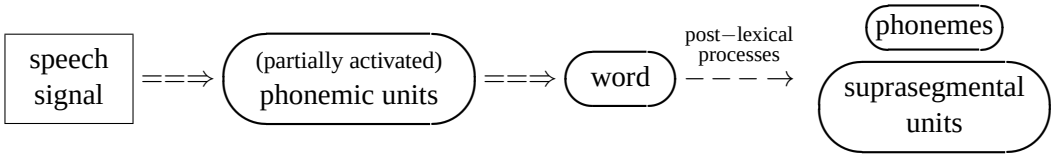


Figure 1.1: Schematic 'flowcharts' of the perceptual process according to the three kinds of models; according to (a) two-step models, coarticulation sensitivity is exhibited in the first-step 'phonemic categorization' process, and phonotactic sensitivity is exhibited in the second-step 'phonotactic repair' process; according to (b) one-step models, both coarticulation and phonotactic sensitivity is exhibited in a single 'phonemic categorization'–'phonotactic repair' process (the **suprasegmental matching** version) or separate but simultaneous 'phonemic categorization' and 'phonotactic repair' processes (the **slot filling** version); according to (c) lexicalist models,

sublexical units are perceived, or how many stages are involved in perceptual processing; see Figure 1.1), our understanding of how speech perception works crucially depends on the relative superiority of those three kinds of models.

In this thesis, perceptual experiments with Japanese listeners testing their coarticulation sensitivity in different settings are employed as the primary method. Perceptual experiments with Japanese listeners have two advantages.

The first advantage comes from the phonological and phonetic characteristics of Japanese, which makes it possible to distinguish among one-step models, two-step models and lexicalist models. It prohibits (non-homorganic) obstruent consonant clusters /C_1C_2/ altogether, and its phonotactics require that a medial vowel /V/ should break up such clusters. Thus, when stimuli are presented that non-native listeners would perceive as /…C_1C_2…/, Japanese listeners perceive /…C_1VC_2…/. However, such /V/ perception by Japanese listeners could arise in two

different ways: through phonemic categorization and through phonotactic repair.

In Japanese, high vowels /i, u/ devoice when both /C_1/ and /C_2/ are *voiceless*, typically leaving coarticulation traces of the devoiced vowel within the physical realization of /C_1/ on the one hand, but no vocalic portion between (the constriction intervals of) the two consonants on the other. Thus a natural expectation would be that, when stimuli such as [ekta] are presented, where [k] has coarticulation traces of /i/, for example, the '[k] portion' would be perceived as a vowel-devoiced realization of, and hence phonemically categorized as, /ki/, rather than as /k/. The identity of the perceived /V/ (in this case, /i/) in such devoicing-based phonemic categorization should, by definition, rely on the coarticulation cues within the '/C_1/ portion'. Coarticulation sensitivity exhibited in this way is called **devoicing-based coarticulation sensitivity** in this thesis. Crucially, /C_1/ has to be voiceless for such /CV/ perception, because it is phonemic categorization enabled by **devoicing-based coarticulation sensitivity**, which leads /CV/ categorization of a voiceless [C] with appropriate coarticulation traces.

On the other hand, phonotactic repair resulting in vowel epenthesis is a common phenomenon cross-linguistically, and in the case of Japanese, /u/ is perceptually epenthesized by default, default in the sense that /u/ is chosen in the absence of coarticulation cues for the identity of the vowel within the physical realization of /C_1/ (Dupoux, Fushimi, Kakehi & Mehler, 1999).[1] The predictions of one- and two-step models differ with respect to whether the identity of the perceived /V/ through phonotactic repair (not phonemic categorization) could depend on the coarticulation cues within consonants; it should be able to in one-step models but not in two-step models. The coarticulation sensitivity in such cases is independent from devoicing, and hence is called **devoicing-independent coarticulation sensitivity** in this thesis.

Devoicing-based coarticulation sensitivity and **devoicing-independent coarticulation sensitivity** are long names and, in addition, easy to misread as each other. Thus they will be

[1] Speaking more precisely, /u/ is the default epenthetic vowel when /C_1/ is not a palatal affricate or dental stop; see Chapter 2 for more details.

abbreviated as **DB-sensitivity** and **DI-sensitivity** in the rest of this thesis.[2]

Suppose that Japanese listeners' coarticulation sensitivity is observed in the identity of /V/ when they perceived /C_1VC_2/ from stimuli that non-native listeners would perceive as /C_1C_2/. The observation could either be interpreted as **DB-sensitivity** (to be exerted in phonemic categorization, resulting in /CV/ categorization of the '[C_1] portion', a possibility compatible both with one- and with two-step models), or as **DI-sensitivity** (to be exerted in phonotactic repair, resulting in a default-overriding /V/ epenthesis, a possibility expected from one-step models but excluded by two-step models). However, the two interpretations could be teased apart by manipulating the voicing of /C_1/; **DB-sensitivity** could be exerted only when /C_1/ is voiceless, and hence coarticulation sensitivity observed with a voiceless /C_1/ could well be interpreted as **DB-sensitivity** exerted in phonemic categorization, but coarticulation sensitivity observed with a voiced /C_1/ could only be interpreted as **DI-sensitivity** exerted in phonotactic repair. Thus examinations of the reality of each of the two kinds of coarticulation sensitivity by manipulating the voicing of /C_1/ would tell us which of one- and two-step models are superior to the other.

Furthermore, high vowels usually devoice only when both [C_1] and [C_2] are voiceless. That means that [...C_1V̥C_2...] are possible as words only when both [C_1] and [C_2] are voiceless. Thus, if **DB-sensitivity** could be reduced to lexical sensitivity, its effects should only be observed with stimuli in which both [C_1] and [C_2] are voiceless (e.g., [ekɨ̥ta], inducing /ekita/ perception); its effects should not be observed with stimuli in which [C_2] is voiced (e.g., [ekɨ̥ma], failing to induce /ekima/ perception). In contrast, if effects of **DB-sensitivity** are observed with stimuli in which [C_2] is voiced (e.g., [ekɨ̥ma], inducing /ekima/ perception), that would mean that **DB-sensitivity** has induced phonetic categorization of the [C_1] portion as /CV/ irrespective of whether the whole stimulus [... C_1V̥ C_2...] is a possible word pronunciation in Japanese, which would argue against lexicalist models. Thus, by seeing whether effects

[2] I owe those names to Martin Corley.

of **DB-sensitivity** are observed when [C_2] is voiced (the sublexical reality of **DB-sensitivity**), we can distinguish between lexicalist vs. non-lexicalist (one- and two-step) models.

In short, the reality of **DI-sensitivity** would distinguish one- vs. two-step models, and the sublexical reality of **DB-sensitivity** would tease apart lexicalist vs. non-lexicalist models. Thus Japanese allows comparisons among the three kinds of models. (For more details, see Chapters 3–4.)

The second advantage of perceptual experiments with Japanese listeners is that the results could be compared and/or combined with some crucially relevant results with Japanese listeners in the literature (Cutler et al.'s 2009 argument for lexicalist models; Dupoux et al.'s 2001 and Mazuka et al.'s 2011 argument against lexicalist models; Dupoux et al.'s 2011 argument for one-step models; Fais et al.'s 2005 results that could be taken as an argument for lexicalist models; Ogasawara & Warner's 2009 observation of a discrepancy between phoneme monitoring and lexical decision). Note that the comparison between lexicalist vs. non-lexicalist models suggested in the previous paragraphs concerns whether coarticulation sensitivity reduces to lexical sensitivity, not whether phonotactic sensitivity reduces to lexical sensitivity; the latter issue could be resolved by combining the previous results in the literature (Duoux et al., 2001; Fais et al., 2005; Mazuka et al., 2011) on the one hand, and the results of the examination of the reality of **DB-sensitivity** on the other. (Again, see Chapters 3–4 for details.)

The organization of the thesis is as follows.

Chapter 2 presents relevant background on Japanese phonetics and phonology, with particular focus on vowel devoicing.

Chapter 3 reviews the previous literature on the three kinds of models. Evidence in support of and against lexicalist models has been provided in the literature (Cutler el al., 2009; Dupoux et al., 2001; Fais et al., 2005); an account that explains all of those observations (based on a conjecture to be confirmed experimentally in Chapter 5) will be suggested in Chapter 3. Coarticulation-sensitive vowel perception from a consonant has been reported in the literature

(Beckman & Shoji, 1984; Dupoux et al., 2011; Ogasawara, 2013; Ogasawara & Warner, 2009; Tsuchida, 1994, Yuen, 2000), and some (Dupoux et al., 2011) was taken as evidence for one-step models; evidence for perceptual reality of sub-syllabic elements has been reported (Berent et al., 2007; Matthews & Brown, 2004; Moreton, 2002), and some of them were taken as potential evidence for two-step models (Dupoux et al., 2011). However, it will be pointed out that such observations in fact do not completely distinguish among the three kinds of models.

Chapter 4 frames the questions emerging from the literature review as experimental questions. As suggested above and elaborated in Chapter 3, the three kinds of models can be compared by examining the sublexical reality of **DB-sensitivity** on the one hand, and the reality of **DI-sensitivity** on the other. The examination of the sublexical reality of **DB-sensitivity** will enable us to determine whether lexicalist accounts of coarticulation sensitivity on the one hand, and of phonotactic sensitivity on the other, would be viable; **DB-sensitivity** would lead us to expect /ekima/ perception from stimuli such as [ekɨ̥ma] or [ekɛ̥ma], only if the sensitivity is real as sublexical sensitivity, because such strings as vowel-devoiced realizations of /ekima/ should not be possible words due to the voiced [m]; furthermore, the (un)reality of **DB-sensitivity** will also enable us to evaluate the arguments for and against a lexicalist account of phonotactic sensitivity by Dupoux et al. (2001) and Fais et al. (2005). The sublexical reality of **DB-sensitivity** would argue against lexicalist models, leaving only two-step and one-step models, while its unreality would support lexicalist models. (See the upper half of Table 1.1, particularly the [ekɛ̥ma] row; an observation of dominant /ekima/ perception would constitute evidence for the sublexical reality of **DB-sensitivity** on the one hand, and against lexicalist models on the other, while an observation of dominant /ekuma/ perception would support lexicalist models.)[3] On the other hand, evidence for the sublexical reality of **DB-sensitivity** (dominant /ekima/ perception against [ekɛ̥ma] stimuli) could be interpreted as evidence for coarticulation sensitivity

[3]An observation of dominant /ekema/ perception against [ekɛ̥ma] stimuli could be interpreted as suggesting either that **DB-sensitivity** is not sublexically real (lexicalist models), or that it is sublexically real but lost the effect of **DI-sensitivity** (one-step models), in which case two-step models should be dropped, with one-step models and lexicalist models remaining as candidates.

Table 1.1: The perception predicted by the three kinds of models, where (i) **DB-sensitivity** is real either at least sublexically (one- and two-step models) or (at most) lexically (lexicalist models), (ii) **DI-sensitivity** is either (A) unreal or (B) real, and (iii) if **DI-sensitivity** is real, its effect are weaker (B1) or stronger (B2) than the effect of **DB-sensitivity**.

stimuli	two-step models	one-step models	lexicalist models
[ekˌima]	/ekima/	/ekima/	(A) /ekuma/ (B) /ekima/
[ekˌema]	/ekima/	(B1) /ekima/ (B2) /ekema/	(A) /ekuma/ (B) /ekema/
[ekˌita]	/ekita/	/ekita/	/ekita/
[ekˌeta]	/ekita/	(B1) /ekita/ (B2) /eketa/	(A) /ekuta/ (B1) /ekita/ (B2) /eketa/
[egˌima]	(A) /eguma/	(B) /egima/	(A) /eguma/ (B) /egima/
[egˌema]	(A) /eguma/	(B) /egema/	(A) /eguma/ (B) /egema/
[egˌita]	(A) /eguta/	(B) /egita/	(A) /eguta/ (B) /egita/
[egˌeta]	(A) /eguta/	(B) /egeta/	(A) /eguta/ (B) /egeta/

in phonemic categorization and coarticulation-insensitive phonotactic repair within two-step models, or as evidence for larger effects of coarticulation sensitivity in phonemic categorization (**DB-sensitivity**) than weak but real effects of coarticulation sensitivity in phonotactic repair (**DI-sensitivity**) within one-step models, and hence does not distinguish one- and two-step models. In order to distinguish them, coarticulation sensitivity that could *not* be interpreted as coarticulation sensitivity in phonemic categorization, i.e., **DI-sensitivity**, needs to be examined; if **DI-sensitivity** turns out to be real, it would argue for one-step models, while its unreality would argue for two-step models; thus the (un)reality of **DI-sensitivity** will enable us to distinguish one-step models vs. two-step models. (See the lower half of Table 1.1.)

Chapter 5 reports Experiments 1–4, which test **DB-sensitivity**. Chapter 6 reports Experiments 5–8, which test **DI-sensitivity**. Chapter 7 reports Experiments 9–10, which are supplementary tests of both **DB-sensitivity** and **DI-sensitivity**.

Chapter 8 summarizes the results and concludes the thesis.

Chapter 2

Japanese Phonetics/Phonology and Vowel Devoicing

2.1 Introduction

This chapter presents an overview of Japanese phonetics and phonology. It is not meant to be a comprehensive overview; it only deals with the segmental phonology and phonetics, together with phonotactic constraints, while a discussion of the prosodic characteristics concerning accents and intonations is not contained at all. Many of the segments exposed, as well as the phonotactic constraints, play vital roles in the perceptual experiments to be discussed in Chapter 3 or reported in Chapters 5–6. Section 2.2 discusses the relevant parts of the Japanese phoneme inventory; among them, /p, b, k, g/, as well as /e, i, u/, are the crucial phonemes in the experiments conducted for this thesis. Section 2.3 illustrates the relevant parts of the Japanese phonotactic constraints. Section 2.4 illustrates place and/or manner alternations of consonants; the reason for employing /p, b, t, g/ but not /t, d/ in the experiments will become clear here. Section 2.5 discusses vowel devoicing; the reason for employing /e, i, u/, but not /a, o/, in the experiments will become clear here. Section 2.6 discusses what the overview (particularly

of the phonotactic constraints and vowel devoicing) suggests concerning perceptual epenthesis, a suggestion that plays a crucial role in the evaluations of most of the previous results in the literature as well as the designs of the experiments conducted for this thesis. Section 2.7 summarizes the characteristics of Japanese that play crucial roles in the following chapters.

2.2 The Phoneme Inventory

Japanese has five short vowels. Following Vance' (1987) transcription on the one hand, and the order Japanese speakers are conventionally taught in elementary schools when they learn writing on the other, let us write them as **/a/** (low non-front non-back), **/i/** (high front), **/u/** (high back), **/e/** (non-high non-low front), and **/o/** (non-high non-low back). It is often assumed that the /u/ phoneme in Japanese is realized as [ɯ]. However, speaking more precisely, according to Kamiyama (2008) and Shibatani (1990), it is the Eastern dialects in which /u/ is realized as [ɯ]; /u/ is rounded in the Western dialects.[1]

In addition, each vowel has corresponding long ones, /aː, iː, uː, eː, oː/.[2] In principle, two different vowels can appear in sequence (e.g., *kao* 'face'), but usually /ei/ and /ou/ are pronounced as [eː] and [oː] respectively, being neutralized with /eː/ and /oː/ respectively.

Japanese has three voiceless and three voiced stop phonemes: /p, t, k/ and /b, d, g/. Among them, (non-geminate) /p/ does not tend to occur in words which are Japanese or Chinese in origin, while it readily occurs in words with other origins (e.g., many loanwords from the Western languages) or onomatopoeic words which are abundant in Japanese. (For more details, see Labrune, 2012, among others.) A perceptual consequence of the special status of (non-geminate) /p/ is assumed by Cutler et al. (2009), as discussed in later chapters.

[1] Personally I have also observed a (non-Western!) native speaker whose /u/ is always rounded.

[2] In fact, how those long vowels should be phonemically transcribed has been rather controversial among phonologists; they could be analyzed as /aa, ii, uu, ee, oo/ (Haraguchi, 1999; Kubozono, 1999; Tsujimura, 1996; Vance, 1987; 2008), as /aR, iR, uR, eR, oR/ (with a special 'lengthening phoneme' /R/; Labrune, 2012), or as /aH, iH, uH, eH, oH/ (with a special 'lengthening phoneme' /H/; Vance, 2008).

In addition, it has three voiceless fricatives:[3] /ɕ/, which is realized as [ɕ]; /s/, which is realized as [s] before /a, e, u, o/ and neutralizes with /ɕ/ before /i/; /h/, which is realized as [h] before /a, e, o/, as [ç] before /i/, and as [ɸ] before /u/. The former two (/ɕ/ and /s/) have voiced counterparts; /ʑ/ and /z/. It also has two voiceless affricates:[4] /t͡ɕ/, which is realized as [t͡ɕ], and /t͡s/, which is realized as [t͡s]. One complication is that the voiceless stop /t/ neutralizes with /t͡ɕ/ and is realized as [t͡ɕ] before /i/ on the one hand, and neutralizes with t͡s/ before /u/ and is realized as [t͡s] on the other. The voiceless affricate /t͡ɕ/ could be said to have a voiced counterpart, but it has completely neutralized with /ʑ/ in the modern language, so let us write it as /ʑ/. Just as /t/ neutralizes with the voiceless fricative /t͡ɕ/ before /i/, /d/ neutralizes with that "voiced counterpart," /ʑ/, which is realized as [d͡ʑ] (or possibly as [ʑ]) before /i/. Similarly, the voiceless affricate /t͡s/ could be said to have a voiced counterpart, but, again, in the modern language, it has neutralized with the voiced fricative /z/ and is realized either as [z] (in utterance-initial positions) or as [d͡z] (elsewhere; Kamiyama, 2008). Furthermore, just as /t/ neutralizes with the voiceless affricate /t͡s/ before /u/ and is realized as [t͡s], /d/ also neutralizes with /z/ before /u/ and is realized either as [d͡z] or as [z].

It also has a glide /j/. Furthermore, two peculiar consonant phonemes traditionally assumed in the Japanese linguistics literature are /N/ and /Q/, which will be explicated in Section 2.3 below.[5]

One important characteristics of Japanese stops is that voiced stops are often subject to lenition; closures are usually so weak, and in the extreme but not rare cases, /g/ and /b/ are often realized as the fricatives [ɣ] and [β] respectively (Kamiyama, 2008; Vance, 2008).[6]

[3]The phonemic and phonetic transcriptions are from Labrune (2012).

[4]Again, the phonemic and phonetic transcriptions are from Labrune (2012).

[5]For the history of the development of the assumption of those two 'special phonemes' with no inherent place of articulation, written with the non-IPA symbols /N/ and /Q/, see Labrune (2012, Chapter 5), among others.

[6]In an older variety, /g/ was realized as [ŋ] in intervocalic positions, and Vance (1987; 2004) assumes that it is still observed in the modern language. This thesis assumes that that variety is now obsolete; in the modern variety, the intervocalic realizations of /g/ are usually [g] or [ɣ].

2.3 Phonotactic Constraints

It is widely assumed that both syllables and morae should be employed in the theoretical description of Japanese phonology. Labrune (2012) argues against this widely held view, claiming that alleged syllabic effects should best be described in terms of morae. However, this section illustrates Japanese phonotactics in syllabic terms, rather than moraic terms. This decision is based on expository convenience, not on the belief that Labrune's argument is invalid. Syllables are larger units than morae, and hence a Labrune-style moraic re-description of Japanese phonotactics, usually stated in syllabic terms, would unnecessarily complicate the exposition, which should best be avoided; whether the alleged syllabic effects (either in theoretical descriptions of Japanese phonology or in perceptual processes) could be re-described in terms of morae is simply beyond the scope of this thesis.

The inventory of syllable types is extremely limited in Japanese. Basically, Japanese syllables are either of the two types: V and CV. Due to influences from Chinese (Vance, 1987; 2000), a very limited class of onset consonant clusters on the one hand, and a very limited class of coda (non-cluster) consonants on the other, are found in the modern language. The onset clusters have to be of the form /Cj/, i.e., the second member of the cluster has to be the glide /j/. The coda has to be either /N/ (the moraic nasal) or /Q/ (a phoneme that is realized either as the first element of a geminate or as a glottal stop).

/N/ is realized either as a nasal homorganic with the immediately following obstruent (e.g., the /N/ in /kiN/ 'gold' realized as [ŋ] if followed by the nominative particle /ga/, as [n̪] if followed by the conjunction /to/,[7] or as [m] if followed by the conjunction /mo/), or as a nasalized vowel otherwise (e.g., the /N/ in /kiN/ being realized as [ĩ] in isolation). /Q/ is realized as a consonant (a mere closure with no release, in the case of a stop) homorganic with the immediately following consonant, or as a glottal stop when not immediately followed by a consonant;[8]

[7] In Japanese, [t] is a dental, rather than an alveolar, stop.

[8] One and the same morpheme ending with /Q/ is subject to allophonic variation depending on the immediately following segment. For example, *teQ* 'iron' is realized as [tepˀ] before /p/ as in /teQpan/ 'steel board', as [tetˀ]

when followed by a consonant, /Q/ is realized as the first member of a geminate. Note that, in the case of /QC/ where /C/ is s stop, a release accompanies /C/, but not /Q/.

One consequence of such phonotactic constraints is that non-homorganic obstruent clusters (e.g., /bz/) are totally banned in Japanese.

Although the above description is based on syllables, the Japanese orthography is based on morae, and each mora is assigned a single, independent 'syllabary' symbols, and for ordinary Japanese speakers, morae are intuitively natural phonological units, but not syllables. (The causal connection between the orthography and the naive intuition is beyond the scope of this thesis.)

2.4 Consonant Alternations

Some consonant phonemes are subject to allophonic alternations when immediately followed by /i/ on the one hand, and by /u/ on the other, as opposed to when immediately followed by the other vowels (/a, e, o/), as seen above. For example, /t, d/ (but not /k, g/ or /p, b/) are neutralized to affricates (or fricatives) immediately before /i/ (e.g., /ti/ and /t͡ɕi/ both realized as [t͡ɕi]; /di/, /ɖi/, /zi/ all realized as [d͡ʑi] word-initially and as [ʑi] word-internally), and to affricates or fricatives before /u/ (e.g., /du/ and /zu/ both realized either as [d͡zɯ], [d͡zu], [zɯ], or [zu]). Recall that /g/ and /b/ are quite often subject to lenition, being realized as voiced fricatives. However, the lenition in question is optional, being a matter of degree, and not dependent on the identity of the immediately following vowel. In contrast, the allophonic alternations exhibited by /t, d/ are obligatory and systematic, conditioned by the identity of the immediately following vowel; /ti/, /tu/, and /ta, te, to/ differ not only with respect to the vowel but also with respect to whether the consonant is an affricate/fricative or a stop (i.e., in /ti/ and /tu/, the consonant is an affricate, whereas in /ta, te, to/, it is a stop), and similarly for /di/, /du/

before /t/ as in /teQtoː/ 'steel tower', and as [tekˀ] before /k/ as in /teQkin/ 'reinforcement steel'.

and /da, de, do/ (i.e., in /di/ and /du/, the consonant is either an affricate or a fricative, whereas in /da, de, do/, it is a stop). Thus, if vowel perception based on consonants coarticulated with vowels should be examined, stimuli of the form /tV, dV/ are rather inappropriate; manner differences of the consonants would be confounded with the vowel differences. For this reason, stimuli involving /k/, /g/, /p/, /b/ will be used in the experiments conducted for the thesis, but not stimuli involving /t/ or /d/.

2.5 Vowel Devoicing

In Japanese linguistics it has long been assumed that high vowels /i, u/ often get "devoiced" in certain phonemic contexts. This assumption originally came from linguists' intuition but has since been examined experimentally (Gaber & Vance, 2000: Kondo, 1997; Ogasawara, 2013; Varden, 1998) or through a corpus-based study (Maekawa & Kikuchi, 2005).[9] The devoicing contexts can be characterized in terms of the voicing of the neighboring consonants. The usual assumption is that the vowels to be devoiced should be flanked by voiceless consonants (or by a voiceless context and a pause), but the voicing of the preceding consonant seems to exert a stronger effect on vowel devoicing than the voicing of the following consonant. According to Maekawa & Kikuchi's (2005) corpus results, the devoicing rates for high vowels are about 89 % if they are surrounded by voiceless consonants, about 17 % if they are preceded by a voiceless consonant and followed by a voiced consonant, and about 1 % if the preceding consonant is voiced (irrespective of the voicing of the following consonant). Thus the preceding consonants seem to be more influential than the following consonants, which accords with the traditional intuition that morae constitute some fundamental units (in some way) in Japanese; a voiceless consonant exerts its devoicing-inducing effect more easily within a mora than across a moraic

[9] Maekawa & Kikuchi (2005) used a large-scale speech database, called The Corpus of Spontaneous Japanese, containing about 7.5 million words of monologues (academic presentation speech on the one hand, and simulated public speaking on the other) spoken by native speakers of so-called Standard, or Common, Japanese.

boundary.[10]

"Devoicing" in Japanese is, speaking precisely, the weakening of the physical realization of a vowel, where the "weakening" is *not* quality neutralization (centralization) or undershooting of supralaryngeal articulatory gestures, but rather vowel duration shortening, sometimes resulting in a total loss (Beckman, 1996; Faber & Vance, 2000; Kondo, 1997; Ogasawara, 2013; Varden, 1998).[11] Ogasawara's (2000:11) list the following three possible cases of "devoicing" in Japanese:

> **deleted:** After the relaese of closure, no acoustic property of the vowel but only frication noise is observed. Coarticulation cues (palatalilzation, when the 'devoiced' vowel is /i/, rather than /u/) are left behind mostly with the preceding consonant, which indicates the underlying presence of the vowel.
>
> **true devoiced:** Neither voice bar nor periodic waves were observed, but there are faint formants visible enough to characterize the vowel.
>
> **low amplitude reduced but voiced vowel:** A voice bar, formant structures, and a small number of periodic waves with very low amplitude are observed, but the acoustic quality of the vowel is quite different from the fully voiced vowel.

The ratio of the 'deleted' cases vs. the other two cases (combined) was 5:1 in Yuen's (2000) production experiments and even 23:1 in Ogasawara's (2013) shadowing experiments. For example, see Figure 2.1; in the bottom panel, the affricate frication is immediately followed by the closure silence of the following velar stop.

[10] Also note that, in the traditional Japanese grammar, morae were simply unanalyzed wholes, not decomposed into consonants and vowels, and hence the voiceless/voiced distinction was a distinction among morae, not among consonants. Thus, for example, given /tabi/ 'travel', it was /ta/, rather than /t/ alone, that was called voiceless ('*sei-on*' in the traditional grammar terminology), and it was /bi/, not /b/ alone, that was called voiced ('*daku-on*'). Such an analysis could be a direct result of the intuition of the traditional grammarians, who were native speakers, or an indirect result of the mora-based orthographies, which were invented by native speakers a bit more than a thousand years ago.

[11] Kondo (1997) reports that duration shortening does not induce centralization.

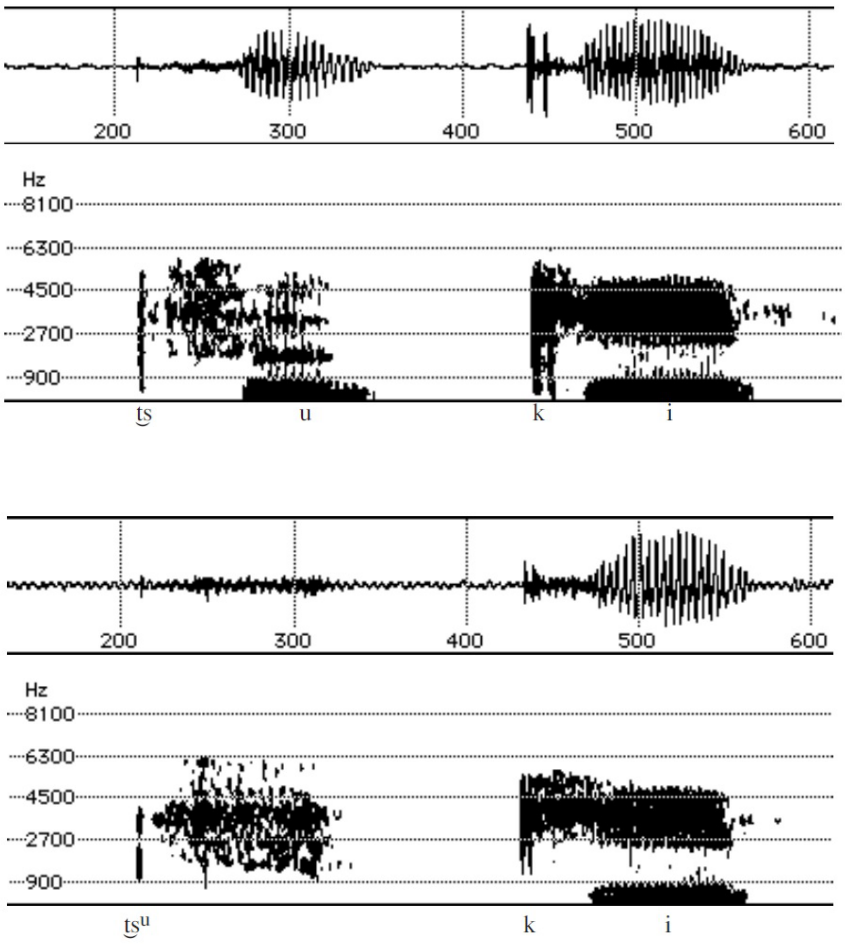


Figure 2.1: Waveforms and spectrograms of a Japanese speaker's productions of *tsuki* 'moon' with voiced /u/ (top panel) and devoiced /u/ (bottom panel) (taken from Varden, 1998:38).

Notably, even when there is no vowel duration, coarticulatory coloring of the vowel is usually present within the immediately preceding consonant, suggesting that articulatory gestures for the vowel are indeed made (and overlap the consonant gestures). For example, see Figures 2.2, which are the waveforms and spectrograms of my own pronunciations of *kisi* 'shore' and *kusi* 'comb', and Figure 2.3, which shows the LPC spectra of the initial 5 ms post-velar-release portions. In Figure 2.2, only one voiced period (corresponding to the vowel in the second mora) is visible in each of the panels; no voicing is visible in the first /ki/ or /ku/ mora. However, the LPC spectra in Figure 2.3 look pretty different, as reflected in the difference between the centers of gravity: 2825 Hz for *kisi* and 1505 Hz for *kusi*; thus the consonantal portions differ depending on the immediately following 'devoiced' vowel.[12] Thus the velar stop in *kisi* differ from that in *kusi*; the former is a velar stop with coarticulation traces of /i/ while the latter is a velar stop with coarticulation traces of /u/. In other words, /-C_1VC_2-/ can be realized either as [-C_1^VVC_2-] or as [-$C_1^V C_2$-], where [C_1^V] refers to [C_1] with coarticulatory coloring of [V].[13] Given that it almost always leaves coarticulation traces within the immediately preceding consonant (Varden, 1998), vowel 'devoicing' in Japanese should be distinguished from true **vowel deletion**; in order for /V/ in /CV/ to be said to have been deleted in the strict sense, the phoneme string should become /C/ and hence its physical realization should be [C] (with no V coloring), while vowel 'devoicing' in Japanese would leave the /CV/ phonemic representations intact, resulting at most in [C^V], but not [C] (with no V coloring), even in the case of what Ogasawara (2013) calls 'deletion'.[14] Thus vowel 'devoicing' in Japanese refers to

[12]For more systematic acoustic studies of coarticulatory traces of a 'devoiced' vowel within the immediately preceding consonant, see Beckman & Shoji (1984) and Tsuchida (1994); Tsuchida observed coarticulation traces not only of the 'devoiced' vowel itself but also of the consonant following the 'devoiced' vowel.

[13]The superscripted V in [C_1^V] is my own notation, although Varden (1998) adopts a similar notation (see Figure 2.1). If /V/ is limited to /i/, the coarticulation could be conceived of as palatalization, in which case it could be written as [C^j], as is done by Vance (1987, 2008). However, as noted below, Nihon Hoosoo Kyookai (1998;227) claims that not only high vowels but also /a, o/ devoice in certain environments. Although the (in)validity of Nihon Hoosoo Kyookai's claim is beyond the scope of this thesis, the superscripted V notation is adopted so that the notation would be general enough to be able to accommodate those cases in which the coarticulation traces could not be conceived of as palatalization.

[14]Kondo's position is very mysterious in this context. She concludes from acoustic measurements that vowel devoicing involves weakening in terms of durations and intensities, but no significant centralization of vowel qualities is observed. However, in Chapter 8 she suddenly jumps to the assumption that vowels are deleted, with some of her

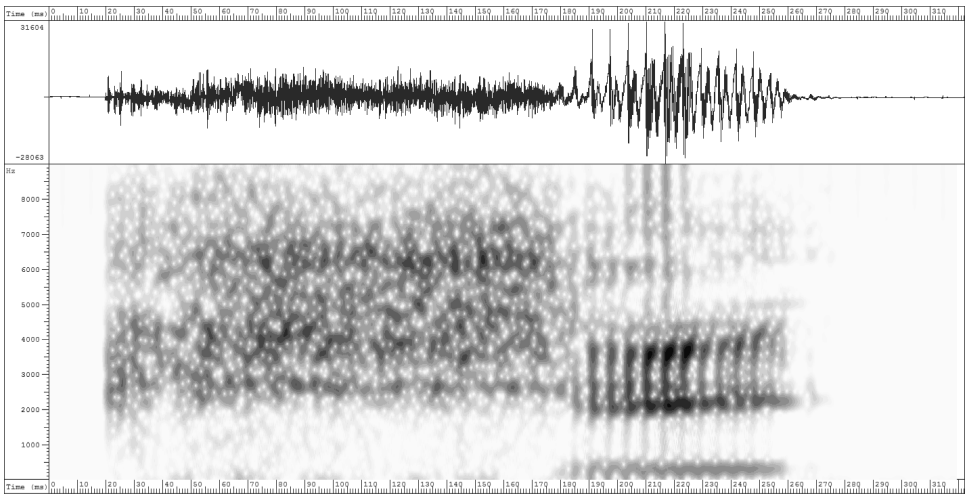


Figure 2.2: The waveforms and spectrograms of my own pronunciations of *kisi* ‘shore’ (the top panel) and *kusi* ‘comb’ (the bottom panel).

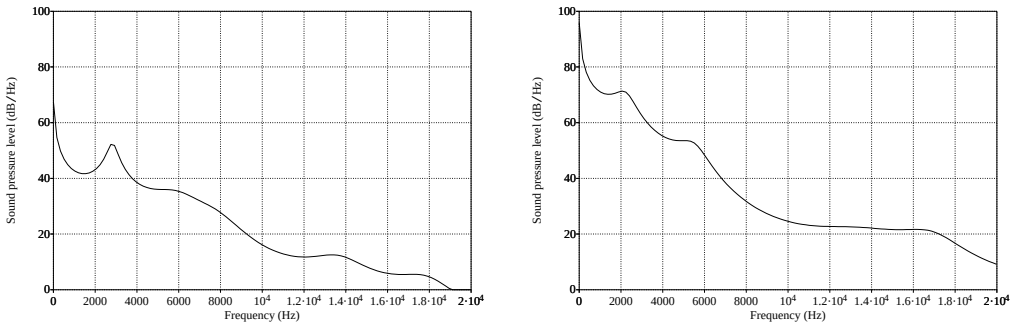


Figure 2.3: The LPC spectra of the initial 5 ms post-velar-release periods of my own pronunciations of *kisi* ‘shore’ (the left panel) and *kusi* ‘comb’ (the right panel).

the fact that vowels can be (i) devoiced or (ii) realized only as coarticulation traces within the preceding consonant, but not deleted in the strict sense.[15]

On the other hand, while it is standardly assumed that it is high vowels /i, u/ that get devoiced, Nihon Hoosoo Kyookai (1998:227) claims that /a, o/ in /ka, ko/ morae tend to get devoiced when the /ka, ko/ morae are word-initial low-pitched morae immediately followed by high-pitched morae. This is simply a claim, not accompanied by any empirical justification, and unfortunately, Maekawa & Kikuchi (2005) did not report the devoicing rates of /a, o/ in such a phonological environment. In contrast, nobody has claimed that /e/ devoices, and indeed, according to Maekawa & Kikuchi's corpus study, the devoicing rate of /e/ is extremely low (1.29 %). Thus the safe conclusion will be that /i, u/ devoice, /e/ does not, while the devoicing possibility for /a, o/ needs more research.

To summarize, in Japanese, typically high vowels "devoice" when flanked by voiceless consonants (or by a voiceless consonant and a pause), where "devoicing" means that the vowel in question becomes either voiceless, or only realized as coarticulation traces (coloring) within the immediately preceding consonant; in either case, coarticulatory traces remain in the immediately preceding consonant (just as when the vowel does not get devoiced), while no voiced portion of the vowel occurs in the acoustic signal.

---

syllable structure descriptions suggesting vowels leave coarticulation traces and others suggesting that they completely delete (with no coarticulation traces), and further argues that a syllable of the form /CCV/ (i.e., a syllable with two consonants in the onset) is phonotactically fine in general in Japanese. Her theoretical motivation for such a jump is clear; she wants to explain the widely accepted belief (which is confirmed by her acoustic measurements) that vowels in consecutive morae do not both devoice.

First she assumes that vowel devoicing results in vowel deletion. She also assumes that the results of devoicing/deletion should obey phonotactics. She further relaxes Japanese phonotactics, so that two-consonant clusters in onsets should be allowed. Now consider /$C_1V_2C_3V_4C_5V_6$/. If only /V_2/ gets devoiced/deleted, the result would be /$C_1C_3V_4$-C_5V_6/ (with the syllable boundary marked with a hyphen); the two-consonant cluster onset would not violate her liberal phonotactics. However, if both /V_2/ and /V_4/ get devoiced/deleted, the result would be /$C_1C_2C_3V_6$/; the three-consonant onset cluster would violate her liberal phonotactics. In general, devoicing/deletion in consecutive morae would result in such three-consonant cluster onsets, and she attempted to account for the ban on devoicing in consecutive morae in terms of the ban on three-consonant cluster onsets.

However, none of her acoustic measurements justify that vowels get deleted in the strict sense.

[15]Nakamura (2003) examined a Japanese speaker's production of devoiced and non-devoiced /su, ki/ syllables with electropalatography (EPG) and electrolaryngography (Lx) and observed that the "lingual gestures for devoiced vowels were retained in the *spatial* domain" although they "were substantially reduced, or shortened, in the *temporal* domain" (p. 55). That is, when /i/ is devoiced, indeed /k/ is manifested as [k^i], rather than a mere [k] with no /i/-coloring.

2.6 Implications for Perceptual Epenthesis

It has long been assumed that the "default" epenthetic vowel in Japanese loanword phonology is /u/. Thus we observe many instances of epenthetic /u/ in loanwords: *desk* in English becomes /desu̲ku̲/, *tab* in English becomes /tabu̲/, *alarm* in English becomes /araːmu̲/, etc. It is also widely assumed that coronal stops and palatal affricates override this default; coronal stops induce /o/ epenthesis (e.g., *try* becoming /to̲rai/), and affricates induce /i/ epenthesis (e.g., *porch* becoming /poːc̟c̟i/).[16]

However, many loanwords involving /k/ betray the expectation from such "rules" (the default for non-coronal stops); for example, indeed /k/ in *Marx*, *weak* or *knock* induce the expected /u/ epenthesis (/marukṵsu/, /wiːku̲/, /noQku̲/), but /k/ in *Mexico* or *text* or *deck* induce the epenthesis of /i/, rather than the expected /u/ (/mekḭsiko/, /tekḭsuto/, /deQkḭ/). The default-overrides with /k/ seem always to be by /i/. Why is this? The above discussion suggests the following obvious candidate answer.

The above discussion of vowel devoicing in Japanese means that /pi, pu/ and /ki, ku/ should often be realized as [p^i, p^u] and [k^i, k^u] respectively;[17] with both /p/ and /k/ being voiceless and both /i, u/ being high vowels. However, /ti, tu/ are *not* realized as [t^i, t^u], not because /i, u/ do not devoice, but rather because /t/ is realized as as affricate before /i, u/.

On the other hand, it is widely known that velar stops exhibit rich place variation and get particularly fronted before /i/ (and to a lesser extent, before /e/) (Vance, 1987; 2008); one possible reason for this is that velar closures and bursts are made with the tongue body and

[16] A usual account of why coronal stops override the default /u/ epenthesis is that /t/ and /d/ are allophonically realized as affricates before /u/; thus the allophonic rules prohibit the default /u/ epenthesis after coronal stops. On the other hand, given that palatal affricate phonemes are realized as palatal affricates before /u/ in Japanese, there does not seem to be an obvious account of why affricates induce /i/, rather than the default /u/, epenthesis; possibly the story concerning voiceless velar stops described below (**DB-sensitivity**) extends to affricates.

[17] It has standardly assumed in Japanese phonology that the Japanese lexicon consists of four different strata: Yamato (native), Sino, foreign, and onomatopoeia (or onomatopoeia and ideophones; Labrune, 2012:13). Based on the observation that non-geminate /p/ does not occur in the Yamato and Sino strata, Cutler et al. (2009) have assumed that vowel devoicing does not apply to /pi, pu/. Given that non-geminate [p] does occur readily in both 'foreign' and 'onomatopoeia' strata and naturally employed in newly invented words, it is not clear why non-occurrence in the Yamato and Sino strata should imply non-devoicing after /p/. However, whether Cutler et al.'s position should be accepted or rejected does not crucially affect the rest of this thesis.

hence are affected much by the tongue position of the immediately following vowel. If such fronting results in salient coarticulation coloring, Japanese listeners are likely to perceive [k^i] as /ki/, which is encountered everyday due to vowel devoicing. For a similar reason, [k^u] must be encountered everyday as a realization of /ku/ and hence should be perceived as /ku/, even if epenthesis played no role. Thus chances are that a voiceless velar burst with sufficiently /i/-like coloring is perceived as /ki/, and a voiceless velar burst with sufficiently /u/-like coloring is perceived as /ku/, at least when the bursts are followed by a voiceless consonant or a pause, as a result of everyday phonemic categorization. Such a consideration readily explains the default-overriding behavior of /k/ in loanwords.

In contrast, /p/ does not seem to induce a similar default-overriding /i/ perception in loanwords. For example, *cake* results in /keiki̲/, but *cape* results in /keipu̲/. Why? In principle, [p^i] could arise as a realization of /pi/. However, since bilabial closures and bursts are *not* made with the tongue, they are not much affected by the tongue position of the vowel, and hence presumably do not carry much coarticulation traces. For example, see Figures 2.4–2.5, in which the waveforms, spectrograms, and LPC spectra of the initial 5 ms post-release portions, of a female native speaker's utterance of *aku, aki, ake* and of *apu, pi, ape* are shown;[18] the centers of gravity computed from the spectra of the initial 5 ms post-release portions were: 3607 Hz (*ki*), 1587 Hz (*ke*), 882 Hz (*ku*); 1911 Hz (*pi*), 1275 Hz (*pu*), and 854 Hz (*pe*). Thus the range of /pi/ (1911 Hz) vs. /pu/ (1275 Hz) was properly contained in the range of /ki/ (3607 Hz) vs. /ku/ (882 Hz), just as was observed by Benneau (1997) for French bursts, and is compatible with the articulation-based assumption that bilabial bursts do not carry as much coarticulation traces as velar bursts.[19]

Then chances are that listeners should not be as good in exploiting the coarticulation traces

[18]The female speaker's utterances were recorded for the abandoned version of the project for this thesis. (Her productions are employed here so that her voluntary efforts with no compensations would not end up in vein.) No apparent devoicing is observed in her productions; the figures, as well as the centers of gravity computed from the spectra, are meant to be example indications of how much coarticulation effects are observed within velars and bilabials.

[19]Benneau's (1997) measurements did not include /ke/ or /pe/.

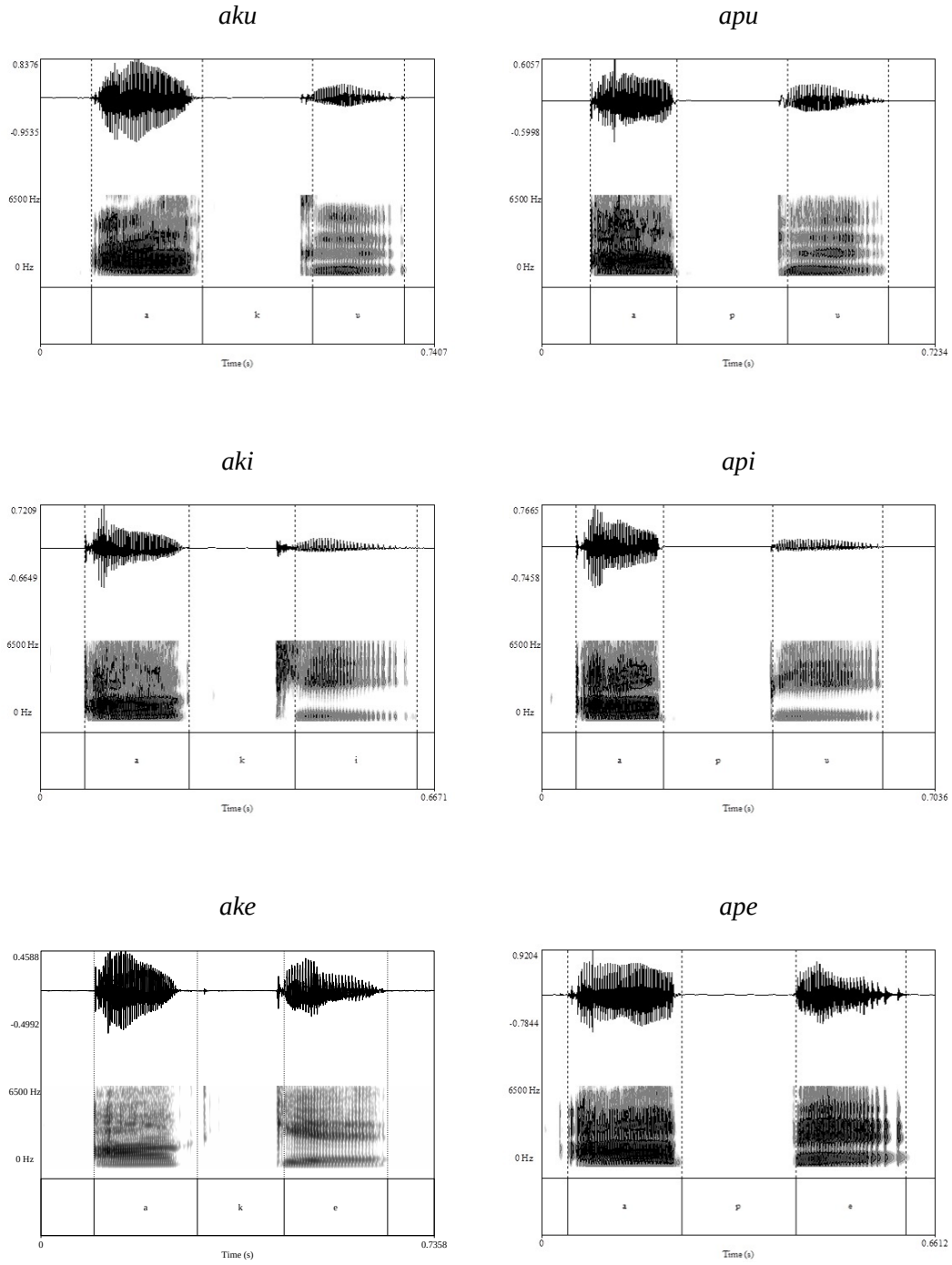


Figure 2.4: Waveforms and spectrograms of a female speaker's utterances of *aku* 'evil' (the left top panel), *aki* 'Autumn' (the left middle panel) , and *ake* (a nonsense word with the high-low pitch; the left bottom panel); *apu* (a non-word; the right top panel), *api* (a non-word; the right middle panel), *ape* (a non-word; the right bottom panel).

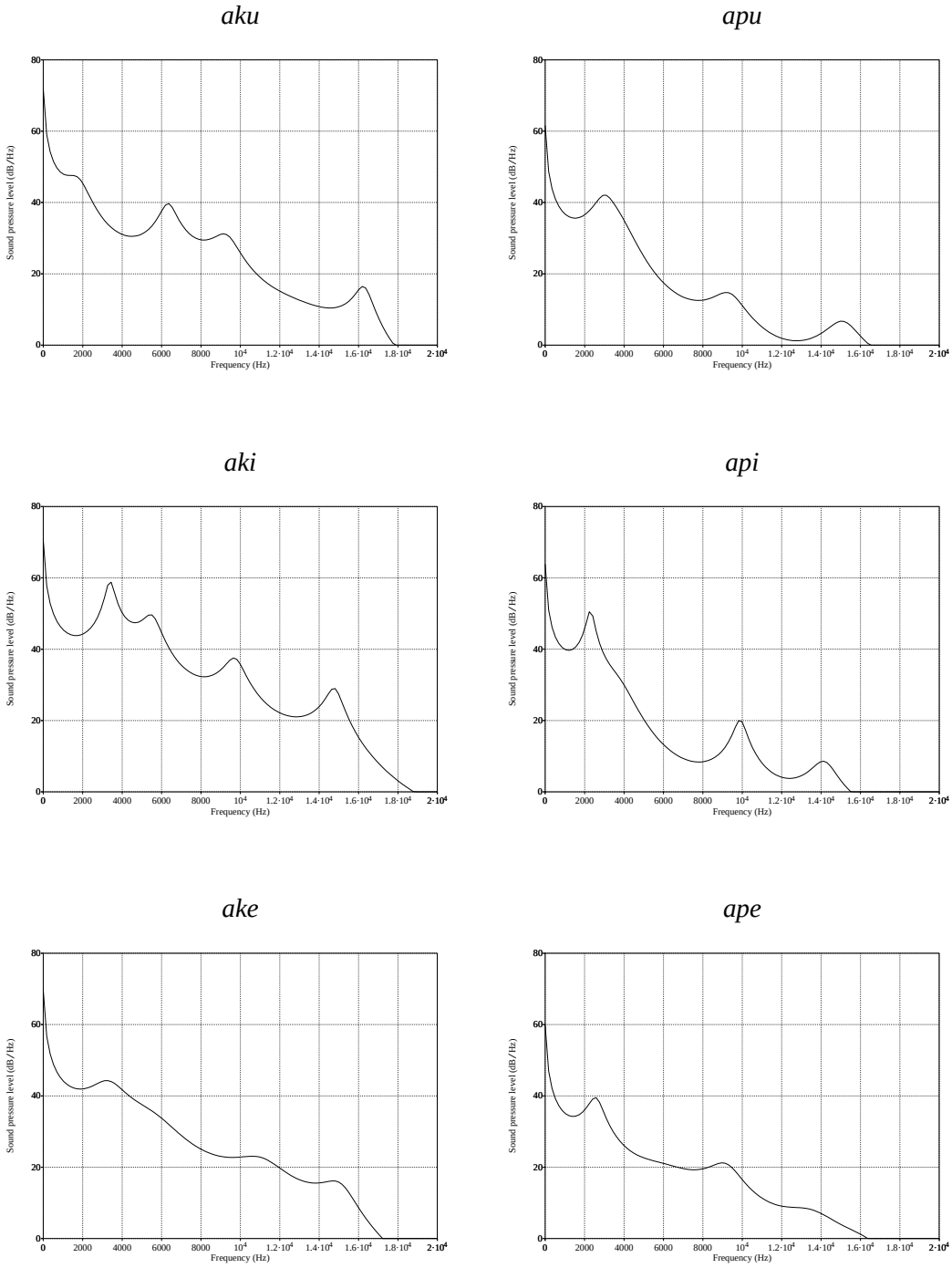


Figure 2.5: LPC spectra of the initial 5 ms post-release portions of a female speaker's utterances of *aku* 'evil' (the left top panel), *aki* 'Autumn' (the left middle panel), *ake* (a nonsense word with the high-low pitch; the left bottom panel); *apu* (a nonsense word; the right top panel), *api* (a nonsense word; the right middle panel), and *ape* (a nonsense word; the right bottom panel).

Table 2.1: Devoicing rates of /pi/ and /ki/, reported by Maekawa & Kikuchi (2005); 'C_1' refers to the immediately preceding consonant, while 'C_2' refers to the immediately following consonant.

C_1	C_2	voiced instances	devoiced instances	the devoicing rate
/k/	/c/	19	62	76.54 %
	/h/	167	65	28.02 %
	/k/	73	476	86.70 %
	/Q/	32	51	61.45 %
	/s/	144	262	64.53 %
	/t/	53	791	93.72 %
/p/	/Q/	118	9	7.09 %

of vowels within bilabials, in which case [p^i] is simply subject to the default /u/ epenthesis, resulting in /pu/ perception. If so, we would expect that devoicing of /i/ in /pi/ tends to be avoided in natural productions by Japanese speakers so that they would not be misheard by listeners. Maekawa & Kikuchi's (2005) corpus results, the relevant portions are presented in Table 2.1, suggest that that is indeed the case; the devoicing rate for /pi/ is considerably lower than that for /ki/.[20] The above observations that (i) [k^i] is a legitimate vowel-devoiced realization of /ki/, and (ii) [k^i] presumably carries enough perceptual cues for /i/, explain the observation of /i/ epenthesis after /k/ in loanwords, if 'sufficiently /i/-like coarticulation' is nothing but *front* coarticulation (putting aside whether 'sufficiently /u/-like coarticulation' coincides with back coarticulation or not); the voiceless velar stops inducing /i/ epenthesis in loanwords are presumably fronted, being next to a front vowel, and hence cues an /i/ to Japanese ears (Takehiko Makino, p.c., 1996).

On the other hand, the assumptions that (i) [p^i] is a legitimate vowel-devoiced realization of /pi/, but (ii) [p^i] presumably does not carry much perceptual cues for /i/, are compatible with the uniform /u/ epenthesis observed after /p/ in loanwords. It is not necessarily that [p^i] should totally fail in inducing /i/ perception; it is just that, presumably, the /i/ perception tends to lose the default /u/ perception, because the /i/-inducing cue is poor (especially when the vowel

[20] As stated above, Cutler et al. (2009) assume that non-geminate /p/ blocks vowel devoicing and attribute the assumed lack of devoicing after non-geminate /p/ to the fact that /p/ does not often appear in native-stratum words. Devoicing of /i/ after /p/ is rather unexpected both from the above discussion and from Cutler et al.'s assumption.

duration is much reduced).

Note that this **devoicing-based phonemic categorization** story suggests the possibility that /ki/ perception induced by [k^i] should not be seen as an instance of epenthesis (perception of something absent in the physical signal) but rather should be seen simply as a result of everyday phonemic categorization. It is not news that, in general, the acoustics-to-phoneme mappings are not one-to-one; for example, one cue may function as a cue for more than one phoneme. Thus it is not a particularly implausible idea that [k^i] cues /k/ as well as an immediately following /i/; the resulting phoneme sequence /ki/ would cause no phonotactic violation in the first place, and hence the /i/ perception should not be seen as an instance of perceptual epenthesis. Thus the vowel devoicing characteristics of Japanese suggest that /i/ perception based on an /i/-coarticulated consonant could be interpreted as a result of coarticulation sensitivity exerted in phonemic categorization (**DB-sensitivity**), instead of coarticulation sensitivity exerted in phonotactic repair (**DI-sensitivity**).

2.7 Chapter Summary

The following characteristics of Japanese will play vital roles in the following chapters.

Japanese disallows non-homorganic obstruent clusters (e.g., /bz/ or /kt/). However, due to the 'devoicing' of high vowels /i, u/ between voiceless consonants (or between a voiceless consonant and a pause), what could be perceived as such clusters by non-native listeners do arise in natural productions, although the coarticulation traces of the 'devoiced' high vowel do remain within the preceding consonant (e.g., [k^it]). Since [C^V] realizations of /CV/ are often heard, where C is a voiceless consonant and V is a high vowel, there is the possibility that [C^V] is perceived by Japanese listeners as /CV/ not as a result of perceptual epenthesis (phonotactic repair) but rather as phonotactics-independent phonemic categorization.

When no coarticulation traces of vowels are found in the preceding consonants, /u/ is

epenthesized by default after those consonants which do not exhibit manner alternations after a high vowel (i.e., non-coronals). Thus a velar (but not dental) stop/plosive with no coarticulation traces of a following vowel tends to be perceived as /ku, gu/. The default status of /u/ epenthesis means that Japanese listeners' /Cu/ perception from [C^u] could be interpreted either in terms of coarticulation-insensitive default /u/ epenthesis, in terms of coarticulation-sensitive /u/ epenthesis (**DI-sensitivity**), or in terms of coarticulation-sensitive phonemic categorization (**DB-sensitivity**); thus it is not very informative with respect to whether phonemic categorization or phonotactic repair is coarticulation-sensitive. However, that in turn means that Japanese listeners' /Ci/ perception from [C^i] (where C is not a coronal) could only be interpreted as evidence for some kind of coarticulation sensitivity.

The reality of coarticulation-based /Ci/ perception, as well as its nature (i.e., whether it should be seen as a result of phonotactic repair or of phonemic categorization) is an empirical issue to be examined later, but given the assumption that velars exhibit richer coarticulation-based place variations than bilabials, such coarticulation-based /Ci/ perception, if real, is expected to be observed with velars, rather than with bilabials. On the other hand, since coronals exhibit manner alternations before a high vowels, Japanese listeners' /Ci/ perception from a coronal with /i/ coarticulation could be interpreted in terms of their sensitivity to the manner of the consonants, rather than to coarticulation. Thus coronals are not appropriate consonant stimuli in an examination of Japanese listeners' coarticulation sensitivity, whether it is devoicing-based or not.

On the other hand, while there is an agreement that /e/ does not exhibit a tendency to devoice, it is not clear whether /a, o/ also exhibit a tendency to devoice (in certain phonological environments). Thus /a, o/ are not appropriate vowel stimuli in an examination of phonemic categorization exhibiting **DB-sensitivity**.

Chapter 3

Literature Review

3.1 Introduction

This chapter critically reviews arguments for or against two-step models, one-step models and lexicalist models of speech perception, and attempts to see to what extent available evidence in the literature distinguishes the three types of models, which offer different answers to the question of how coarticulation sensitivity, phonotactic sensitivity and lexical sensitivity relate to each other

Section 3.2 reviews classical literature on phonotactic sensitivity. Section 3.3 reviews some arguments for and against lexical reductions of phonotactic sensitivity (Dupoux et al., 2001; Fais et al., 2005; Mazuka et al., 2011). The possibility of Japanese listeners' **DB-sensitivity** was suggested by the production characteristics of Japaneses (section 2.6 above). It will be seen that, if it is indeed perceptually real, **DB-sensitivity** would explain away the alleged evidence for a lexicalist reduction of phonotactic sensitivity. Since Dupoux et al.'s and Mazuka et al.'s evidence against a lexicalist reduction of phonotactic sensitivity would remain intact, the available evidence would then favor a non-lexicalist account of phonotactic sensitivity, provided that **DB-sensitivity** is real.

The primary aim of Section 3.4 is to review evidence concerning the reality of **DB-sensitivity**,

as well as an alternative, lexicalist account. Subsection 3.4.1 reviews classical attempts to demonstrate **DB-sensitivity** (Beckman & Shoji, 1984; Tsuchida, 1994); it will be seen that the attempts were not fully successful in the sense that the observed sensitivity could alternatively be interpreted as **DI-sensitivity**. Subsection 3.4.2 examines a more recent and more successful demonstration of **DB-sensitivity** (Ogasawara & Warner, 2009), which (under the interpretation employed in this thesis) suggests that it is real both sublexically and lexically (in the sense to be explicated below, which favors one- and two-step models and disfavors lexicalist models) and, in turn, suggests the need for a revision to the Merge model (Norris et al., 2000; Norris & McQueen, 2008), a revision compatible with Ogasawara's (2013) observations. However, it is only *more* successful than the classical attempts as a demonstration of **DB-sensitivity**, not a complete success. Subsection 3.4.3 critically reviews Cutler et al.'s (2009) claim of a lexicalist reduction of vowel perception from (what they analyze as) a consonant in general, which would imply a lexicalist reduction not only of **DB-sensitivity** but also of phonotactic sensitivity; it also reviews Kingston et al.'s (2011) results, cited as supporting evidence by Cutler et al.; it will be seen that neither Cutler et al.'s or Kingston et al.'s results entail Cutler et al.'s conclusion. Section 3.4 as a whole thus suggests the possibility that **DB-sensitivity** is real both sublexically and lexically, which would defend Dupoux et al.'s (2001) and Mazuka et al.'s (2011) claim of the sublexical reality of phonotactic sensitivity.

Section 3.5 examines evidence concerning the choice between two-step models and one-step models. Subsection 3.5.1 conceptually examines one-step models and points out that they could be implemented in two different versions (**suprasegmental matching** and **slot filling**). Subsection 3.5.2 examines what Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) and Dupoux et al. (2011) saw as potential pieces of evidence against one-step models; it will be pointed out that they are in fact compatible with both versions of one-step models. However, it will also be pointed out that Matthews & Brown's (2004) results are rather incompatible with the **suprasegmental matching** version (but not with the **slot filling** version) of one-step mod-

els. Subsection 3.5.3 examines Dehaene-Lambertz et al.'s (2000) and Dupoux et al.'s (2011) argument for one-step models; it will be seen that they are inconclusive. Section 3.5 as a whole thus suggests that the choice between one-step models and two-step models is still open.

Section 3.6 concludes the chapter, illustrating what has to be examined in order to distinguish among two-step models, one-step models, and lexicalist models.

3.2 Classical Studies on Phonotactic Effects

Speech deviating from native norms is frequently assimilated perceptually to native categories (cf. Best et al., 1988). Such assimilation induced by cross-linguistic differences in phonemic inventories has been studied extensively. However, such assimilation is not limited to single phonemes; assimilation induced by cross-linguistic differences in phonotactics has been exemplified in the literature.

Brown and Hildum (1956) observed that American English listeners' transcription of phonotactically admissible pseudo-words (e.g., /prol/) was better than of phonotactically inadmissible non-words (e.g., /tlib/), and the error patterns (unspecified in the paper) suggest assimilations of illicit phoneme strings to licit strings.

Massaro and Cohen (1983) (mentioned in Chapter 1) obtained clearer instances of phonotactically induced assimilation. They observed that the identification function for an /l/–/r/ continuum varied depending on whether it was in /s_V/ and /t_V/ contexts; the continuum was more easily perceived as /l/ in the former and as /r/ in the latter, which was interpreted in terms of the legality of /sl/ and /tr/ and the illegality of /sr/ and /tl/ in "syllable-initial" position. They also compared /bl/, /br/, /dr/, and */dl/ clusters and found that, while the first consonants within the clusters affect the perception of the second, the second also affect the perception of the first. They regarded this finding as evidence against the view that speech perception is a process of perceiving individual phonemes in a serial fashion and for the view that syllables are

the discrete units against which the incoming speech signal is matched (p. 347), presupposing that the relevant phonotactics is a set of syllable-level constraints sanctioning such onsets as /sl/ but banning such onsets as /sr/.[1]

Adapting Best's terminology (Best et al., 1988), Segui et al. (2001:198) call such phonotactically induced assimilation "structural perceptual assimilation" and classified them into the following three cases (with my own label for each in bold):

(1) cases where cues for (or acoustic correlates of) phonotactically illicit phonemes are simply ignored. (**perceptual deletion**)[2]

(2) cases where phonemes not articulated and hence with no cues or acoustic correlates in the signal are perceived. (**perceptual epenthesis**)

(3) cases where a phoneme string is perceived as another phoneme string. (**perceptual conversion**)

For the ease of reference, let's call them 'perceptual deletion', 'perceptual epenthesis', and 'perceptual conversion' respectively, as indicated above.

3.2.1 Perceptual Conversion

In the above terminology, Massaro and Cohen's (1983) observations of phonotactics-sensitive discrimination functions (mentioned above) will be classified as instances of perceptual conversion. Further instances of perceptual deletion or epenthesis will be illustrated below.

Another example of perceptual conversion is provided by Hallé et al. (1998). They examined French listeners' perception of stop-liquid clusters in a word- or syllable-initial position. The tendency to transcribe phonotactically illegal /dl/ and /tl/ as phonotactically legal /gl/

[1] However, it is not necessarily clear whether their presupposed assumption is correct. According to some phonologists and its experimental endorsement by Treiman et al. (1992), a seemingly syllable-initial /s/ in English is a "stray" segment (Blevins, 1995:223ff) and does not belong to a syllable, no matter whether followed by /l/ or /r/ (or any other phoneme; see the references in Berent et al., 2007:594), while according to Morelli's (2003) OT-based argument, /s/ + stop clusters form a syllable onset cross-linguistically.

[2] 'Ignored' may be too strong; alternatively, listeners are less sensitive to such cues, in which case it should be called **reduced perceptual sensitivity**.

and /kl/ was observed, as compared to other stop-liquid clusters which do not violate French phonotactics, a pattern replicated in a forced choice task. They further observed, in a gating experiment, that listeners' initial dental judgments of /dl, tl/ were later revised into velar judgments of /gl, kl/; when /l/ was perceived, their initial dental judgments changed into velar judgments, which was interpreted as a phonotactic effect. Their perception of velars in such clusters was further confirmed in a phoneme-monitoring experiment, and a slower RT to the target /k/ in /tl/-stimuli (which was already observed to be perceived as /kl/) than in /kl/-stimuli was interpreted as suggesting that the assimilation (conversion, in the above terminology) of the illegal /tl/ clusters to legal /kl/ clusters incur an additional cost of processing (Segui et al., 2001). Such assimilation (conversion) was also observed in lexical decision (*ibid.*); phonotactically illegal nonwords tended to be judged to be words (68.3 %) significantly more often than phonotactically legal nonwords were so judged (19.6 %), suggesting that the perceptual assimilation from phonotactically illegal ones to legal ones cannot be explained away in lexical terms. Furthermore, in a cross-modal repetition priming experiment, priming effects in lexical decision on visual target real words were observed when the phonotactically illegal auditory primes could be perceptually assimilated to the visual targets, to the same degree as when the auditory primes coincided with the visual targets.

3.2.2 Perceptual Deletion

Experimental results described by Kakehi et al. (1996) could be seen as perceptual deletion, if 'perceptual deletion' could be understood in a broader sense so as to include 'reduced sensitivity' (cf. footnote 2 on page 32). The perceptual cues for the place of C, where C is either /p/, /t/ or /k/, are distributed in the speech signal: for /V_1CV_2/, the preclosure formant transitions, the burst, and the postclosure formant transitions all offer such cues. They observed Japanese listeners' disadvantage, as compared to Dutch listeners, in exploiting the preclosure cues for the stop place, a difference naturally interpreted in terms of whether /V_1C/ is legal (Dutch) or ille-

gal (Japanese); only when it is legal in the native language do the listeners need to rely solely on the preclosure cues for stop place. It was not that Japanese listeners were totally unable to utilize such cues, but the disadvantage they exhibit demonstrates phonotactics-dependent reduction of sensitivity to perceptual cues.

3.2.3 Perceptual Epenthesis

Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) observed perceptual epenthesis. They examined the perceptual nature of the epenthesis of /u/, often found in loanword phonology in Japanese. As seen in Chapter 2, Japanese bans obstruent clusters except (1) in the case of geminates (i.e., /Q/ followed by a consonant), or (2) when the first of the two consecutive consonants is the moraic nasal /N/. A usual "repair strategy" to cope with obstruent clusters violating such a phonotactic constraint is to insert a vowel, typically /u/, between the obstruents. In Dupoux, Kakehi, Hirose, Pallier, & Mehler's experiments, auditory stimuli containing obstruent clusters violating the constraint (e.g., *ebzo*) were presented to Japanese and French listeners, and they observed that Japanese listeners tended to report the perception of /u/ between the obstruents significantly more often than French listeners, which was interpreted in terms of the fact that Japanese, but not French, bans such clusters. Their cross-linguistic design demonstrated the language-dependent nature of epenthesis, as opposed to possible "universal properties of phonetic perception" (p. 1569) independent of L1 phonotactics. The observation of Japanese listeners' failure, as opposed to French listeners' success, in speeded ABX discriminations (between, say, *ebzo* and *ebuzo*), a task involving no production nor an explicit mention of the vowel, was interpreted as arguing against a production account of epenthesis (i.e., against the idea that an epenthetic vowel was reported due to listeners' internal articulation of the stimuli), as well as an orthography account (i.e., against the idea that an epenthetic vowel was reported due to listeners' knowledge of Japanese orthographies, which do not allow such consonant clusters), on the assumption that discriminations would not involve listeners' 'inner speech' or

conscious categorization.[3]

3.3 Arguments for and against Lexicalist Reductions of Phonotactic Sensitivity

As reviewed above, Massaro & Cohen (1983) observed that the identification function for /l/–/r/ continuum is dependent on phonotactic permissibility. Massaro & Cohen interpreted this result in terms of phonotactics-driven perceptual assimilation. However, McClelland & Elman (1986) proposed an alternative lexical effect interpretation of this observation, according to which the /l/–/r/ continuum tended to be perceived as /l/ before /s/ than before /t/, in Massaro & Cohen's experiments because, given that there are many words beginning with /sl/ but none with /tl/, the phoneme perception units for /sl/ receives stronger feedback from the lexicon than those for /tl/ does.[4] Given that phonotactics are theoretically derived from the lexicon as sublexical regularities, a lexical effect interpretation could similarly be proposed for supposedly phonotactics-driven epenthesis, such as the ones observed by Dupoux, Kakehi, Hirose, Pallier, & Merler (1999), reviewed above. For example, /...bz.../ is unattested in words but /...buz.../ is attested in actual words (e.g., *buzoku* 'tribe'), and hence Japanese listeners might tend to perceive *ebzo* as /ebuzo/, as observed by Dupoux, Kakehi, Hirose, Pallier, & Merler (1999), because /ebuzo/ is more similar to actual words than /ebzo/. Thus a natural question, given Dupoux, Kakehi, Hirose, Pallier, & Merler's (1999) results, would be whether the observed perceptual epenthesis is sublexical or lexical in nature.

A mixture of observations favoring and disfavoring lexicalist reductions of phonotactic sensitivity is found in the literature; an account that explains all of those observations is called for.

[3] Indeed, there remains the possibility that 'internal articulation' or orthographic patterns have been 'hard-wired' into listeners' perceptual behavior, in which case the results of perceptual experiments are affected indirectly by "internal articulation" or orthographic patterns even when the task does not explicitly involve internal articulation or orthography. However, whether such an indirect effect of "internal articulation" or orthography should be seen as real is beyond the scope of this thesis.

[4] See Massaro & Cohen (1991) for a reply.

The goal of this section is to review those observations and point out that, if **DB-sensitivity** is assumed to be perceptually real (either sublexically or lexically), it would resolve the seemingly conflicting observations, in favor of the sublexical reality of phonotactic sensitivity.

3.3.1 A Potential Argument for a Lexicalist Account of Phonotactic Sensitivity

Fais et al. (2005) argued that Dupoux, Kakehi, Hirose, Pallier, & Mehler's (1999) results should be accounted for in terms of the frequencies of the attested sound patterns, rather than discrete phonotactic constraints. They studied well-formedness ratings for auditory stimuli by Japanese listeners, who were shown target nonword (e.g., *neeku* /niːku/) in an orthographic form and evaluated auditory stimuli (e.g., *neek* [niːk] or *neeku* [niːku]) as pronunciations of the visually presented nonword. A phonotactic account, according to Fais et al., leads us to expect that an epenthetic vowel should be perceptually inserted even in potentially vowel-devoicing environments and hence that the stimuli without a vowel (called 'noncanonical forms') should be perceived as having an epenthetic vowel and should receive the same ratings as the stimuli with a vowel (called 'canonical forms'). However, they observed that noncanonical forms (e.g., *neek* /niːk/) were rated consistently lower than canonical forms (e.g., *neeku* /niːku/). Moreover, they observed that different noncanonical forms received different ratings (a higher rating for *keets* than for *neek*), which was attributed to different likeliness of vowel devoicing (more frequent vowel devoicing after an affricate than after a stop). They claimed that those results argue against a phonotactic account of epenthesis.

Under Fais et al.'s (2005) own interpretation, such results suggest that Japanese listeners' vowel perception from a consonant should not be seen as a result of phonotactic constraints (as claimed by Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999) but rather as a reflection of how often the sound patterns are encountered. If such frequencies could be taken as the likeliness of the sound patterns as possible words, their results seem to favor lexicalist models. However, there are two problems with the view that such results argue for a lexicalist reduction

of phonotactic sensitivity.

First, vowel perception based on 'noncanonical forms' could be due to devoicing-based phonemic categorization of auditory [C] as /CV/, to a phonotactic repair epenthesizing /V/, or to both. In other words, there exists the possibility that, while phonotactic sensitivity is real as sublexical sensitivity, as a result of which [C] does give rise to a /CV/ percept, devoicing-based familiarity with attested sound patterns operates as an *additional* factor, as a result of which different /CV/ percepts (with /V/ epenthesized by sublexical phonotactic repair) differ with respect to the goodness as exemplars of /CV/. Thus their results do not necessarily argue against the sublexical reality of phonotactic sensitivity.

The second problem with Fais et al.'s (2005) experiment concerns their unjustified assumption that the 'epenthesized' vowel should always have been the one they assumed, i.e., /u/. They observed that the auditory stimuli *neek* ([niːk], according to their transcription) received a lower rating than the auditory stimuli *neeku* ([niːkɯ]), both evaluated against the visually presented form *neeku* /nɪːku/, where the stimuli were uttered by a bilingual English-Japanese speaker. However, possibly, /k/ in such environments was fronted, in which case, without a stronger cue for the backness of the following vowel supplied by a following vocalic /u/ portion in *neeku*, /k/ in *neek* might have given a /ki/ percept (**DB-sensitivity**), contradicting the visually presented form *neeku*. Thus the lower rating of *neek* could well be due to the conflict between /ki/ percepts and the visually presented form /ku/, a possibility they failed to notice. Thus, although they interpreted listeners' different ratings as evidence for their sensitivity to the distinctions between the presence vs. the absence of an epenthetic vowel (/u/), it is also possible to interpret the ratings rather as evidence for the coarticulation-sensitive nature of their vowel perception (**DB-sensitivity**). Similarly, if [k] in *neek* carries more front coarticulation cues than [ts] in *neets*, with only the former tending to induce /i/ percepts, a lower rating for auditory *neek* as a realization of visually presented *neeku* than for auditory *keetsu* as a realization of visually presented *keetsu* would also follow, because of the conflicts between perceived /i/ and printed

/u/ in the former. Remember Kakehi et al.'s (1996) remark that phonemic cues are temporally distributed in the speech signal. The possible existence of such a cue for a front vowel within the /k/ burst is another example of such temporal distribution, but ironically Fais et al. failed to note such complexity of acoustics-to-phoneme mapping, in spite of their emphasis on "acoustic forms."

With those two problems, Fais et al.'s (2005) results would not necessarily constitute evidence against the existence of sublexical phonotactic constraints.

3.3.2 Arguments against a Lexicalist Reduction of Phonotactic Sensitivity

In contrast, Mazuka et al. (2011) and Dupoux et al. (2001) argued for the sublexical nature of perceptual epenthesis.[5] However, Dupoux et al. also report seeming counterexamples to their claim of the sublexical nature of perceptual epenthesis. As discussed below, Dupoux et al.'s own account of the seeming counterexamples has shortcomings and probably invalid, and hence, without a better account of the seeming counterexamples, their claim of the non-lexical nature could not be accepted. In fact, **DB-sensitivity** (if real) would offer a better account of the seeming counterexamples. That means that, if **DB-sensitivity** is assumed to be real, their argument for the sublexical nature of phonotactic sensitivity (perceptual epenthesis) could be defended.

Mazuka et al. (2011) asked whether the acquisition of the phonological grammar, responsible for the kinds of perceptual epenthesis by Japanese listeners as observed by Dupoux, Kakehi, Hirose, Pallier, and Mehler (1999), should be seen as through the lexicon (i.e., through the comparison of underlying lexical representations and surface word forms) or prelexical (i.e., incomplete but robust fragments of the native phonology can be bootstrapped from a bottom-up analysis of distribution of segments). They examined the discriminations between /ebzo/- vs. /ebuzo/-type stimuli by Japanese-learning (and control French-learning) infants at the ages of

[5] Dupoux, Fushimi, Kakehi, & Mehler (1999) is a previous version of Dupoux et al. (2001).

8 and 14 month old and observed the successful discriminations by the 8 month old age group on the one hand, and the unsuccessful discriminations by the 14 month old group on the other, a result that suggests that the phonological grammar (responsible for perceptual epenthesis) is acquired between 8 and 14 months of age. Assuming that the vocabulary size of a 14 month old infant is around 40, they argue that the phonological grammar could not be acquired through a large lexicon at this age and hence should be seen as acquired prelexically.

In fact, Jusczyk et al. (1993) and Juszyk and Luce (1994) had already observed that, between 6 and 9 months, English-learning infants develop a preference for phoneme sequences that would count as phonotactically licit (non)words over those that would count as illicit. Presumably, the vocabulary size must be smaller between 6 and 9 months than around 14 months, Jusczyk et al.'s (1993) and Juszyk and Luce's (1994) observation could also be taken as evidence for the pre-lexical nature of phonotactic sensitivity *in general*, in which case Mazuka et al.'s (2011) observations should be seen as evidence for the pre-lexical nature *specifically of* perceptual epenthesis by Japanese listeners.

While Mazuka et al.'s (2011) argument was based on the developmental pattern, Dupoux et al. (2001) attempted to argue for the non-lexical nature of perceptual epenthesis based on experimental results with adult listeners. They compared two groups of nonwords containing phonotactically illegal obstruent clusters which had only one lexical neighbor. One group, called the u-Set, consisted of those whose lexical neighbors can be obtained if /u/ is epenthesized (e.g., *sokdo*, with the neighbor *sok*u*do* 'speed), while the other group, called the non-u-Set, consisted of those whose lexical neighbors can be obtained if a non-/u/ vowel is epenthesized (e.g., *mikdo*, with the neighbor *mik*a*do* 'emperor'). They examined Japanese listeners' transcriptions and found that the rates of /u/ epenthesis did not differ across the two sets, although a lexical neighborhood effect account would lead us to expect a significantly higher rate of non-/u/ epenthesis in the non-u-Set. In short, whether the results of epenthesis would constitute a real word did not have an effect. They also examined responses to a lexical judgment task, with

which lexical effects should be larger, and observed that (1) the u-Set members tended to elicit a "word" response (though not as often as real words with no need for epenthesis, used as controls), (2) the non-u-Set members clearly tended to be judged as "nonwords," (3) the u-Set members were responded to as fast as real words with a physical vowel portions, and (4) the responses to the non-u-Set members were *not* slower as responses to nonwords with a physical vowel portion. If perceptual /u/ epenthesis is a lexical effect, the lexical neighbors should be activated by the non-u-Set members and hence should elicit "word" responses, against the (2) observation; the activations of the lexical neighbors should make "nonword" responses to the non-/u/-Set members slower than nonwords with a physical vowel portion, against the (4) observation. Again, whether the results of epenthesis would constitute a real word did not have an effect. Thus those observations point toward the non-lexical nature of /u/ epenthesis.

One notable aspect of Dupoux et al.'s (2001) results, however, is that /i/ transcription was observed for *reksi* and *riksi* in the non-u-Set, resulting in the existing words *rekisi* and *rikisi* respectively. If the non-lexical nature of phonotactic sensitivity leads us to expect the default /u/ epenthesis, they would constitute counterexamples. Dupoux et al. resisted a lexicalist interpretation of this result by pointing out that other stimuli in the non-u-Set did not exhibit a non-/u/-epenthesis so that the resulting percepts would constitute an existing word, and suggesting that the results with *reksi* and *riksi* should be interpreted in terms of the "independent observation" that voiceless stop-fricative clusters tend to elicit /i/-epenthesis (pp. 500–501). It is not clear at all where that "independent observation" came from, and it is rather hard to accept, in light of the fact that many such clusters do induce /u/ epenthesis in loanwords (e.g., /taQk̲u̲su/ 'tax', /faQk̲u̲su/ 'fax', /riraQk̲u̲su/ 'relax', /poteto tiQpusu/ 'potato chips', /pepusi/ 'Pepsi'). Even if it was accepted, that would still leave 11 of their stimuli (including *reksi* and *riksi*), out of 38, as candidates for /i/-epenthesis, and hence we would be left with no account of why those two (but not others) elicited /i/ responses.

If Dupoux et al.'s (2001) own account of *reksi* and *reksi* is not viable, and if there is no

alternative account, the /i/ transcriptions observed with those stimuli would constitute evidence for a lexicalist reduction of phonotactic sensitivity (while the results with the other non-/u/-set stimuli argue against a lexicalist reduction). However, there does exist an alternative, straightforward account, a **DB-sensitivity** account, according to which the voiceless velar stops in *reksi* and *riksi* were fronted and hence induced /i/ perception as a result of devoicing-based phonemic categorization, rather than as a result of phonotactic repair. In fact, among their stimuli, *reksi* and *riksi* were exactly those in which /k/ was probably fronted. Thus their argument for the non-lexical nature of perceptual epenthesis (or vowel perception based on consonants) would be valid if **DB-sensitivity** is perceptually real and employed as an account of the seeming counterexamples. Also note that, putting aside those loanwords whose origins are presumably non-perceptual (Irwin, 2011; Smith, 2006),[6] default-overriding /i/ 'epenthesis' after /k/ in loanwords does seem to accord with the expectation from **DB-sensitivity**; /i/ 'epenthesis' when /k/ is presumably fronted (e.g., /keːki̲/ 'cake'; /deQki̲/ 'deck'; /suteːki̲/ 'steak'), in contrast to /u/ 'epenthesis' when /k/ is presumably not fronted (e.g., /boQku̲su/ 'box'; /ruQku̲/ 'look').[7]

To summarize this subsection, the overall patterns of the results by Mazuka et al. (2011) and Dupoux et al. (2000) suggest the non-lexical nature of perceptual epenthesis by Japanese listeners. However, Dupoux et al. also observed potential counterexamples. While Dupoux et al.'s own account of them is rather inviable, **DB-sensitivity** offers a better account, linking the potential counterexamples to loanword patterns and Japanese listeners' everyday linguistic experience, thereby more successfully defending Dupoux et al.'s claim of the non-lexical nature of perceptual epenthesis.

Note that, besides the relative success or failure, Dupoux et al.'s own account on the one

[6]Irwin (2011) traces the historical origins of many loanwords philologicaly and finds that usually an 'expert' of the source language dictates a particular epenthetic pattern based on his or her own phonemic analysis (not his or her own perception) of the source language, and the general public not familiar with the source language through auditory media simply followed. Such doublets as /eki̲supuresu/ vs. /eku̲supuresu/ 'express' (with the latter being preferred by bigger companies) or /teki̲suto/ vs. /texu̲suto/ 'text' (with the latter being a very new form) presumably reflect such dictations; as far as I am aware, most Japanese teachers of English dictate /u/ epenthesis after /k/, and the /u/ versions presumably reflect such dictations.

[7]Note that the word-final velars are not followed by a fricative in /keːki̲/ 'cake' or in /deQki̲/ 'deck' but yet induce /i/ 'epenthesis'.

hand, and the **DBsensitivity** account on the other, support the non-lexical nature of perceptual epenthesis in rather different ways. For Dupoux et al., *reksi* and *riksi* should both violate phonotactics and hence be subject to perceptual epenthesis; their 'voiceless stop–fricative' account claims that the observed /i/ perception should be seen as an instance of epenthesis which overrides the default thanks to the particular consonantal phonemes. In contrast, according to the **DB-sensitivity** account, the observed /i/ perception should be seen as phonemic categorization of fronted velar bursts and hence *not* as an instance of epenthesis, because *reksi* and *riksi* should not violate phonotactics (when the velar bursts are phonemically categorized as /ki/); thus /i/ perception from the velar stop in *reksi* or *riksi* is not a result of phonotactic repair and hence should not bear on the question of whether perceptual epenthesis (phonotactic repair) is lexical or not.[8]

3.3.3 Summary of the Section

This section has reviewed Fais et al. (2005), Mazuka et al. (2011) and Dupoux et al. (2001). Mazuka et al.'s (2011) observations, as well as Dupoux et al.'s (2001) observations except the *reksi* and *riksi* results, argue against a lexicalist reduction of phonotactic sensitivity. On the other hand, Fais et al.'s (2005) results, as well as Dupoux et al.'s (2001) *reksi* and *riksi* results, seemingly argue for a lexicalist reduction, but it was pointed out that they only constitute unconvincing evidence, particularly if **DB-sensitivity** is perceptually real. Thus, provided that **DB-sensitivity** is perceptually real, the observations by Fais et al., Mazuka et al., and Dupoux et al. would not conflict; a coherent interpretation would be that phonotactic sensitivity is sublexical, while phonemic categorization results due to **DB-sensitivity** were wrongly counted as instances of phonotactic repair.

[8] This account is orthogonal to the question of whether such phonemic categorization (based on **DB-sensitivity**) is lexical or sublexical (a question to be discussed later). The point here is simply that, if the fronted velar bursts are subject to **DB-sensitivity** (which may or may not be lexical in nature), the *reksi* and *riksi* results should not be seen as results of phonotactic repair, in which case they would be simply irrelevant to the question of whether phonotactic repair is lexical or sublexical.

3.4 Evidence for DB- (rather than DI-)Sensitivity

This section reviews evidence (mostly) concerning **DB-sensitivity**, as opposed to **DI-sensitivity**. Subsection 3.4.1 reviews classical attempts to demonstrate its reality (Beckman & Shoji, 1984; Tsuchida, 1994). The attempts were successful only in the sense that Japanese listeners' coarticulation sensitivity was confirmed; they were unsuccessful in the sense that they did not show that the coarticulation sensitivity in question is devoicing-based. Subsection 3.4.2 reviews more successful evidence for **DB-sensitivity** provided by Ogasawara & Warner's (2009) observations; under the interpretation employed in this thesis (but not in theirs), their observations suggest that **DB-sensitivity** is real both sublexically and lexically, i.e., its lexicon-independent effect and lexicon-dependent effect are both real; however, their results still leave some (small) room for an interpretation in terms of **DI-sensitivity**. Subsection 3.4.3 critically reviews Cutler et al.'s (2009) claim of a lexicalist account of vowel perception from consonants in general, which would imply a lexicalist account not only of **DB-sensitivity** but also of phonotactic sensitivity (against Dupoux et al., 2001, and Mazuka et al., 2011); it also reviews Kingston et al.'s (2011) results, cited as supporting evidence by Cutler et al. It will be argued that neither Cutler et al.'s or Kingston et al.'s results entail Cutler et al.'s conclusion. Thus this section as a whole suggests that **DB-sensitivity** is real both sublexically and lexically, while defending the sublexical reality of phonotactic sensitivity (as concluded in the previous section).

3.4.1 Classical Potential Evidence for DB-Sensitivity

The defense in the previous section of Dupoux et al.'s (2001) and Mazuka et al.'s (2011) argument for the sublexical nature of phonotactic sensitivity relies on the assumed perceptual reality of **DB-sensitivity**. This section reviews classical attempts to demonstrate the perceptual reality of **DB-sensitivity**; it will be seen that the attempts were successful as demonstrations of Japanese listeners' coarticulation sensitivity, but rather unsuccessful as demonstrations that it

is devoicing-based.

Beckman & Shoji (1984) presented Japanese listeners with natural productions of [ɕu̥] and [ɕi̥] on the one hand, and synthetic versions of [ɕ] created by modeling the fricative portions of [ɕu] and [ɕi] on the other.[9] They observed Japanese listeners' significantly successful identification of the devoiced or missing vowels /i/ or /u/.

Similarly, Tsuchida (1994) excised the [ɕ] portions[10] from natural speech of /ɕi/ and /ɕu/, which were presented to Japanese listeners; again, the listeners exhibited significantly successful identification of the missing vowels /i/ or /u/.

Both Beckman & Shoji (1984) and Tsuchida (1994) interpreted the observed successful vowel identifications in terms of **DB-sensitivity**. However, strictly speaking it does not necessarily have to be, because the observed identifications could be taken as a reflection of devoicing-independent vowel recovery from coarticulation cues (**DI-sensitivity**). One crucial difference between /i/ perception from a consonant due to **DB-sensitivity** on the one hand, and /i/ perception from a consonant due to **DI-sensitivity** on the other, is that the former is a result of phonemic categorization of front coarticulation, which does not necessarily have to be due to a following /i/, whereas the latter is a result of the *recovery* of the missing /i/ after the consonant from the coarticulation traces in the consonant; in other words, the former does not necessarily require an underlying /i/ after the consonant, while the latter is crucially dependent on an underlying /i/ after the consonant. Since Beckman & Shoji's and Tsuchida's results could be interpreted either as due to phonemic categorization of front coarticulation or as recovery of the missing /i/, they could be interpreted either as evidence for **DB-sensitivity** or as evidence for **DI-coarticulation sensitivity**.

That means that their results are not enough to validate the argument suggested in the previous section for the sublexical nature of phonotactic sensitivity, an argument which crucially

[9] Beckman & Shoji (1984) used [ʃ] as the IPA transcription symbol for the fricative, but [ɕ] should be more appropriate (Kamiyama, 2008; Vance, 1987; 2008).

[10] Tsuchida also transcribes ɕ] as [ʃ].

relied on the perceptual reality of **DB-sensitivity**. Recall that, according to the argument, /k/ in Fais et al.'s (2005) *neek* stimuli, in Dupoux et al.'s (2001) *reksi* and *reksi* stimuli, or in the sources for such loanwords as /deQki/, tended to be phonemically categorized as /ki/, in which case no phonotactic repair would be involved. However, in none of those was /k/ followed by an underlying /i/; thus, in order for that argument to be validated, the reality of **DB-sensitivity** (which would induce /ki/ perception irrespective of whether an underlying /i/ exists after /k/) should be demonstrated, rather than the reality of **DI-sensitivity** (which could induce /ki/ perception only when an underlying /i/ followed /k/). To the extent that they could be interpreted in terms of **DI-sensitivity**, rather than **DB-sensitivity**, Beckman & Shoji's (1984) and Tsuchida's (1994) results (with voiceless fricatives, rather than voiceless velar stops) are not enough to validate the argument in the previous section for the sublexical nature of phonotactic sensitivity.

3.4.2 The Lexical/Sublexical Nature of DB-Sensitivity

Stronger evidence for the perceptual reality of **DB-sensitivity** is provided by Ogasawara & Warner (2009). Although their experiments and arguments have various problems (as will be seen below), some of their observations suggest, under the interpretation employed in this thesis, that **DB-sensitivity** is real both sublexically and lexically (i.e., its non-lexically-driven and lexically-driven aspects are both real), which in turn suggests the need for a revision to the Merge model (Norris et al., 2000; Norris & McQueen, 2008), as seen below.

Ogasawara & Warner (2009) conducted two /i/ monitoring experiments and two lexical decision experiments, all with native listeners of Japanese.[11] Since the stimuli in the lexical decision experiments constituted real words only if a medial /i/ was perceived, all the experiments examined listeners' /i/ perception (in a phonemic processing context in the /i/ monitoring experiments; in a lexical processing context in the lexical decision experiments).

[11] They also conducted one /i/ monitoring experiments with English listeners, which is ignored in this review.

The purpose of Ogasawara & Warner's (2009) experiments was to compare the effects of the following three factors on the perception of /i/ by Japanese listeners:

(A) allophonic appropriateness (devoiced /i/ being more appropriate than non-devoiced /i/ between voiceless consonants; the other way around in other contexts)

(B) acoustic strength (weaker cues and shorter durations for devoiced /i/ vs. stronger cues and longer durations for non-devoiced /i/)

(C) phonotactics (phonotactics-driven sensitivity to the existence of a devoiced /i/ within consonant clusters)

According to Ogasawara & Warner, the 'vowel' portions in their 'vowel-devoiced' stimuli (in their term, "reduced stimuli") lacked not only voicing but also formant structures, and they explicitly say (p. 378) that vowels are 'deleted' in such stimuli (where 'deletion' is clearly meant to be deletion in a broad sense, i.e., an underlying /i/ with coarticulation traces in the acoustic signals). Thus (B) is somewhat hard to understand. If their descriptions of the stimuli are correct, (B) should rather read: 'only coarticulation traces within the preceding consonants vs. vowels in addition to coarticulation traces'. Given the default status of /u/ epenthesis for all of their stimuli, then, successful /i/ perception should be seen as reflecting the listeners' coarticulation sensitivity.

Broadly classified, two kinds of stimuli in three environments were compared:[12]

[12] As seen above, according to Ogasawara & Warner's (2009) descriptions of the "reduced stimuli", voewls are 'deleted', with coarticulation traces within the preceding consonants. Thus the superscript notation [C_1^i] notation is employed for their "reduced stimuli" in this review. Since such coarticulation traces must exist even when vowels are voiced, the superscript notation is also employed for their "unreduced stimuli."

(1) devoicing environment:

a. [...C_1^ii C_2V...] (e.g., /hokito/), where both /C_1/ and /C_2/ are voiceless, and /i/ is produced as a voiced vowel. (“unreduced stimuli”)

b. [...$C_1^i$$C_2$V...] (e.g., /hokito/), where both /C_1/ and /C_2/ are voiceless, and /i/ is ‘devoiced’. (“reduced stimuli”)

(2) voicing environment:

a. [...C_1^ii C_2V...] (e.g., /tadʒiga/), where both /C_1/ and /C_2/ are voiced (and not nasal), and /i/ is produced as a voiced vowel. (“unreduced stimuli”)

b. [...C_1^i C_2V...] (e.g., /tadʒiga/), where both /C_1/ and /C_2/ are voiced (and not nasal), and /i/ is ‘devoiced’. (“reduced stimuli”)

(3) nasal environment:

a. [...C_1^ii C_2V...] (e.g., /kedanida/), where /C_1/ is a nasal, /C_2/ is voiced, and /i/ is produced as a voiced vowel. (“unreduced stimuli”)

b. [...C_1^i C_2V...] (e.g., /kedanida/), where /C_1/ is a nasal, /C_2/ is voiced, and /i/ is ‘devoiced’. (“reduced stimuli”)

The natural productions by Ogasawara, who is a native speaker of Tokyo Japanese, were used as stimuli with no significant editing.[13]

Ogasawara & Warner (2009) expected that the strengths of the effects of **allophonic appropriateness** and of **acoustic strengths** could be compared by examining listeners’ relative ease (RTs and error rates) with **unreduced stimuli** vs. with **reduced stimuli** in the **devoicing environment**. Because the production grammar dictates that /i/ should be ‘devoiced’ in the **devoicing environment**, **reduced stimuli** should sound more natural than **unreduced stimuli**

[13]Speaking more precisely, cross-splicing was applied to half of the stimuli in their Experiment 3 and Experiment 5, but they report that it had no significant effect on the results.

in the **devoicing environment**. Thus, as an effect of **allophonic appropriateness**, we would expect that /i/ pereption should be easier in **reduced stimuli** and than in **unreduced stimuli**, in the **devoicing environment**. In contrast, as an effect of **acoustic strength**, we would expect that /i/ perception should be easier in **unreduced stimuli** than in **reduced stimuli** (in whatever environment). Since the expectations from **allophonic appropriateness** and from **acoustic strengths** are opposite in the **devoicing environment**, they reasoned that the strengths of their effects could be compared by examining whether /i/ perception turns out to be easier or harder in **reduced stimuli** than in **unreduced stimuli** in the **devoicing environment**. For example, if the effect of **allophonic appropriateness** is stronger than that of **acoustic strengths**, easier /i/ perception should be observed in [hok^{i}to] than in [hok^{i}ito]; if the effect of **allophonic appropriateness** is weaker than that of **acoustic strengths**, easier /i/ perception should be observed in [hok^{i}ito] than in [hok^{i}to].

On the other hand, the **nasal environment** is meant to examine the effects of **phonotactics**; Ogasawara & Warner (2009) reasoned that the **nasal environment** stimuli should contrast with the **devoicing** and **voicing environment** stimuli in that /i/ perception failure should not induce a phonotactic violation in the **nasal environment**, because the underlined consonant cluster in /kedaNda/ with no /i/ between /N/ and /d/ is phonotactically fine in Japanese, in contrast to the underlined cluster in /hokto/ or /tagda/; thus phonotactics should encourage listeners' sensitivity to the existence of a devoiced /i/ within the **reduced stimuli** in the **devoicing** and the **voicing environment**, but not in the **reduced stimuli** in the **nasal environment**.

The general pattern observed in their experiments were the following (with two exceptions to be discussed below), where the success of /i/ perception is measured in terms of RT's and miss rates:[14]

[14] In fact, there were three, rather that two, exceptions. The one that is not discussed below is that the differences of miss rates between the **reduced** and the **unreduced stimuli** did not significantly differ in the **voicing environment** in Experiment 1 if "outliers" are included in the analysis, but obeyed the general pattern if they are excluded from the analysis. Ogasawara & Warner (2009) seem to favor the analysis with "outliers," but if they are indeed "outliers," it is not clear why they should be included in the analysis. Furthermore, even if their inclusion was admitted methodologically, it is not clear what theoretical advantage they would thereby gain. Thus this 'exception'

- The perception of /i/ was equally successful with the **reduced** and the **unreduced stimuli** in the **devoicing environment**.

- The perception of /i/ was significantly less successful with the **reduced stimuli** than with the **unreduced stimuli** in the **voicing** and the **nasal environment**.

- The perception of /i/ with **reduced stimuli** was more successful in the **devoicing environment** than in the **voicing environment** on the one hand, and in the **voicing environment** than in the **nasal environment** on the other.

Before considering what implications those results would have with respect to one-step, two-step and lexicalist models, first let us consider Ogasawara & Warner's (2009) own interpretations.

In Ogasawara & Warner's (2009) interpretation, /i/ perception was equally successful with the **reduced** and the **unreduced stimuli** in the **devoicing environment** because the effect of (B) (= acoustic strength), discouraging /i/ perception with the **reduced stimuli**, and the effects of (A) (= allophonic appropriateness) and (C) (= phonotactics), encouraging /i/ perception with the **reduced stimuli**, canceled each other; if |X| refers to the magnitude of the effect of a factor X, then,

$$|B| = |A|+|C|$$

with B being a negative effect, while A and C being positive effects, on /i/ perception. Also, in their interpretation, /i/ perception was less successful with the **reduced stimuli** than with the **unreduced stimuli** in the **voicing environment** because the effects of (A) (= allophonic appropriateness) and (B) (= acoustic strength), each discouraging /i/ perception with the **reduced stimuli**, won over the effect of (C) (= phonotactics), encouraging /i/ perception with the **reduced stimuli**, that is,

is ignored here.

$$|A| + |B| > |C|$$

Finally, /i/ perception was less successful with the **reduced stimuli** than with the **unreduced stimuli** in the **nasal environment** because the effect of (A) (= allophonic appropriateness) and (B) (= acoustic strength) both discouraged /i/ perception with the **reduced stimuli**, while the effect of (C) (= phonotactics) was not applicable, with the underlined consonant cluster in /kedaNda/ with no /i/ between /N/ and /d/ being phonotactically fine, that is,

$$|A| + |B| > 0$$

In short, the following approximations of the magnitudes of the effects of the three factors follow from such an interpretation:[15]

$|B| = |A| + |C|$ (from the devoicing environment results)

$|A| + |B| > |C|$ (from the voicing environment results)

$|A| + |B| > 0$ (from the nasal environment results)

Under such interpretations, more successful /i/ perception with the **reduced stimuli** in the **devoicing environment** than in the **voicing environment** seems to follow; the magnitude of the negative effect of reduction (devoicing) on /i/ perception performance in the **devoicing environment** can be approximated by

$$|B| - (|A| + |C|) = 0$$

i.e., a negative effect of (B) (= acoustic strength), minus

positive effects of (A) (= allophonic appropriateness) and (C) (= phonotactics)

while that in the **voicing environment** can be approximated by

$$(|A| + |B|) - |C|$$

i.e., negative effects of (A) (= allophonic appropriateness)

and (B) (= acoustic strengt),

minus a positive effect of (C) (= phonotactics)

[15] Such approximations in the form of arithmetic formulas are my own reformulations of their interpretations.

which, with substitution of |B| with |A| + |C|, amounts to

$$|A| + |(|A| + |C|) = 2|A|$$

and hence the negative effect in the **devoicing environment** (= 2|A|) should be greater than the negative effect in the **devoicing environment** (= 0); similarly, the negative effect in the **nasal environment** can be approximated by

$$|A| + |B|,$$

i.e., negative effects of (A) (= allophonic appropriateness)
and (B) (= acoustic strength)

which, again with the substitution of |B| with (|A| + |C|), amounts to

$$|A| + (|A| + |C|) = 2|A| + |C|$$

which is greater than the negative effect in the **voicing environment** (= 2|A|), and hence the more unsuccessful /i/ perception with the **reduced stimuli** in the **nasal environment** than in the **voicing environment**.

Unfortunately, however, there are several reasons to question the validity of their interpretation. For one thing, the above calculations presuppose that the /i/ perception performance is a direct reflection of the negative effect of 'reduction' ('devoicing'), which in turn presupposes that the 'baseline' performance with the **unreduced stimuli** did not differ significantly across the three environments. However, those presuppositions are not confirmed; their graphs rather suggest that the 'baseline' performance did differ across the three environments. If the 'baseline' performance with the **unreduced stimuli** differs across the three environments, a quantitative comparison of the /i/ perception performance with the **reduced stimuli** across the three environments could not be interpreted solely in terms of the effects of (A), (B) and (C) in the three environments.

Of course, that does not mean that qualitative comparisons between /i/ perception performance with the **reduced** vs. the **unreduced stimuli** across the three environments is mean-

ingless. To repeat, as such qualitative comparisons, they report that /i/ perception was equally successful with **reduced** and **unreduced** stimuli in the **devoicing environment**, but was less successful with **reduced stimuli** than with **unreduced stimuli** both in the **voicing** and the **nasal environment**. The problem of unequal 'baseline' performance with **unreduced stimuli** does not make their interpretation invalid with respect to such qualitative observations. However, there exists another reason to doubt their interpretation of the **nasal environment** results.

Because 'environment' and 'reduced/unreduced' are meant to be orthogonal, the **reduced** and the **unreduced stimuli** in the **nasal environment** should differ only with respect to whether /i/ is devoiced or not, which was indeed the case according to their transcription; the nasal was [ɲ], which is an allophonic realization of /n/ in front of /i/. (Our notation [n^{i}] should be understood as [ɲ].) However, in order for the **nasal environment** to function in the way they intended it to, we have to assume that [ɲ] in a **reduced stimulus** is phonemically categorized as /N/, which is a distinct phoneme from /n/, as seen in Chapter 2 and is admitted by Ogasawara & Warner (2009); only if it was categorized as /N/ would the **nasal environment reduced stimuli** conform to phonotactics without the perception of a vowel after the nasal. The problem is that /N/ could be realized as [ɲ] only when it is immediately followed by another palatalized consonant; as seen in Chapter 2; /N/'s place of articulation is assimilated to the following consonant (when there is one). Speaking more specifically, [ɲ] is a realization of (i) /n/ before /i/, (ii) /nj/ before any vowel, and (iii) /N/ before another palatalized consonant; /N/ is realized as [m] before a bilabial (e.g., /seNpoː/ → [sempoː] 'strategy'), as [n] before an alveolar (e.g., /keNri/ → [kenɾi] 'rights'), as [n̪] before a dental (e.g., /seNtoː/ → [sen̪n̪toː] 'battle'), as [ŋ] before a velar (e.g., /teNki/ → [teŋki] 'weather'), and [ɲ] only before another [ɲ] (e.g., /seNniN/ → [seɲɲiĩ] 'thousand people; full-time; Taoism Xian').[16]

[16]The /n/ phoneme in /seNniN/ is realized as [ɲ] because it is immediately before /i/; the first /N/ assimilates its place to this palatalized nasal.

On the other hand, the word-final /N/ is not followed by a consonant. It is transcribed here as a nasalized vowel, following Kamiyama (2008) etc.. McQueen et al. (2001:109) assume that /N/ is realized as [ŋ], rather than a nasalized vowel, in word-final position. However, which is correct is not relevant in this discussion.

However, in none of Ogasawara & Warner's (2009) **nasal environment reduced stimuli** is the nasal followed by another palatalized consonant. Thus Japanese listeners could phonemically categorize the [ɲ]'s in the **nasal environment reduced stimuli** only if they were insensitive to the allophonic inappropriateness of [ɲ]'s. In fact, as noted above, [ɲ] is not only an /i/-conditioned allophone of /n/ but also the only allophone for /nj/ (irrespective of the following vowel), and loanwords from French suggest the possibility of Japanese listeners' sensitivity to palatalization of a nasal, inducing the default /u/ epenthesis: /ɕanpaːnjɯ̲/ 'campagne', /burugoːnjɯ̲/ 'bourgogne', etc. Possibly, the acoustic properties varying with nasal place contrasts are not particularly prominent (Raphael, 2005:196) and hence Japanese listeners may be rather insensitive to palatalization of nasals, but possibly, as suggested by loanwords from French, they may be sensitive; which is indeed the case is an empirical question, an answer for which should not be assumed a priori.[17]

Thus the above considerations leave as valid only the observations that /i/ perception was (i) equally successful with **reduced** and **unreduced stimuli** in the **devoicing environment** but (ii) was less successful with **reduced stimuli** than with **unreduced stimuli** in the **voicing environment**. (If the nasal in the **reduced stimuli** in the **nasal environment** is categorized as /ɲ/, rather than /N/, the **nasal environment** is nothing but a special kind of **voicing environment**, and (ii) subsumes it.) Note that vowel perception in question is that of /i/, rather than the default

[17] Otake et al. (1996) examined Japanese listeners' perception of [m, n, n̠, ŋ] realizations of /N/ and observed that (i) monitoring for a nasal was faster and more accurate for various realizations of /N/ than for /n/, (ii) mismatch of the realizations of /N/ and the following contexts results in lower naturalness ratings, slower monitoring for a nasal, and slower and more erroneous monitoring for the post-nasal consonant, and (iii) the faster and more accurate nasal monitoring for /N/ than for /n/ persists even when the realizations of /N/ mismatched with the following contexts. The (ii) observation suggests Japanese listeners' sensitivity to place variations within nasals, but the (iii) observation suggests that the sensitivity does not prevent /N/ perception. However, they did not examine neither [ɲ] or whether a given nasal is more easily categorized as /N/ or /n/.

Also note that the physical realizations of /N/ tend to be considerably longer (possibly more than twice as long) than those of /n/ (2.39 times on average in Sato's 1993 results; 2.38 times on average in Otake et al.'s 1996 results). In Otake et al.'s (1996) stimuli, the durations of the nasals (both /n/ and /N/) were kept intact as produced by a native speaker, and hence presumably the /N/ portions were quite (and significantly) longer than the /n/ portions; thus the durations is likely to have encouraged listeners to interpret the various realizations of /N/ as /N/ (the *moraic* nasal) rather than /n/ (or non-moraic onset consonants in general). However, since the nasal portions in the reduced **nasal environment** stimuli were produced as [ɲi̯], their durations are likely to have been rather short, possibly functioning as a cue for the non-moraic status for the nasal portions.

If Otake et al.'s (1996) (ii) observation is combined with the duration considerations above, it is rather unlikely that Ogasawara & Warner's [ɲ] portions were categorized as /N/.

epenthetic /u/; thus the (i) observation (i.e., equally successful /i/ perception with with **reduced** and **unreduced stimuli** in the **devoicing environment**) suggests that coarticulation-based /i/ perception is as easy as vowel-based /i/ perception, and the contrast between the **devoicing environment** and the **voicing environment** suggests that the coarticulation sensitivity in question is **DB-sensitivity**, rather than **DI-sensitivity**.

Now we are ready to ask the question of whether **DB-sensitivity** is lexical or sublexical in nature. As noted above, the above characterization of the general pattern of their overall results have two exceptions. The two exceptions in question seem to point toward the conclusion that **DB-sensitivity** is perceptually real both lexically and sublexically.[18]

One of the exceptions concerns Ogasawara & Warner's (2009) manipulation in the second /i/ monitoring experiment and in the second lexical decision experiment. Recall that their stimuli were generally of the form [...C_1^ii C_2V...] (unreduced stimuli) and [...C_1^i C_2V...] (reduced stimuli). They prepared two kinds of **voicing environment** unreduced/reduced stimuli: those in which C_1 is voiced but C_2 is voiceless (**pre-voiced stimuli**), and those in which C_1 is voiceless but C_2 is voiced (**post-voiced stimuli**). They classified those two kinds of stimuli under **voicing environment** based on the assumption that devoicing-based perception should be possible only if both C_1 and C_2 are voiceless. This assumption is based on the following two ideas:

> **The voicing requirement in production:** Vowels could devoice only if both C_1 and C_2 are voiceless.
>
> **Production-perception symmetry:** Devoicing-based perception mirrors devoicing patterns in production.

However, the reduced **post-voiced stimuli** resulted in no RT or miss rate difference in /i/ monitoring from the unreduced **post-voiced stimuli**, patterning with the **devoicing environment**

[18] Ogasawara & Warner (2009) themselves seem to favor a lexicalist interpretation of their overall results (see footnote 20 below), but the lexicalist interpretation was only suggested without an argument, and their reasoning has a conceptual flaw (again, see footnote 20 below).

results, while the results with the **pre-voiced stimuli** exhibited a trend toward the general pattern of the **voicing environment**, i.e., larger RT's and miss rates with the reduced than the unreduced stimuli.[19] In order to make sense of the **post-voiced stimuli** results, Ogasawara & Warner say that "when the following consonant is what makes the environment a voicing environment, listeners may detect [/i/] before they have enough information about the following consonant to realize that vowel reduction is inappropriate" (p. 392). Putting aside the dubious assumption that phonemes are processed one by one in a sequential fashion, this claim amounts to admitting that voiceless C_1 suffices for devoicing-based perception irrespective of the voicing of C_2; if C_2's suppression of devoicing-based perception is too late, C_2 simply can not have an effect. To say the least, observationally, **post-voiced stimuli** exhibited the **devoicing environment** pattern, and the unbiased interpretation would be that it is the voicing of C_1, but not of C_2, that affects the possibility of devoicing-based /i/ monitoring.

This result betrays the expectation from the combination of 'the voicing requirement in production' and 'production-perception symmetry'. Thus at least one of them must be wrong. This result itself does not decide which is wrong, but there exists a reason to believe that it is the a priori assumed 'production-perception symmetry' that is wrong, as will be seen below.

The second exception to the general pattern (no effect of reduction/devoicing in the **devoicing environment** on the one hand, and with **post-voiced stimuli** on the other; worse perception of /i/ in **reduced stimuli** than in **unreduced stimuli** in the **voicing environment** except with **post-voiced stimuli**) concerns the results of lexical decisions, which exhibited a slightly different pattern, with respect to the **devoicing environment** on the one hand, and with respect to the **post-voiced stimuli** on the other. While /i/ monitoring was equally successful across the **reduced** and the **unreduced stimuli** in the **devoicing environment**, lexical decision based on /i/ perception was observed to be *easier* with **reduced stimuli** than with **unreduced stimuli**;

[19] Ogasawara & Warner (2009) attributed the failure to reach significance to the smaller number of the **pre-voiced stimuli**, which was half of the number of the **voiced environment** stimuli in Experiment 1.

On the other hand, the **post-voiced** stimuli exhibited the **voicing environment** pattern in lexical decision in Experiment 5, which will be discussed below.

reduction/devoicing had a positive effect in the **devoicing environment** when the task is lexical decision. The difference between /i/ monitoring and lexical decision concerning the **devoicing environment** could be attributed to the fact that /h/ was employed as C_1 only in the /i/ monitoring experiments. However, that would not explain the observed difference between **reduced** and **unreduced stimuli** in lexical decision. Rather, a more promising alternative interpretation would be to assume that the mental lexicon contains prototypical phonetic realizations of words, and lexical decision involves the matching between the incoming speech signals with those stored prototypical realizations. If /i/ is devoiced in the stored prototypical realizations of those words (e.g., /akikan/, prototypicaly stored as [ak^{i}kan]) in the **devoicing environment**, the **reduced stimuli** should result in a better match than the **unreduced stimuli**. Adopting this interpretation, let us call the easier lexical decision with devoiced vowels "the prototype effect."

Ogasawara & Warner's (2009) lexical decision results seem to help us decide why voiceless C_1 was observed to lead to devoicing-based perception of /i/ even when C_2 is voiced in their /i/ monitoring experiments. As noted above, the observed asymmetry between the perceptual effect of the voicing of the preceding consonant (C_1) and of the following consonant (C_2), could be interpreted in two ways: (i) vowels do 'devoice' *in production* as long as C_1 is voiceless, even if C_2 is voiced, and (ii) indeed vowels resist devoicing *in production* when C_2 is voiced, but voiceless C_1 alone is enough to induce devoicing-based *perception*. The (i) interpretation leads us to expect that **post-voiced** stimuli should pattern with the **voiceless environment** stimuli both in /i/ monitoring and in lexical decision. In contrast, the (ii) interpretation (coupled with the assumption of the prototype effect) leads us to expect that /i/ monitoring and lexical decision results with their **post-voiced stimuli** (e.g., /sekidome/) should be different; they should pattern with the **voiceless environment** stimuli in /i/ monitoring but with the **voiced environment** stimuli in lexical decision, because lexical decision based on stored prototypical realizations should mirror production patterns. In fact, the prediction under the (ii) interpre-

tation was exactly what was observed by Ogasawara & Warner; while /i/ monitoring results with **post-voiced stimuli** exhibited the general pattern of the **devoicing environment** (no difference between **reduced** and **unreduced stimuli**), lexical decision results with them exhibited the general pattern of the **voicing environment** (more difficult /i/ perception against **reduced stimuli** than in **unreduced stimuli**). To the extent that it supports the (ii) interpretation, such a result suggests both the sublexical reality and and the lexical reality of **DB-sensitivity**; it suggests the sublexical reality because a voiceless consonant [C^i] followed by a voiced consonant is suggested to induce the *perception* of /Ci/, although [C^i] in this context is not observed as a realization of /Ci/ in the *productions* of actual words; it also suggests the lexical reality because the prototype effect is nothing but a manifestation of **DB-sensitivity** at the lexical level.

One implication of the prototype effect is that it poses a problem to feed-forward models of the relation between sublexical and lexical processing like Merge (the original version presented by Norris, et al., 2000; or the Merge B version based on Shortlist B, both presented by Norris & McQueen, 2008). Merge distinguishes two kinds of phoneme units: phoneme *perception* units and phoneme *decision* units. The perception units are assumed to mediate the speech signal and lexical access; the speech signal is first processed by the phoneme perception units, whose output is fed as input to the lexicon; the decision units are task-specific, specifically constructed 'on the fly' when coping with phonemic perception tasks, and receive input both from the phoneme perception units and the lexicon; hence lexical effects on presumably sublexical tasks (such as phoneme monitoring) could be accounted for as an effect of the input from the lexicon to the decision units, without assuming a feedback loop from the lexical units to the phonemic perception units.

The problem is that, under this general architecture, lexical access is assumed to be always mediated by phonemic perception. If so, the **reduced/unreduced** distinction should be able to affect lexical access only to the extent that the success of the initial phonemic perception is sensitive to that distinction. However, according to Ogasawara & Warner's (2009) results,

/i/ monitoring performance in the **devoicing environment** is not sensitive to the distinction, where /i/ monitoring presumably taps on phonemic perception. Thus, according to the versions of Merge as presented by Norris et al. (2000) or Norris & McQueen (2008), the ease of lexical access should not differ across the **reduced** and the **unreduced stimuli**, contrary to Ogasawara & Warner's lexical decision results (the prototype effect). Thus the observed prototype effect suggests that a modification to Norris et al.'s (2000) 'Merge A' or Norris & McQueen's 'Merge B' may be required.

The easiest remedy would be to add a 'direct route' from the speech signal to the lexicon; the prototype effect would then be due to the processing through this route. Note that the originally posited route from the speech signal to phonemic perception units should be retained so that the production-perception asymmetry would be explained. If the latter route (i.e., from the speech signal to the phonemic perception units) is also part of an 'indirect' route for lexical access, leading from the speech signal to the lexicon through phonemic perception units, the resulting model would have dual routes for lexical access, as in the Race model (Cutler & Norris, 1979, cited by Norris et al., 2000).[20]

Such a modification is also in line with Ogasawara's (2013) results. She conducted a shadowing experiment with Japanese listeners, in which similar **reduced** and **unreduced stimuli** in the **devoicing** and the **voicing environment** were the targets to be shadowed. (Again, the

[20] Ogasawara & Warner (2009) also discuss the implications of their own results for Merge.

As stated above, Merge distinguishes two kinds of phoneme units: phoneme *perception* units and phoneme *decision* units. In the case of Ogasawara & Warner's (2009) experiments, then, /i/ monitoring is presumably a phonemic task and hence would involve both the perception units and the decision units, but lexical decision is non-phonemic and hence would not involve the decision units. Then what would the decision units look like?

Ordinary listeners are not consciously aware of the **reduced/unreduced** distinction (and that's why, in Ogasawara & Warner's 2009 experiments, for example, stimuli had to be uttered by a phonetically trained speaker with a subsequent instrumental confirmation of devoicing). Thus, in Ogasawara & Warner's /i/ monitoring experiments, listeners were instructed to monitor for /i/, which subsumes both devoiced and non-devoiced [i]. Put in Merge, then, there was only one phoneme *decision* unit for 'i', while the number of the corresponding phoneme *perception* units may be one (/i/) or two (devoiced and non-devoiced one).

However, Ogasawara & Warner (2009) suggest the inclusion of both non-devoiced and devoiced [i] 'as categories in the phoneme decision module' in order to account for the results of phoneme monitoring in non-words; responses to devoiced [i] was facilitated by 'comparison to real words that have material in common with the non-words'. However, listeners were only asked to monitor for /i/. Thus assuming devoiced and non-devoiced categories as *decision* units does not accord with their experimental task (or with what naive listeners could do). Note that 'decision' in the citation must not be a typo; if they meant phoneme perception units, the idea of an effect of the comparisons to real words would imply feedback, betraying the basic philosophy of Merge.

manipulated vowel was /i/.) She observed that (i) the shadowing RT's were slower to **reduced stimuli** than to **unreduced stimuli** in the **voicing environment**, and (ii) the shadowing RT's to **reduced** and **unreduced stimuli** did not differ in the **devoicing environment**. Both observations exhibit the general pattern observed by Ogasawara & Warner (2009). Ogasawara (2013) interprets such results as evidence for direct access from the speech signal to the lexicon (although she does not admit a separate 'indirect' route through sublexical processing.)

Before closing the discussion of Ogasawara & Warner's (2009) results, one note is necessary. Recall that Dupoux et al.'s (2001) and Mazuka et al.'s (2011) claim of the sublexical nature of phonotactic sensitivity could be defended if **DB-sensitivity** is perceptually real; thus the perceptual reality of **DB-sensitivity** would argue against lexicalist models. However, as stated in 3.3.2, Beckman & Shoji's (1984) and Tsuchida's (1994) results do not establish the perceptual reality of **DB-sensitivity** because the observed coarticulation sensitivity could be taken as the recovery of underlying /i/ independently from devoicing (**DI-sensitivity**). The same problem would apply to Ogasawara & Warner's results if they are seen as an attempt to demonstrate the reality of **DB-sensitivity**. The only suggestion of the devoicing-based nature of the coarticulation sensitivity observed by Ogasawara & Warner comes from the comparison between the **devoicing environment** and **voicing environment**; coarticulation sensitivity was observed only in the **devoicing environment**. However, the lack of observed coarticulation sensitivity in the **voicing environment** could be due to the stimuli. Ogasawara and Warner report the difficulty in uttering the reduced stimuli in the **voicing environment** that Ogasawara felt, which is natural for a native speaker of Japanese, in which vowels do not devoice in the **voicing environment**. That difficulty suggests the possibility that the coarticulation cues within the reduced stimuli in the **voicing environment** were less natural or robust compared to the reduced stimuli in the **devoicing environment**. If that possibility is real, then, the observed difference between coarticulation sensitivity in the **devoicing environment** and in the **voicing environment** could be interpreted not only in terms of the devoicing-based nature of the sensitivity but also in

terms of devoicing-independent exploitation of the sufficient amounts of coarticulation cues in the reduced **devoicing environment** stimuli vs. the insufficient amounts of coarticulation cues in the reduced **voicing environment** stimuli. Thus the above interpretations of Ogasawara & Warner's results, according to which **DB-sensitivity** is real both sublexically and lexically, are still tentative; they rely on the assumption that the observed coarticulation sensitivity is devoicing-based, which requires confirmation.

3.4.3 Cutler et al.'s (2009) claim entailing lexicalist reductions

Under the above interpretation, Ogasawara & Warner's (2009) results suggest that devoicing-based /i/ perception is real both sublexically and lexically. However, Cutler et al. (2009) conducted word spotting experiments with Japanese listeners and interpret the results as showing that "during prelexical processing vowels are not automatically restored or inserted" (p. 1701) after (what they assumed to be) a phonotactics-violating consonant. This claim challenges not only the above interpretation of Ogasawara & Warner's results but also Dupoux et al.'s (2001) and Mazuka et al.'s (2011) claim, reviewed above, of the sublexical reality of perceptual epenthesis. Cutler et al. also cite Mash and colleagues' results, reported fully by Kingston et al. (2011), as supporting evidence for their claim.

In this subsection, Cutler et al.'s (2009) results are first reviewed, after which Kingston et al.'s (2011) results are reviewed; it will be seen that neither results entail Cutler et al.'s conclusion.

Cutler et al. (2009)

Cutler et al.'s (2009) experiments were conducted as a follow-up of McQueen et al.'s (2001) experiments, and both Cutler et al.'s and McQueent et al.'s experiments are based on Norris et al.'s (1997) Possible Word Constraint (PWC). The PWC states that those lexical segmentations of the incoming speech signal are disfavored which would produce segmented portions that

are impossible as words; for example, if a stimulus is of the form /$X_1X_2X_3$/, and if /X_1/ is impossible as a word but /X_1X_2/ and /X_3/ are possible as words, listeners prefer to segment it as /X_1X_2–X_3/, rather than /X_1–X_2X_3/. What are impossible as words would differ cross-linguistically, but it is assumed that a consonant alone could not be a word in any language. Thus, for example, spotting *apple* in the stimulus *fapple* should be more difficult than spotting it in the stimulus *vufapple*, according to PWC, because the former would result in a consonant *f* as the segmentation residue, which could not be a word, while the latter would result in *vuf* as the segmentation residue, which could be a word. What if a similar experiment was conducted in Japanese? Consider the word *agura* (to sit cross-legged). Spotting it in *tagura* would result in a consonant residue *t*, which should hence be more difficult than spotting it in *oagura*, which would result in the residue *o*, which is a possible word.[21] This expectation was born out in McQueen et al. (2001) experiments, but McQueen et al.'s real interest was in whether, in addition to the ban on consonants as possible words, morae also function as segmentation units in Japanese; if they do, spotting *uni* (sea urchin), for example, in *gyaNuni* and in *gyaouni* should be easier than spotting it in *gyabuni*, because the stimuli were *gya-N-u-ni*, *gya-o-u-ni*, and *gya-bu-ni*, with the mora boundaries indicated with a hyphen, and only in the former two cases would the required segmentation boundary coincide with a mora boundary. The results supported the assumption that morae also function as segmentation units in Japanese.[22]

However, what is crucial for our present concern is McQueen et al.'s (2001) observation that spotting targets such as *bikini* ('bikini') or *saru* ('monkey') was significantly slower when they were followed by consonants (the consonant context; *bikinip* or *sarup*) than when they were followed by vowels (the vowel contexts; *bikinia* or *sarua*), although no significant differ-

[21] In fact, *o* is not only a possible word but also a real word in Japanese (meaning 'tail'); it could also be interpreted as an honorific prefix, and *oagura* could be interpreted as an honorific form of *agura* (which is probably rather morphologically incorrect in this particular case, for some reason).

[22] If lexical parses should respect mora boundaries, then spotting *agura* is predicted to be more difficult in *ta-gu-ra* than in *o-a-gu-ra*, independently of the assumed 'impossible word' status of a consonant. Thus more elaboration is needed for the precise nature of the PWC than has been provided here. See the General Discussion section of McQueen et al. (2001).

ence with respect to error rates was observed. Note that stimuli such as *bikinip* or *sarup* would constitute a phonotactic violation in Japanese, with the phonotactics-violating consonant /p/; Cutler et al. (2009) argue that, if Japanese listeners automatically repair phonotactic violations by epenthesizing /u/, as Dupoux and colleagues have argued, the *bikinip/sarup* kinds of stimuli should have been perceived as /bikinipu/ or /sarupu/, in which case the consonant context stimuli should perceptually have incurred no PWC violation, and hence Japanese listeners' word spotting performance should not differ across the 'consonant' and the 'vowel' contexts.[23]

Cutler et al. (2009) further extended this reasoning, focusing on the distinction between those contexts in which seeming phonotactic violations could arise from vowel devoicing (**possible-devoicing context**) and those in which they could not (**impossible-devoicing context**). According to their interpretation, McQueen et al.'s (2001) results had already negated automatic phonotactics-driven vowel perception, so they turned to the question of whether devoicing-based vowel perception is real and automatic.

In their Experiment 1, they compared Japanese listeners' word spotting performance (e.g., spotting *asa* 'morning') in the **possible-devoicing context** (e.g., *asaf*), the **impossible-devoicing context** (e.g., *asap*), and the **vowel context** (e.g., *asau*, *asafu* or *asapu*). If devoicing-based perception results in vowel perception, word spotting should be equally easy in the **possible-devoicing environment** as in the **vowel context**, according to their reasoning. However, they observed that word spotting was more difficult both in the **possible-devoicing** and in the **impossible-devoicing context** than in the **vowel context** not only with respect to RT's but also with respect to error rates; this result was interpreted as suggesting that devoicing-based vowel perception is not real or automatic.

In their Experiment 2, spotting performance with similar targets (e.g., *asa*) were compared across slightly different kinds of the **possible-devoicing context** (*asafte*), the **impossible-**

[23] As seen immediately below, Cutler et al. (2009) assume that /p/'s special status in Japanese phonology (i.e., it only rarely occurs in native stratum words) blocks vowel devoicing after /p/. However, they assume that phonotactic repair is not sensitive to such a special status of /p/.

devoicing context (*asapdo*), and the **CV context** (*asazu*), and they observed equally slower RT's in the former two contexts than the latter. Cutler et al. (2009) also interpret this result as suggesting that devoicing-based vowel perception is not real or automatic.

While the contexts following the targets were compared in their Experiment 1 and Experiment 2, the contexts preceding the targets were compared in their Experiment 3, in which two kinds of targets were used: those which begin with a vowel (/VCV/ targets; e.g., *asa*) and those which begin with a consonant (/CVCV/ targets; e.g., *sake* 'sake'). The target factor (/VCV/ vs. /CVCV/) was crossed with the preceding context factor (/CCVCC–/ such as *myoji*– or *nyagu*– vs. /CCVC–/ such as *myoch*– or *nyak*–). Spotting *asa* or *sake* in *myojiasa* or *myojisake* should be easy, while spotting *asa* in *myochasa* should be difficult (with the mora boundary not coinciding with the segmentation boundary).[24] The real question is whether spotting /CVCV/ targets (e.g., *sake*) in the /CCVC–/ context (e.g., *nyak–sake*) would be easy or hard, because the preceding context portions could be perceived as /CCVCV–/ (e.g., *nyaku*–) through devoicing-based vowel perception. They observed that (i) /CVCV/ targets (e.g., *sake*) were significantly missed more often in the /CCVC–/ context (*nyak–sake*) than in the /CCVCV–/ context (*nyagu–sake*), (ii) but was not as often as /VCV/ targets (e.g., *asa*) were missed in the /CCVC–/ context (*myoch–asa*), (iii) RT's with /CVCV/ targets (e.g., *sake*) did not significantly differ across the /CCVC–/ context (*nyak–sake*) and the /CCVCV–/ context (*nyagu–sake*). Indeed, the (ii)–(iii) results accord with the idea that vowels *are* perceptually restored or inserted after the final /C/ in the /CCVC–/ context, but the (i) result is expected rather from the idea that vowels are *not* perceptually restored or inserted between *nyak* and *sake*. Cutler et al. (2009) interpret such results as supporting the idea that devoicing-based vowel perception is not real

[24] Speaking more precisely, Cutler et al. (2009) attribute the expected difficulty with spotting *asa* in *mochasa* to the observation that vowels do not devoice before another vowel and hence *mochasa* would not count as a vowel-devoiced realization of *mochVasa* (where 'V' is either /i/ or /u/). However, even if a voiceless consonant alone tends to give rise to the perception of a high vowel, as a result of devoicing-based perception, the cued vowel should be either /i/ or /u/, and that should conflict with the immediately following vocalic portion, with a clear cue for /a/. The perception of /i/ or /u/ should be based on coarticulation cues within the consonant, which is presumably weaker than the cue provided by the vocalic portion, in which case coarticulation-based /i, u/ perception should lose the vocalic-portion-based /a/ perception. Thus devoicing-based perception does not predict that *mochasa* should be perceived as *mochVasa*.

or automatic

Cutler et al. (2009) further classified the **possible-devoicing context** stimuli into the following two sets:[25]

> **Set 1:** Those for which relatively many words (≥ 10) matched a sequence linking the /CCVC-/ context and the /CVCV/ target, if /u/ is perceptually inserted immediately after the context (e.g., *kusa* 'grass' in *nyak(u)-sake*, from which listeners were expected to spot *sake*).
>
> **Set 2:** Those for which relatively few words (≤ 5) matched such a sequence (e.g. *shuha* in *gashuharu*, from which listeners were expected to spot *haru* 'spring').

They observed that (i) words were harder to spot in the /CCVC–/ than in the /CCVCV–/ contexts both in Set 1 and in Set 2, but (ii) the relative difficulty with the /CCVC–/ context was more severe in Set 2 than in Set 1.[26] They interpret those observations as suggesting that, when spotting *sake* in *nyaksake*, for example, **(a)** non-target words such as *kusa* are lexically activated, which lexically helps /ku/ perception; **(b)** with /ku/ perceived, segmenting *nyak(u)sake* into *nyak(u)* and *sake* would not be penalized by the PWC, **(c)** in which case the activation of *kusa* (in the lexical parse *nya-k(u)sa-ke*) and the activation of *sake* (in the lexical parse *nyak(u)-sake*) compete, but because *kusa* is not fully supported by the bottom-up cues, with /u/ being absent in the signal, *sake* wins. According to this story, **(d)** devoiced vowels are restored only lexically, and **(e)** how successful such a restoration will depend on how many words would lexically help the restoration (in the (a) step). Thus they conclude that devoiced vowels are not restored pre-lexically; they are restored only lexically.

This story is conceptually interesting in the sense that the recognition of winning words in lexical competition (e.g., *sake*) is partially due to the activation of losing words in lexical

[25] *Shuha*, an example of matching sequences for Set 2, was presented by Cutler et al. (2009:1701) as a real word in Japanese, but they do not indicate what that word means; I have to confess I do not recognize it as a real word in Japanese, although I am a native speaker.

[26] The (i) observation comes from the main effect of 'context' and the (ii) observation from the interaction between 'context' and 'Set'; both were significant by participants, but not by item.

competition (e.g., *kusa*). Note that Cutler et al.'s (2009) conclusion directly conflicts with the idea of sublexical devoicing-based vowel perception, as well as the sublexical reality of phonotactic repair. They claim that vowels (irrespective of whether they are /u/ or /i/ or something else) are not perceptually inserted in the **possible-devoicing context** without lexical supports, which amounts to denying the sublexical reality of devoicing-based vowel perception from a consonant; given that **DB-sensitivity** presupposes devoicing-based vowel perception from a consonant, then, it seems to imply the denial of the sublexical reality of **DB-sensitivity**. Furthermore, it also seems to imply the denial of the sublexical reality of phonotactic sensitivity. As far as Cutler el al.'s experimental task is concerned, devoicing-based vowel perception and phonotactics-based vowel epenthesis are indistinguishable, so if their results suggest that vowels were not sublexically restored or inserted, that would deny not only the sublexical reality of devoicing-based vowel perception (a presupposition for **DB-sensitivity**) but also the sublexical reality of phonotactics-based vowel epenthesis (phonotactic sensitivity).

However, Cutler et al.'s (2009) results do not entail their conclusion, as will be seen below.

In fact, Cutler et al.'s (2009) experiments have several problems. For one thing, they assume that /p/ does not tolerate the devoicing of the immediately following vowel, but the only reason for this assumption is that /p/ does not occur in the native stratum.[27] However, it is not clear at all why non-occurrence in the native stratum implies the impossibility of the devoicing of the immediately following vowel; /p/ does comfortably occur in the loanword- as well as onomatopoeic-stratum, and unless the possibility of devoicing is shown to depend on the stratum, their assumption concerning /p/ is not justified.[28] A yet more puzzling is their assumption that /f/ allows the devoicing of the immediately following vowel; if they mean a

[27] Although they do not state it, things are a bit more complicated. First, non-geminate /p/ is disfavored not only in the native stratum but also in the Sino stratum. For another, it is only disfavored; although rare, non-geminate /p/ does occur in the native as well as the Sino stratum (according to Labrune, 2012).

[28] Note the same authors' remark in McQueen et al. (2001:108) that loanwords 'are marked in the orthography (through the use of *katakana* script), they are phonologically fully incorporated into" Japanese; furthermore, the results observed in their Experiment 1, in which many of the word spotting targets were loanwords, were replicated in their Experiment 3, in which loanword targets were avoided.

voiceless labio-dental fricative by /f/, it does not occur in Japanese in the first place, irrespective of the stratum; if non-occurrence of /p/ in the native stratum implies the impossibility of devoicing, the same should apply to /f/. Thus their distinction between **possible-devoicing** and **impossible-devoicing context** does not make a good sense with respect to /p/ and /f/, although their classification of consonants other than /p/ and /f/ into those two contexts seem to be reasonable.

Furthermore, their assumption concerning lexical competition (**(c)** above) is inconsistent with Ogasawara & Warner's (2009) lexical decision results (the prototype effect). That is, in the case of stimuli such as *nyaksake*, according to Cutler et al. (2009), the activation of *kusa* (in the lexical parse *nya-k(u)sa-ke*) and the activation of *sake* (in the lexical parse *nyak(u)-sake*) compete, but because *kusa* is not fully supported by the bottom-up cues, with /u/ being absent in the signal, *sake* wins; this lexical competition story is simply imagined, with no direct experimental support. However, Ogasawara & Warner's (2009) lexical decision results rather suggest that vowel-devoiced stimuli are more easily recognized as words than vowel-non-devoiced stimuli, which suggests that *ksa* should be more easily recognized as /kusa/ than *kusa*. If so, both /kusa/ and /sake/ should be equally activated from *nyaksake* stimuli, and hence there should be no reason for the activation of /kusa/ to lose the activation of /sake/.[29]

More crucially, if the results of Cutler et al.'s (2009) experiments are accepted, they do not necessarily imply Cutler et al.'s conclusion. Their observations were simply that, interpreted in terms of the PWC, '…C_1-C_2…' segmentations were not as easy as '…C_1V-C_2…' segmentations; under the PWC, which dictates that the ease of segmentation should depend on the perception of a vowel between 'C_1' and 'C_2', such observations would simply mean that vowel perception with '…C_1C_2…' is not as easy as with '…C_1VC_2…', not necessarily the denial of vowel perception with '…C_1C_2…' altogether; their results could be interpreted by assuming

[29] We could also note that the 'Set 1 vs. Set 2' division is solely based on type frequencies of competing words, with no consideration of their token frequencies or Japanese listeners' familiarities with them.

that vowels are indeed perceived, but with some difficulty.

In fact, there are several empirical reasons that such an interpretation should not be dismissed. First consider presumably devoicing-based vowel perception. As noted above, Ogasawara & Warner's (2009) /i/ monitoring results suggest that /Ci/ perception from [C^{i}] is equally easy as /Ci/ perception from [Ci], while their lexical decision results suggest that /Ci/ perception from [C^{i}] is easier than /Ci/ perception from [Ci]. However, the relative ease of such presumably devoicing-based vowel perception with a phonemic task could vary. For example, as noted above, Beckman & Shoji's (1984) and Tsuchida's (1994) results could be, strictly speaking, interpreted as results of **DI-sensitivity**, but they could well be due to **DB-sensitivity**. The experimental task in their experiments was vowel identification, which is a phonemic task, but Beckman & Shoji report more difficulty of vowel identification from vowel-devoiced stimuli than from non-devoiced stimuli, while Tsuchida reports the opposite; identification was significantly better when the vowels were devoiced than when they were not. While a comparison between Ogasawara & Warner's /i/ monitoring and lexical decision results suggests that the relative ease of such devoicing-based vowel perception depends on the experimental task, a comparison among Ogasawara & Warner's /i/ monitiring results, Beckman & Shoji's (1984) and Tsuchida's (1994) vowel identification results suggests that the relative ease cannot be completely predicted from whether the task is sublexical or lexical. One obvious candidate reason for the difference among Ogasarawa & Warner's, Beckman & Shoji's and Tsuchida's phonemic experiment results is that the amount or quality of coarticulation cues within the stimuli differed; **DB-sensitivity** depends on the amount or quality of coarticulation cues within the consonantal stimuli by definition. Thus Cutler et al.'s results could possibly be attributed to weak coarticulation cues in the [k] portion of the *nyak–* context stimuli.

Next consider phonotactic sensitivity. By definition, vowel perception from a consonant caused by phonotactic sensitivity involves some sort of epenthesis operation, whatever the operation is exactly like. Thus /CV/ perception from [C] due to phonotactic sensitivity should be

more costly than /CV/ perception from [CV], because the former should involve an additional perceptual process of epenthesis. Thus more difficult vowel perception with "…C_1C_2…' than with '…C_1VC_2…' is just as expected.

Furthermore, the lexical effects claimed by Cutler et al. (2009) could be interpreted as additonal facilitation provided by the lexical level, independently of the functioning of the phoneme perception units, if the general architecture of models such as Merge are accepted. (Recall that the existence of various lexical effects reported in the literature was not conceived of as evidence that phoneme perception units should not be assumed in Merge.)

In short, more difficulty in vowel perception within 'C_1C_2' than within 'C_1VC_2' or some lexical effect helping vowel perception does not imply that vowel is not perceived at all at the sublexical level. Although the precise nature of the lexical effect claimed by Cutler et al. (2009) needs a more careful investigation,[30] their experimental results do not argue against the sublexical reality of **DB-sensitivity** or of phonotactic repair.

Kingston et al. (2011)

Cutler et al. (2009) also cite the experimental results by Mash, Kawahara, Kingston, Brenner-Alsp, and Chambless, presented at a conference, as supporting their conclusion that devoiced vowels are not perceived sublexically. Next let us consider what Mash and colleagues' results do and do not imply concerning whether devoiced vowels are sublexically perceived, based on the journal paper version, Kingston et al. (2011).

Mann (1980) observed that English listeners' identification of stops intermediate between /d/ and /g/ is affected by the preceding liquid; they tend to perceive /g/ more often when the stop is preceded by /l/ than when it is preceded by /r/. Mann and Repp (1981a; 1981b) reported a similar effect of fricatives on the identification of stops intermediate between /t/ and /k/; English listeners tend to perceive /k/ more often when the stop is preceded by /s/ than when it

[30] As noted above, Cutler et al.'s interpretation conflicts with Ogasawara & Warner's (2009) lexical decision results (the prototype effect).

is preceded by /ʃ/. A simple observational generalization is that a preceding front consonant C_1 (the context phoneme) encourages the perception of a following back consonant C_2 (the target phoneme). Let us simply refer to those observations 'compensation effects' (given that those observations have come to be known under the name 'compensation for coarticulation').

Three possible accounts could be imagined for such compensation effects:

The categorization account: The perceived front phoneme categories of C_1 somehow encourage the categorization of C_2 as a back phoneme.

The articulatory account: The articulatory gestures are recovered from the speech signal; front C_1 results in more front articulation of C_2, and listeners tend to attribute the acoustic properties of C_2 signaling front articulation to such compensation effects of C_1, rather than to the inherent articulation of C_2 (Mann, 1980; Mann & Repp, 1981a; 1981b; Fowler, 2006).

The auditory account: The acoustic properties of the speech signal gives rise to speech perception without the mediation of articulation recovery; the acoustic properties of front C_1 auditorily induce more backward perception of C_2 (Lotto & Holt, 2006; Lotto & Kluender, 1998).

Note that both **the articulatory** and **the auditory account** share the view that it is the physical acoustic properties of C_1, rather than its perceived phonemic category, that is (observationally) responsible for the effect; thus they contrast with **the categorization account**.

According to Kingston et al. (2011), their experiments were conducted in order to tease apart the predictions of **the articulatory** and **the auditory account**; **the categorical account** was not considered as a candidate in designing their experiments. They focused on the observation that /si/ and /su/ in Japanese are usually realized as [ɕ$^{\text{i}}$] and [s$^{\text{ɯ}}$] respectively.[31] Crucially,

[31] In Kingston et al.'s notation, these realizations are [ʃi̯] and [su̯] respectively. However, their [ʃ] should read [ɕ], and given that the speakers were from Tokyo, their [su] should read [sɯ]. Furthermore, it is generally agreed that in /CV/, where /C/ is a fricative, the devoicing of /V/ results in a prolonged fricative noise, with no 'whispered vowel'; thus they should more properly be transcribed as [ɕ$^{\text{i}}$] (or [ɕ˕]) and [s$^{\text{ɯ}}$] (or [s̟]) respectively.

acoustic vestige of the devoiced vowels remains within the fricative; thus, in the case of /si/, [ɕ] is relatively back but the /i/ vestige is relatively front, while in the case of /su/, [s] is relatively front but the /u/ vestige is relatively back. Assuming that the front/back distinction of the fricatives is acoustically larger than the front/back distinction of the vestige, they assume that **the auditory contrast** account predicts that the /si, su/ should only exhibit the compensation effect due to the fricative, on the perception of the immediately following /t/–/k/ continuum; more /k/ perception after [ɕ^{i}] should be observed. In contrast, given that articulatory gestures of both the fricatives and the devoiced vowels are present, they assume that **the articulatory account** leads us to expect that both the fricative and the devoiced vowel should exert a compensation effect on the perception of the immediately following /t/–/k/ continuum, and the effects of the fricative and of the devoiced vowel should cancel each other out; the net effect should be no compensation effect.

Kingston et al.'s (2011) experiments were conducted with Japanese and English listeners, and the results turned out to bear our the former expectation; both Japanese and English listeners exhibited more tendency for /k/ perception after [ɕ^{i}] than after [s^{ɯ}]. It is this result that Cutler et al. (2009) saw as evidence that Japanese listeners do *not* sublexically perceive devoiced vowels; if they did, they should have exhibited a compensation effect due to the devoiced vowels (in contrast to English listeners).

However, note that Cutler et al.'s (2009) interpretation of this result implicitly assumes **the categorical account**, which states that the compensation effect reflects the perceived phonemic identity of C_1 (the context phoneme), not the acoustic properties. In fact, this is rather against Kingston et al.'s (2011) position; in their conception, **the articulatory account** predicts that both the fricative and the devoiced vowel should exert a compensation effect on the perception of the immediately following stop continuum, and the above result betrays this prediction; **the articulatory account** could be defended if Japanese listeners are assumed to ignore the articulations of 'devoiced' vowels, but Kingston et al. call that assumption *ad hoc* (p. 521;

italicization by Kingston et al.); for Kingston et al., Beckman & Shoji's (1984) and Tsuchida's (1994) results had already shown that Japanese listeners do perceive devoiced vowels.

Indeed, Kingston et al.'s (2011) position could well be wrong. Past research suggests that it is not a simple matter to decide whether the perceived phonemic category or the spectral properties of C_1 is responsible for compensation effects.

For example, Mann and Repp (1981a; Experiment 5) varied spectral properties of C_1 and examined (i) the perceived category of C_1, and (ii) the perceived category of C_2. They observed that the perceived category of C_2 (the compensation effect) is affected both by within-category spectral differences of C_1 as well as the perceived category of C_1. However, Mann and Repp (1981b; Experiment 2) failed to observe the effect of the perceived category of C_1 independent from within-category spectral differences. Furthermore, Mann (1986) examined the compensation effect of liquid C_1, which had been observed with English listeners by Mann (1980), on the identification of the immediately following /d/–/g/ continuum. In Mann's (1986) experiments, this effect was examined both with English and with Japanese listeners. The context liquids were English /l/ and /r/, which do not exist as separate phonemes in Japanese and are, as is well known, very hard to distinguish for Japanese listeners. However, the same kind of compensation effect was observed with Japanese listeners that was observed with English listeners.[32] Mann (1986) interprets this result as suggesting the existence of a pre-phonemic perceptual stage, which is not affected by the native language. In other words, according to Mann's view, the compensation effect is a reflection of the assumed universal pre-phonemic stage, not of phonemic processing. If so, Kingston et al.'s (2011) result in question would be simply irrelevant to the issue of whether Japanese listeners phonemically perceive devoiced vowels, contrary to Cutler et al.'s (2009) assumption.

However, the idea of the universal pre-phonemic stage is probably too strong. In fact,

[32]Speaking more precisely, this result was obtained with two groups of Japanese listeners: those who were advanced L2 learners of English, who exhibited good /l/–/r/ discriminations, and those who were not, who exhibited only chance-level discriminations. However, when it comes to the compensation effect in question, both of those two groups exhibited the compensation effect.

Kingston et al. (2011; Experiment 2) report that Japanese listeners failed to exhibit a compensation effect to voiced vowels in the first place, in contrast to English listeners, who did exhibit such an effect. This observation has two implications. First, the lack of a compensation effect of 'devoiced' vowels in [ɕ^{i}] and [s^{ɯ}] with Japanese listeners could be attributed to the lack of a compensation effect of voiced vowels; [ɯ] and [i] are simply not the right kinds of vowels to induce a compensation effect with Japanese listeners (whether voiced or 'devoiced'). Second, the contrast between Japanese and English listeners with respect to the presence vs. the absence of the compensation effect with the same voiced vowels suggests that the processing stage responsible for compensation effects is not completely universal. The first implication casts a further doubt on Cutler et al.'s (2009) interpretation; if the lack of the compensation effect due to 'devoiced' vowels should imply that Japanese listeners do not perceive devoiced vowels, the lack of the compensation effect with voiced vowels should imply that they do not perceive voiced vowels either, a conclusion that nobody would accept. On the other hand, the second implication (i.e., the non-universality of the processing stage responsible for compensation effects) would best be illustrated with Fowler's (2006:168) position,[33] according to which listeners parse the speech signal into articulatory gestures, but it

> is well established, however, that listeners are not ideal parsers ... Sometimes, they pull out too much; sometimes, they pull out too little ... The conditions under which each outcome is observed have not been determined.

Indeed, Kingston et al.'s (2011; Experiment 3) results with respect to voiced vowels' compensation effects on the immediately following stop continuum are quite complex; both English and Japanese listeners exhibited less /t/ identifications after /u/ than after /o/ (while Japanese listeners did not exhibit such an effect when /u/ and /i/ are compared); long vs. short vowels resulted in no compensation effect with Japanese listeners, but in rather idiosyncratic results

[33]Here, **the articulatory account** is adopted simply for an expository purpose.

with English listeners (more /t/ responses after short /u/ than long /u/, but more /t/ responses after long /e/ than short /e/). The results with the long vs. short vowels suggests that what Fowler calls 'the conditions under which each outcome is observed' are not a simple function of phonological learning; long vs. short vowels are phonemically contrastive in Japanese but did not affect Japanese listeners' performance, whereas English listeners exhibited sensitivity to a distinction which does not seem to be phonemically contrastive.[34]

To summarize the crucial points of the above discussion, Mann's (1986) results with English /l, r/ with Japanese listeners strongly suggest that compensation effects are not a simple function of the perceived phonemic identities of the target phonemes, which is also suggested by Kingston et al.'s (2011) failure to observe a compensation effect with voiced /i, u/ with Japanese listeners; thus Kingston et al.'s failure to observed the compensation effect with devoiced vowels inducing does *not* support or refute sublexical perception of devoiced vowels, contrary to Cutler et al.'s (2009) interpretation.

3.4.4 Section Conclusion

Ogasawara & Warner's (2009) observations of the contrast between the **devoicing environment** vs. the **voicing environment**, between the **pre-voiced** and the **post-voiced stimuli**, and between /i/ monitoring vs. lexical decision, point toward both the sublexical and the lexical reality of **DB-sensitivity**. To the extent that the sublexical reality is suggested, they would rather argue against lexicalist models, which claim that coarticulation sensitivity *reduces* to lexical sensitivity. On the other hand, neither Cutler et al.'s (2009) or Kingston et al.'s (2011) results necessarily imply the denial of the sublexical reality of **DB-sensitivity** (or of sublexical reality of phonotactic sensitivity). Thus the above discussion seems to favor the sublexical reality of **DB-sensitivity**.

[34]This consideration argues against Kingston et al.'s (2011) idea, which they themselves call speculative, that English listeners exhibited sensitivity to voiced /i/ vs. /u/, produced by a Japanese speaker, because they assimilated the /u/ stimuli to English /u/ (a kind of **the categorical account**). If assimilation to native phonemic categories is at work, why the (idiosyncratic) length effects with English listeners (but not with Japanese listeners)?

However, the difficulty with interpreting Beckman & Shoji's (1984) and Tsuchida's (1994) results as evidence for **DB-sensitivity** partially applies to Ogasawara & Warner's (2009) results too. Indeed, Ogasawara & Warner's observation of the contrast between the **devoicing environment** vs. the **voicing environment** seems to suggest that the coarticulation sensitivity they observed was **DB-sensitivity**, rather than **DI-sensitivity**, but there remains the room for an interpretation in terms of **DI-sensitivity** because the /i/ perception could be seen as a recovery of the devoiced /i/, rather than a phonemic categorization (interpretation) of front coarticulation within a consonant on the one hand, and the contrast between the effects of reduction/devoicing with **devoicing** and **voiced environment** could be interpreted in terms insufficient coarticulation in the **voicing environment reduced stimuli**, rather than in terms of the devoicing-based nature of the sensitivity, on the other. Thus the reality of **DB-sensitivity** still needs a further confirmation.

Another problem that remains is what determines the sublexical devoicing-based ease of /CV/ perception from [C]. Beckman & Shoji's (1984) vowel identification results and Cutler et al.'s (2009) word spotting experiments suggest that /CV/ perception from [C] is more difficult than /CV/ perception from [CV], Tsuchida's (1994) vowel identification results suggest the opposite, and Ogasawara & Warner's (2009) /i/ monitoring results suggest the lack of difference. The amount or quality of the coarticulation traces within the consonantal stimuli was suggested above as a possible cause for such conflicting results, but it is only a conjecture at this point; it needs to be examined empirically.

3.5 One-step Models vs. Two-step Models

Thus far we have asked whether phonotactic sensitivity and/or coarticulation sensitivity could be reduced to lexical sensitivity. The answer was rather negative, but the answer depends on the assumed reality of **DB-sensitivity** (as opposed to **DI-sensitivity**), which has to be confirmed

yet. If it is confirmed to be real, then, lexicalist models should be denied. That would leave two-step models and one-step models as candidates. So we turn now to evidence found in the previous literature relevant to the choice between two-step models and one-step models.

Subsection 3.5.1 conceptually examines one-step models and points out that they could be implemented in two different versions (**suprasegmental matching** and **slot filling**). Subsection 3.5.2 examines what Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) and Dupoux et al. (2011) saw as potential pieces of evidence against one-step models; it will be pointed out that they are in fact compatible with (both versions of) one-step models. It will also be pointed out that Matthews & Brown's (2004) results (not cited by Dupoux et al., 2011) constitute evidence against the **suprasegmental matching** version. However, they are fully compatible not only with two-step models but also with the **slot filling** version of one-step models. Thus none of the empirical results reviewed in this subsection constitute evidence against one-step models as a whole. Subsection 3.5.3 examines Dehaene-Lambertz et al.'s (2000) and Dupoux et al.'s (2011) argument for one-step models; it will be seen that they are inconclusive. Section 3.5 as a whole thus suggests that the choice between one-step models and two-step models is still open.

3.5.1 Two Implementations of One-step Models

As noted above, Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) confirmed the perceptual nature of phonotactic epenthesis. At the end of their discussion Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) raised the question of how such epenthesis perceptually arises.

They pointed out two possibilities. One is that incoming speech is first analyzed as individual phonemes, and the resulting phoneme string then gets parsed into such units as syllables (Church, 1987). This is the view of what we, following Dupoux et al. (2011), call two-step models, according to which the speech signal is first perceived as (a string of) phonemes in a phonotactics-independent way, and phonotactic effects arise as a result of subsequent phono-

tactic repairs of the string.

What Dupoux and colleagues (Dehaene-Lambertz et al., 2000; Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Dupoux et al., 2011) saw as an alternative is that phonotactic repair is, say, pre-compliled; suprasegmental units such as syllables, rather than phonemes, are acquired, and the initial perceptual stage consists of a matching of the incoming speech signal against the repertoire of the acquired suprasegmental units (such as syllables) rather than individual phonemes; because only phonotactically admissible suprasegmental units are within the listeners' acquired repertoire, the incoming speech signal is assimilated to a phonotactically admissible pattern.[35] In other words, according to such a **suprasegmental matching** view, phonotactic effects are, somewhat metaphorically speaking, 'pre-complied' in the reportoire of suprasegmental units acquired in infancy. This view claims that phonemic identification (as observed in an identification experiment) is in fact a result of the listeners' post-hoc analysis of the perceived suprasegmental unit. This view denies the existence of two separate and sequential perceptual stages, the first for phonemic categorization and the second for phonotactic repair, as assumed by two-step models; rather, perception is claimed to consist of a single stage of the matching between the speech signal and the suprasegmental unit repertoire. Thus such a view implements one-step models.

However, although not proposed in the previous literature, another implementation of one-step models is possible, which we could call **slot filling**.[36] First note that two-step models

[35]Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) mention Mehler et al.'s (1990) "template matching of incoming speech signals with syllables" as an instance of such a thesis. Speaking more precisely, Mehler et al.'s claim is multi-faceted; they claim (i) some discrete sublexical perceptual units should be posited, to which the incoming speech signal is mapped, (ii) the units in question are larger than phonemes, (iii) phonemes are recognized through a post-perceptual analysis of the suprasegmental units. They carefully remain neutral with respect to what the units exactly are; they may or may not correspond to syllables posited in metrical phonology, or may differ cross-linguistically. However, the suprasegmental units in question are assumed to be the domain of phonotactic regularities. In fact, whether the Church-style parsing perspective or such a suprasegmental matching perspective should be employed, what suprasegmental units capture perceptually relevant phonotactic constraints is a separate issue to be investigated (Berent et al., 2007; Kabak, 2003; Kabak and Idsardi, 2003, 2007; Moreton, 2002).

[36]In fact, my idea of **slot filling** has its origin in my misunderstanding of Kakehi et al. (1996); somehow I misunderstood their experimental findings and wrongly thought that Kakehi et al. claimed **slot filling**. If my misunderstanding of their experimental results is partially due to the way the results were presented in Kakehi's invited lecture at a conference (the date or details of which I can not recall), possibly any advantage of the idea of **slot filling** should partially be attributed to Kakehi.

On the other hand, Alice Turk called my attention to Shattuck-Hufnagel (1979). In Shattuck-Hufnagel's (1979)

embodies three claims:

- Phonemic segments are perceptually real. (Perceptual reality of phonemic segments)
- Suprasegmental units (which function as domains for relevant phonotactic constraints) are also perceptually real. (Perceptual reality of suprasegmental units)
- Suprasegmental units are constructed based upon already perceived phonemic segments. (The dependency of suprasegmental unit perception on phonemic segment perception)

The **suprasegmental matching** version of one-step models denies the perceptual reality of phonemic segments (and hence the dependency of suprasegmental unit perception on phonemic segment perception); phonemes are assumed to be recognized only as a result of post-perceptual meta-analysis of suprasegmental units. However, the **slot filling** version admits the perceptual reality of phonemic segments on the one hand, and of suprasegmental units on the other, but denies the dependency of suprasegmental unit perception on phonemic segment perception; rather, according to the **slot filling** version, the perception of phonemic segments depends on the construction of suprasegmental units.

More specifically, according to the **slot filling** version, listeners are equipped with phonotactics-respecting *structured frames* of suprasegmental units, with empty slots to be filled in by phoneme-sized segments;[37] if the suprasegmental unit in question is the syllable in the conventional phonological sense, it should have slots named, say, 'two consonant positions in the onset'. Speech perception then consists of filling those slots. According to such a view, phonotactic effects should arise because of the constitution of suprasegmental frames, which are abstract

frame-and-filler model of speech production (not perception), phonemic segments are production units but only as fillers for phonotactics-respecting syllabic frames. Thus the **slot filling** version of one-step models could be conceived of as a perception version of Shattuck-Hufnagel's frame-and-slot model of speech production.

[37] Alternatively, the structured frames could be conceived of as structural templates whose phonemic constituent portions are underspecified (rather than being empty slots), in which case phonemic perception is a process of filling in the pieces of information underspecified in the templates, in the manner of the so-called unification-based grammar formalism employed for syntax and semantics (Pollard & Sag, 1987; 1994; Shieber, 1986). The 'employ slot' implementation and the 'underspecification' implementation do not differ with respect to their empirical predictions, as far as this thesis is concerned. In the remainder of the thesis, the 'empty slot' implementation is employed.

structured entities, which, in the case of languages such as modern Japanese, would look like

$$/C(j)V\left\{\begin{array}{c} Q \\ \\ N \end{array}\right\}/$$

or, in the case of languages such as ancient Japanese, which prohibits onset clusters and codas,

/CV/

Thus the success of phonemic perception should depend on whether it succeeds in filling an appropriate slot. However, in order to fill a slot, a slot has to be provided by a suprasegmental structural frame. Thus phonemic perception depends on the perceptual construction of suprasegmental units.

For example, suppose [keb] is heard by a listener of, say, ancient Japanese, which bans codas altogether. Also suppose that the relevant phonotactic domains are syllables. According to two-step models, the perception of /k/, /e/ and /b/ is independent of their syllabic positions; /keb/ is perceived in the first step as an unstructured phoneme string, and is structurally parsed in the second-step; the second-step parsing of /keb/ into syllables either deletes /b/ (an illicit coda) or epenthesizes a vowel after /b/ (so that it would constitute a licit onset of the second syllable). In contrast, according to the **slot filling** version of one-step models, listeners attempt to build up syllables from the outset. A speech signal consists of at least one syllable, so one structural syllabic frame will be posited as soon as the onset of the speech signal arrives. Listeners' next task is to fill in the onset and nucleus slots. The phonemic cues for /k/ and /e/ are successfully exploited to fill in the onset and the nucleus slot respectively. However, as far as only one structural syllabic frame is assumed, there will be no way to exploit the perceptual cues for /b/, because the unit does not have a coda slot. Thus, those cues could be ignored (perceptual deletion). Alternatively, if those cues are too salient to be ignored, the only way to accommodate /b/ would be to construct an additional, second structural syllabic frame, in which case the perceptual cues for /b/ would be exploited to fill in the onset slot of this second

syllabic frame. However, the nucleus slot of this second syllabic frame has to be filled, in which case some vowel (probably /u/) is perceptually inserted (perceptual epenthesis).

In the sense that suprasegmental units (such as syllables) are claimed to be structured entities with phonemic constituents, the **slot filling** version of one-step models on the one hand, and two-step models on the other, share the view that a speech signal is suprasegmentally parsed. However, the assumed nature of the parsing operations differs. According to two-step models, phonemic parsing (parsing of the speech signal into phonemes, i.e., phonemic categorization) precedes suprasegmental parsing; thus the speech parser is assumed to be strictly bottom-up and serial. In contrast, according to the **slot filling** version of one-step models, phonemic parsing and suprasegmental parsing are parallel and interactive; for example, the first syllabic frame will presumably be assumed without the perception of a specific phoneme (top-down information flow), and, in the above perceptual epenthesis scenario, the perceptual cues for /b/ function as cues for the existence of /b/ and of the existence of the second syllabic frame (with an onset slot to accommodate it) at the same time (bottom-up information flow), but, as soon as the second syllabic frame is assumed, that would lead listeners to fill in the nucleus slot, this time by epenthesis (top-down information flow).

Not only the **suprasegmental matching** version but also the **slot filling** version does count as an implementation of one-step models to the extent that the crucial or defining characteristic of one-step models is that they deny the idea that phonotactic repair is a post-phonemic operation.

The **slot filling** version of one-step models differs from the **suprasegmental match** version with respect to whether phonemes are recognized as legitimate perceptual entities. In other words, they differ with respect to how phonemes are assumed to be recognized by the listener. According to the **suprasegmental matching** version, phonemic perception is in fact derived from the perception of suprasegmental units as a result of a post-perceptual meta-analysis process (Mehler et al., 1990). In contrast, according to the **slot filling** version, phonemes are

perceived as phonemes at the initial stage of perception; it is just that, to use a metaphor based on the university enrollment system in the U. K., phonemic perception is just like 'conditional offer of acceptance'; phonemes (applicants) are perceived (accepted) only if they succeed in filling in some slots (only if they attain the required IELTS or TOEFL scores).[38]

Another difference between the **suprasegmental matching** version and the **slot filling** version concerns what is expected to happen with unattested clusters. For example, suppose a language in which a coda is generally allowed but somehow a /b/ coda is unattested. In the **suprasegmental matching** version, suprasegmental units are assumed to play the role conventionally assumed to be played by phonemes in phonemic categorization. Thus, just as unattested phones are expected to be assimilated to attested phonemic categories, unattested suprasegmental units are assumed to be assimilated to attested suprasegmental units. Because the unattested /keb/ should not be within the acquired repertoire of suprasegmental units, a [keb] stimulus should be assimilated to /ke/ (perception deletion) or /kebV/ (perceptual epenthesis) or /keC/, where C is not /b/ (perceptual conversion). In contrast, according to the **slot filling** version, what is acquired is not a repertoire of specific suprasegmental units but rather structural frames, such as /CVC/ in this particular hypothetical language. With a coda structurally permitted, then, a [keb] stimulus should not incur perceptual deletion of /b/ or perceptual epenthesis after /b/ or perceptual conversion of /b/ to another consonant (in the absence of a further constraint on the coda position within the structural frame).

[38] A comparison with syntactic notions might further help the reader's understanding of the **slot filling** version.

In syntax, two broad classes of theories exist: those based on 'categories' and those based on 'grammatical functions/relations'. Such notions as 'Noun' or 'Noun Phrase' are examples of 'categories', while such notions as 'subject' and 'object' are examples of 'grammatical functions/relations'. Consider the following sentences:

(i) The boy likes the girl.

(ii) The girl likes the boy.

The boy is a Noun Phrase irrespective of in what sentence it appears, but it is the subject in (i) but not in (ii); whether something is a subject or not is determined not by its inherent properties but rather with respect to what function it serves in the sentential context (or its relation to the other elements in the sentence).

For two-step models, phonemic perception resembles syntactic categories; it does not depend on their structural relations with other phonemes; for the **slot filling** version of one-step models, phonemic perception resembles grammatical functions/relations more than syntactic categories; it does depend on its structural relations with other phonemes.

Thus, while the **suprasegmental matching** version denies perceptual access to sub-syllabic elements (assuming that the relevant units are syllables, rather than bi-phones or morae, etc.), the **slot filling** version does allow such access.

3.5.2 (Non)arguments against the Two Versions of One-step Models

Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) refer to Pallier et al.'s (1993) results (presented below) as a possible support for two-step models. Although the reason for interpreting Pallier et al.'s results in such a way was not explicitly stated, probably the reason was that their results seemed to argue against the **suprasegmental matching** version; under their assumption that two-step models and the **suprasegmental matching** version of one-step models are the only candidates, that would imply a support for two-step models. Furthermore, Dupoux et al. (2011:208–209) take Kabak & Idsardi's (2007) Berent et al.'s (2007) and Moreton's (2002) results as potential evidence for two-step models and against one-step models.

This subsection reviews those results and points out that (i) Pallier et al.'s (1993), Kabak & Idsardi's (2007), Berent et al.'s (2007) and Moreton's (2002) results are compatible with both version of one-step models, while (ii) Matthews & Brown's (2004) (not cited by Dupoux et al., 2011) indeed are incompatible with the **suprasegmental matching** version but are fully compatible with the **slot filling** version; thus they could constitute evidence only against the **suprasegmental matching** version at best, but not against the core idea of one-step models, to the extent that they are fully compatible with the **slot filling** version. Thus this subsection as a whole defends the core idea of one-step models. The crucial point is whether the given results constitute evidence for listeners' perceptual access to sub-syllabic elements (on the view that the suprasegmental units assumed in one-step models are syllables); such evidence would argue against the **suprasegmental matching** version, but not against the **slot filling** version, of one-step models.

A non-argument against one-step models 1: Pallier et al. (1993)

Pallier et al.'s (1993) goal was to argue for the perceptual reality of syllables as sublexical units. They conducted phoneme detection experiments with French and Spanish listeners, manipulating listeners' attention by varying the probability of target phonemes' positions within the stimuli; the assumption is that listeners' attention will be attracted to a certain position if the probability of target phonemes in that position is increased. More specifically, trials were divided into 'inductors' and 'tests', and and one or two inductor trials were presented before each test trial; the inductor trials' role was to attract listeners' attention to a specific position within the stimuli. For example, if the inductor trial is *b* detection within *dou-blure*,[39] listeners' attention would be drawn to the onset of the second syllable, and a subsequent test trial of *p* detection within *ca-price* would count as phoneme detection in the attended position, whereas if the inductor is *b* detection within *sub-merge*, the same test trial would count as phoneme detection within a non-attended position. The results suggested that

(i) the detection of /C_1/ (e.g., 'p') within /…C_1-C_2…/ (e.g., *cap-ture*) is faster if the listeners attend to the coda of the first syllable within the stimuli than if they attended to the onset of the second syllable within the stimuli,

(ii) the detection of /C_1/ (e.g., 'b') within /…C_1C_2- …/ (e.g., *ta-bleau*) is faster if they attended to the onset of the second syllable than if they attended to the coda of the first syllable.

Pallier et al. draws the conclusion from such observations that listeners do construct syllabic representations.[40] If such results were interpreted as effects of attention allocation on the *codas* or *onsets* of syllables, they would suggest the perceptual reality of such syllable-internal

[39] In the descriptions of their results, a hyphen indicates a syllable boundary.

[40] They further argue for the non-lexical nature of such results. However, because their results are reviewed here only because they are cited by Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) as possible evidence against one-step models, their argument for the non-lexical nature is omitted here. (They do not contradict the conclusion above of the sublexical reality of phonotactic sensitivity; if the relevant phonotactic domains are syllables, then, their conclusion is already subsumed by the above conclusion.)

positions as *codas* or *onsets*. Thus, to the extent that the suprasegmental units relevant for **suprasegmental matching** are syllables, the results would seem to argue against **suprasegmental matching**, because they would suggest that the perceptually real suprasegmental units should not be atomic wholes, as claimed by the **suprasegmental matching** version of one-step models, but rather structured objects, as claimed by two-step models (or the **slot filling** version of one-step models).

However, Pallier et al.'s (1993) results do not necessarily argue against the **suprasegmental matching** version of one-step models. Note that the manipulation of attention was accomplished by the frequency of the positions (the first syllable's coda and the second syllable's onset) hosting the targets within the inductor trials. With such a manipulation, the attention could be interpreted either as attention on the first syllable's coda vs. the second syllable's onset, or as attention on the first syllable vs. the second syllable. Under the second interpretation, which is compatible with Pallier et al.'s conclusion that syllables are perceptually real, Pallier et al.'s results are compatible with the **suprasegmental matching** version of one-step models, according to which suprasegmental units such as syllables are *the* perceived objects, and phonemes are recognized only through a post-perceptual meta-analysis of the perceived suprasegmental objects. Under the **suprasegmental matching** version of one-step models, then, the listeners' attending to the first syllable would means that the post-perceptual meta-analysis of the first syllable was given priority, and their attention to the second syllable would mean that the meta-analysis of the second syllable was given priority. Then the detection of the target phoneme should be faster when it is in the syllable which is meta-analyzed with priority.

Thus, against Dupoux, Kakehi, Hirose, Pallier, & Mehler's (1999) construal, Pallier et al.'s (1993) results do not specifically argue for two-step models (or the **slot filling** version of one-step models) and are fully compatible with the **suprasegmental matching** version of one-step models, although further experiments employing the attention allocation technique may or may not produce results incompatible with the **suprasegmental matching** version of one-step

models.[41]

A non-argument against one-step models 2: Kabak & Idsardi (2007)

Kabak & Idsardi's (2007) goal was to determine whether Japanese listeners' /u/ epenthesis observed by Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) should be attributed to (1) a sequential ban on consonant clusters (e.g., */bz/, hence [ebzo] being perceived as /ebuzo/), or to (2) a syllable structure (coda) constraint (e.g., */b/ as a coda, hence [ebzo] being perceived as /ebuzo/). Noting that (1) and (2) are rather hard to tease apart in Japanese. they resorted to Korean.

In Korean, /k/ and /l/ do appear in codas but are subject to certain phonological constraints; stops, including /k/, are subject to the nasalization constraint, which requires that stops should be realized as a nasal before another nasal (e.g., /...k-m.../ ⟶ [...ŋ-m...], with a hyphen indicating a syllable boundary); laterals are subject to the lateralization constraint, which requires the following nasals should be realized as another lateral (e.g., [...l-n...] ⟶ [...l-l...]). In other words, syllables with coda /k/ or /l/ can be realized as [...k̚] or [...l] when and only when they do not violate such constraints.[42]

On the other hand, strident consonants (e.g., /c/, /cʰ/, and /s/) neutralize to the unreleased [t] in codas.[43]

Thus, for example, both [pʰak̚-ma] and [pʰas-ma] are impossible surface forms, with the former violating the nasalization constraint and the latter violating the obligatory coda neutralization. If Korean listeners' perceptual epenthesis is incurred by ill-formed consonant *sequences*, both kinds of stimuli should induce epenthesis, because both [k̚-m] and [s-m] are ill-formed consonant sequences; however, if their perceptual epenthesis is incurred by ill-formed

[41] See the final chapter of this thesis.

[42] In Korean, coda stops are unreleased, and hence [...k̚], rather than [...k].

[43] When a syllable of the form /...C/ (e.g., /picʰ/ 'light', to be realized as [pit̚] when produced in isolation) is followed by another syllable of the form /V.../ (e.g., /i/, a nominative case marker), they are re-syllabified to /...-CV.../ (/pi-cʰi/) in Korean; in such a case, C is no longer in a coda position and hence a strident C does not neutralize to [t] ([pit͡ɕʰi]).

codas, only the latter kinds of stimuli (such as [p^{h}as-ma]) should induce epenthesis, because syllables such as [p^{h}ak$^{\neg}$] are fine as syllables. In an AX discrimination experiment, Kabak & Idsardi (2007) observed that Korean listeners fail to discriminate only the latter kinds of stimuli from their corresponding epenthetic versions. Thus they argue that epenthesis is driven by syllable structure constraints, rather than consonant sequence restrictions, because, if Korean listeners' epenthesis was driven by consonant sequence restrictions, they should have perceptually epenthesized a vowel in both kinds of stimuli, in which case their discriminations should have failed with the former kinds of stimuli as much as with the latter kinds of stimuli.

Dupoux et al. (2011:208–209) seem to interpret Kabak & Idsardi's (2007) argument as a threat to (the **suprasegmental matching** version of) one-step models, on the assumption that Kabak & Idsardi's results would suggest listeners' sensitivity to syllable-internal elements (codas). However, upon a closer reflection, such a result is compatible with (the **suprasegmental matching** version of) one-step models.

According to the **suprasegmental matching** version, a speech signal (i.e., the surface realizations) has to be matched against the repertoire of suprasegmental units (e.g., syllables). According to Kabak & Idsardi's (2007) description of the *production* grammar of Korean, /p^{h}ak/ can be realized as [p^{h}ak$^{\neg}$], which means that, under the **suprasegmental matching** version, [p^{h}ak$^{\neg}$] should be a good match with the /p^{h}ak/ syllable in the repertoire. However, the /p^{h}as/ syllable is realized as [p^{h}at$^{\neg}$], which means that it is [p^{h}at$^{\neg}$] that should be a good match with the /p^{h}as/, and [p^{h}as] should probably not match well with any syllable in the repertoire, resulting in perceptual resyllabification of [p^{h}as] into [p^{h}a] and [s…], accompanied by epenthesis after [s]. Thus Kabak & Idsardi's result does not conflict with the **suprasegmental matching** version of one-step models. (Of course, it is compatible with the **slot filling** version too.)

A non-argument against one-step models 3: Moreton (2002)

Moreton's (2002) goal was to determine whether perceptual assimilation of non-native clusters should be seen as a result of listeners' unfamiliarity with the clusters or as a result of abstract phonological constraints.

Both /dl/ and /bw/ are unattested onset clusters in English. However, Moreton (2002) has a theoretical reason to assume that /dl/ is more marked as an onset than /bw/ universally (a reason that is not directly relevant to the purpose of this thesis). Thus, Moreton reasoned, if English listeners exhibit more resistance to /dl/ perception than to /bw/ perception, that could only be because their perception is subject to the universal markedness, which is an abstract phonological constraint, not a matter of (un)familiarity with the clusters.

In Moreton's (2002) first experiment, synthetic non-word /CCæ/ monosyllables were presented to English listeners, where the first C was a continuum between [g] and [d] or between [g] and [b], and the second was a continuum between [l] and [w]. The task was to identify the clusters as /dw, dl, gw, gl/ or /bw, bl, gw, gl/. Significant bias was observed only against /dl/ identification.

In order to see whether this result should be interpreted in terms of the marked status of the /dl/ *onset*, rather than the /dl/ *sequence*, a second experiment was conducted in which a synthetic [æ] was added to the beginning of the /CCæ/ stimuli employed in the first experiment. The resulting stimuli would presumably be syllabified as /æC-Cæ/ (with the hyphen indicating a syllable boundary), and hence, if the result of the first experiment should be interpreted in terms of the marked status of /dl/ *onset*, no bias against /dl/ identification should be observed; in contrast, if the result of the first experiment should be interpreted in terms of the /dl/ *sequences*, the bias should persist. The result supported the former prediction; the bias went away.

However, listeners' perceptual bias against /bw/ as compared to /dl/ could be interpreted as suggesting listeners' perceptual bias either against /bw/ *onsets* or against /bw–/ *syllables*. If it

was interpreted as a bias against /bw/ *onsets*, that would imply that listeners have perceptual access to such syllabic positions as onsets, which is not expected under, and hence would constitute evidence against, the **suprasegmental matching** version of one-step models. However, if it was interpreted as a bias against /bw–/ *syllables*, it would not necessarily imply listeners' perceptual access to sub-syllabic positions such as onsets and hence would be fully compatible with the **suprasegmental matching** version of one-step models. Moreton's (2002) argument against an account of the bias in terms of unfamiliarity with unattested sound patterns would be unaffected by the choice of those two alternative interpretations, but the fact that both interpretations are possible means that Moreton's results do not distinguish among two-step models and the two versions of one-step models (a task Moreton was not concerned with).[44]

A non-argument against one-step models 4: Berent et al. (2007)

Berent et al.'s (2007) goal was similar to Moreton's (2002); it was to see the perceptual effects of the markedness of onset clusters in the universal sonority hierarchy. Since this thesis is not concerned with the typological utility of the sonority hierarchy, let us only concentrate on those aspects of their results that are directly relevant to the choice between two-step models and (the two versions of) one-step models.

Berent et al. (2007) conducted syllable counting, AX discrimination, and identity priming experiments with English and Russian listeners, with three kinds of stimuli classified according to the constitutions of the initial onset clusters:

- small sonority rise stimuli (e.g., /bnif/), whose initial onset clusters (underlined) are the least marked with respect to sonority

[44] In order to defend (the **suprasegmental matching** version of) one-step models, Dupoux et al. (2011) points out the possibility that marked clusters are harder to articulate than unmarked clusters and hence a schwa-like element is likely to have been inserted within the cluster in the experimental stimuli, in which case the presence of the schwa-like element could be responsible for alleged epenthesis. However, the /dl/ portions in Moreton's (2002) two experiments were synthetic and physically the same across the two experiments. Thus Dupoux al.'s (2011) suggestion does not apply to Moreton's results. As seen above, (the **suprasegmental matching** version of) one-step models could be defended without appealing to the postulated articulatory difficulty.

- sonority plateau stimuli (e.g., /bdif/), whose initial onset clusters (underlined) are more marked than 'small sonority rise' but less marked than 'sonority fall'

- sonority fall stimuli (e.g., /lbif/), whose initial onset clusters (underlined) are the most marked with respect to sonority

The onset clusters in the three kinds of stimuli are all unattested in English but are all attested in Russian; Russian syllables are more liberal than English syllables with respect to sonority. Berent et al. were interested in whether the likelihood of perceptual epenthesis within onset clusters is sensitive to the sonority difference, irrespective of the clusters are attested or not.

In the syllable counting experiments, the task was to count the number of the syllables contained in the stimuli, and the results suggest that (i) English listeners tend to erroneously count the monosyllabic stimuli as two (interpreted as suggesting perceptual epenthesis of a vowel within the clusters),[45] and the error rates increase as the sonority status becomes more marked, and (ii) Russian listeners are generally accurate, but still, their error rates are modulated by the sonority status in the same direction.

In the AX discrimination experiments, the task was discrimination between /CCVC/ vs. /CəCVC/, and the results suggest that (i) English listeners tend to fail (suggesting perceptual epenthesis of a schwa) more with sonority plateau stimuli than with sonority rise stimuli, (ii) their responses get slower as the sonority status becomes more marked, and (iii) Russian listeners' responses were slower with sonority fall stimuli than with sonority plateau stimuli.

In the identity priming experiments, listeners were presented with a series of two stimuli, for each of which they had to make a lexical decision. In the critical trials, the targets were preceded by identical primes (e.g., /lbif/–/lbif/; /bdif/–/bdif/) or their epenthetic counterparts

[45] It is often assumed that nasals could constitute a nucleus (e.g., *button* being syllabified into /bʌ/ and /tn̩/). If that assumption was applied to 'small sonority rise stimuli' such as /bnif/, the stimuli would be counted as two syllables not because of epenthesis between /b/ and /n/, but rather as /n/ being perceived as a syllabic /n/. However, the 'syllabic nasal' assumption would not apply to 'sonority plateau stimuli' such as /bdif/ or 'sonority fall stimuli' such as /lbif/, and hence counting 'sonority plateau/fall stimuli' as two syllables could not be interpreted in terms of syllabic nasals.

(e.g., /ləbif/–/lbif/; /bədif/–/bdif/); perceptual epenthesis should make the latter case true identity priming situations and hence the magnitude of identity priming should reflect perceptual epenthesis. English listeners' behavior with sonority plateau stimuli was compared with their behavior with sonority fall stimuli, and identity priming was observed with sonority plateau stimuli but not with sonority fall stimuli; for English listeners, the sonority fall stimuli targets benefit from an epenthetic prime (e.g., /ləbif/–/lbif/) as much as from an identity prime (e.g., /lbif/–/lbif/), but such was not the case with sonority plateau stimuli.

All those results suggest that English listeners' perceptual epenthesis differs among unattested clusters on the one hand, and, more weakly, that Russian listeners' epenthesis differs among attested clusters on the other.[46] However, the sonority status *of onsets* could also be interpreted as the sonority status *of syllables*. For example, listeners' perceptual bias against 'sonority fall stimuli' such as /lbif/ as compared to 'sonority plateau stimuli' such as /bdif/ could be interpreted as a bias against a /lb/ *onset* or against a /lb–/ *syllable*. If it is interpreted as a bias against a /lb/ onset, that would suggest listeners' perceptual access to sub-syllabic elements (in this case, onsets), in which case either two-step models or the **slot filling** version of one-step models should be chosen over the **suprasegmental matching** version of one-step models. However, if it is interpreted as a bias against a /lb–/ syllable, that would be fully compatible with the **suprasegmental matching** version of one-step models. Because both interpretations are possible, Berent et al.'s (2007) results do not distinguish among two-step models and the two versions of one-step models.[47]

[46] Alice Turk (p. c.) points out the possibility that what Berent et al. (2007) analyze as 'small sonority rise' consonant clusters (/bn/) do exist in English listeners' syllabic repertoire, because the second syllable of words such as *open* could be realized as [pn̩]. That would explain the difference between English listeners' perceptual behavior with 'small sonority rise stimuli' (/bn/) vs. with the other two kinds of stimuli, but that would not explain the difference between their behavior with 'sonority plateau stimuli' vs. with 'sonority fall stimuli', neither of which involve /n/.

[47] Recall that Dupoux et al. (2011) attempted to defend (the **suprasegmental matching** version of) one-step models by appealing to the possibility that marked clusters are harder to articulate than unmarked clusters and hence a schwa-like element is likely to have been inserted within the cluster in the experimental stimuli, in which case the presence of the schwa-like element could be responsible for alleged epenthesis. However, Berent et al.'s (2007) stimuli included utterances by native speakers of Russian, which does allow the marked clusters in question; it is rather unnatural to assume that native speakers of a language allowing such clusters have difficulty producing such clusters and insert a schwa-like element. As noted above, (the **suprasegmental matching** version of) one-step

An argument against the suprasegmental matching version: Matthews & Brown (2004)

As seen above, Pallier et al.'s (1993), Kabak & Idsardi's (2007), Moreton's (2002) and Berent et al.'s (2007) results are all compatible not only with two-step models and the **slot filling** version of one-step models but also with the **suprasegmental matching** version of one-step models. However, Matthews & Brown's (2004) results are rather hard to interpret in terms of the **suprasegmental matching** version of one-step models, but are fully compatible both with two-step models and with the **slot filling** version of one-step models.

Matthews & Brown (2004) had multiple goals in conducting their cross-linguistic speeded AX discrimination experiment with Japanese and Thai listeners: (1) to examine whether perceptually relevant phonotactics should be seen as familiarity with the attested phoneme sequences encountered in the native language, or as structural constraints stated in terms of such notions as 'consonant' and 'vowel', and (2) to examine whether phonotactic effects differ depending on whether the listeners are involved in pre-phonological or phonological processing. We have already noted that /u/ is the default epenthetic vowel in Japanese, but, according to Matthews & Brown, /a/ is the default epenthetic vowel in Thai. Their stimuli involved /-kt-/, /-pt-/, and /-bd-/ clusters, and they examined Thai listeners' discrimination between such /-CC-/ clusters and /-CaC-/ on the one hand, and Japanese listeners' discrimination between /-CC-/ clusters and /-CuC-/ on the other. It is the results concerning (1) that are relevant here, so this review concentrates on (1).[48]

In Thai, according to them, CC clusters are structurally fine, but /–bd–/ clusters are unattested. Thus, if perceptually relevant phonotactics should be seen as familiarity with the attested phoneme sequences (or the probabilities of phoneme sequences), not abstract structural constraints, Thai listeners should epenthesize the default /a/ within the /–bd–/ clusters and hence fail to discriminate /–bd–/ and /–bad–/. In contrast, if perceptually relevant phonotactics

models could be defended without appealing to the assumed articulatory difficulty.

[48]The results come from Thai listeners' behavior; Japanese listeners' behavior serves as a control.

are abstract structural constraints on C- and V-slots within some relevant phonotactic domains (e.g., syllables or bi-phones), according to which consonant clusters are fine (whatever the consonants are), Thai listeners should not epenthesize /a/ within /–bd–/ and hence should succeed in discriminating /–bd–/ and /–bad–/. Matthews & Brown (2004) observed that Thai listeners did *not* epenthesize /a/ within /–bd–/ clusters, suggesting that their perception is guided by whether consonant clusters in general are permitted or not, rather than familiarity vs. unfamiliarity with specific clusters.[49] Based on this result Matthews & Brown argue that perceptually relevant phonotactics should be seen not as familiarity with the phoneme sequences encountered in the native language, but rather as structural constraints stated in terms of such notions as 'consonant' and 'vowel'.

This result suggests that listeners do parse the speech signal in terms of C- and V-slots. Such a result is rather hard to interpreted in terms of the **suprasegmental matching** version of one-step models, which claims that perception consists of the matching of the speech signal with the repertoire of the suprasegmental units such as syllables; suprasegmental units (syllables) containing /–bd–/ must not be in the acquired repertoire and hence should perceptually assimilate to another within the acquired repertoire (/–bad/–).[50]

However, this result is fully compatible not only with two-step models but also the **slot filling** version of one-step models, both of which claim that constituents of phonotactic domains are indeed perceived. In the case of two-step models, /–bd–/ sequences perceived in the first phonemic categorization stage does not violate the structurally stated phonotactic constraint and hence should not incur perceptual epenthesis. In the case of the **slot filling** version of one-step models, structural frames are available with appropriate slots for the /b/ and /d/ phonemes

[49]The Thai listeners were exactly those who, in a separate perception study, exhibited /a/ epenthesis when the relevant stimuli violate structurally stated phonotactic constraints. Thus the possibility of non-perceptual nature of /a/ epenthesis in Thai can be excluded.

[50]Again recall Dupoux et al.'s (2011) suggestion of the possibility of a schwa-like element in marked consonant cluster stimuli, which could be responsible for listeners' epenthetic behavior against such stimuli. Such a suggestion does not apply to Matthews & Brown's (2004) results, because the crucial observation was Thai listeners' *non*-epenthetic behavior.

and hence there should be no need for perceptual epenthesis.

Conclusion from the subsection

In this subsection, Pallier et al.'s (1993), Kabak & Idsardi's (2007), Moreton' (2002), and Berent et al.'s (2007) results on the one hand, and Matthews & Brown's (2004) results on the other, were reviewed. Those results (except those by Matthews & Brown's) were cited by Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) or by Dupoux et al. (2011) as potential evidence against one-step models, because (i) such results seem to suggest listeners' perceptual access to sub-constituents of the suprasegmental units relevant for phonotactics, and (ii) the **suprasegmental matching** version was assumed to be the only possible version of one-step models. It was pointed out above that Pallier et al.'s (1993), Kabak & Idsardi's (2007), Moreton' (2002) and Berent et al.'s (2007) results do not argue against (the **suprasegmental matching** version of) one-step models because they do not necessarily constitute evidence for listeners' perceptual access to sub-constituents of syllables; thus they are fully compatible not only with two-step models but also with both versions of one-step models. In contrast, Matthews & Brown's (2004) results do seem to constitute evidence for listeners' perceptual access to sub-constituents of syllables or bi-phones and hence seem to argue against the **suprasegmental matching** version of one-step models. However, that does not mean that Matthews & Brown's results argue against one-step models as a whole; they are fully compatible with the **slot filling** version of one-step models. Thus their results could constitute evidence against the **suprasegmental matching** version of one-step models at best; not only two-step models but also the **slot filling** version of one-step models remain as valid candidates.[51]

[51] On the other hand, Moreton's (2002), Berent et al.'s (2007) and Matthews & Brown's (2004) results suggest that listeners' behavior could not be explained solely in terms of (un)familiarity with attested patterns and hence argue against lexicalist reductions too. Pitt (1998) had also argued that Massaro & Cohen's (1983) results could not be accounted for in terms of attested phoneme sequence frequencies.

Such results, if valid, argue against the view that phonotactics should be reduced to relative frequencies (probabilities) of phoneme sequences, with "phonotactically prohibited sequences" being the extreme cases of zero probability (Pierrehumbert, 1994, 2001, 2003), as far as perception is concerned. Of course, this does not deny the claim that relative probabilities have some effect (Bailey & Hahn, 2001; Coleman and Pierrehumbert, 1997; Frisch et al., 2000; Hay et al., 2000; Juszcyk et al., 1994; Luce and Large, 2001; Pitt and McQueen, 1998; Treiman et al.,

3.5.3 Arguments for one-step models

Next we consider arguments for one-step models by Dehaene-Lambertz et al. (2000) on the one hand, by Dupoux et al. (2011) on the other. It will be argued that their results are inconclusive as arguments for one-step models; Dehaene-Lambertz et al.'s have failed to provide a cohesive interpretation of their overall results (which this thesis cannot provide, either), and Dupoux et al.'s results allow an alternative and cohesive interpretation in favor of two-step models.

Dehaene-Lambertz et al. (2000)

Dehaene-Lambertz et al. (2000) employed a mismatch detection task and measured ERP's (event-related potentials), and based on the results, argue for one-step models.[52] A series of precursor items /igumo/ (or /igmo/), repeated three times each uttered by a different female speaker, was presented, after which the test item, uttered by a male voice, was presented; the test item was /igmo/ (or /igumo/) in the experimental condition and /igumo/ (or /igmo/) in the control condition. The task was to detect a change in the stimuli (mismatch detection). In other words, in the experimental condition, listeners heard /igumo, igumo, igumo, igmo/ or /igmo, igmo, igmo, igumo/, and pressed a button when the underlined stimuli are heard, while in the control conditions, they heard /igumo, igumo, igumo, igumo/ or /igmo, igmo, igmo, igmo/. The behavioral data replicated Dupoux, Kakehi, Hirose, Pallier, & Mehler's (1999) results; Japanese listeners, but not French listeners, tended to fail in mismatch detection (because, according to their interpretation, /u/ was perceptually epenthesized by Japanese, but not by French, listeners).

The electro-physiological data were rather complicated. They first identified three temporal regions where French listeners exhibited significant differences across the experimental

2000; Vitevitch and Luce, 1998, 2005; Vitevitch et al., 1999); it is just that frequencies do not exhaust (perceptually relevant) phonotactics.

[52] Dehaene-Lambertz et al. (2000) only have the **suprasegmental matching** version in mind (which they call 'Coarse Coding Models'). However, the (in)validity of their argument for one-step models does not depend on the choice between the **suprasegmental matching** and the **slot filling** version.

and the control conditions. The ERP components in the first time window were interpreted as Mismatched Negativity (MMN), which reflect "very early speech processing" (p. 636) and are "elicited when an acoustical mismatch is detected" (p. 642). As for the components in the second window, they suggest the possibility of interpreting them as a phonological mismatch (PMM), but they are not definitive (see below). The component in the third window was interpreted as a late positive complex (LPC), which "is known to be due to the conscious detection of a less frequent stimulus and is modulated by the response decision" (p. 643). They then examined Japanese listeners' ERP components in those three windows. According to two-step models, phonotactic repairs should follow the stage of epenthesis-free perception, a stage where the mismatching target should have sounded different from the precursor items, but according to one-step models, such a stage should not exist; Dehaene-Lambertz et al. (2000) reasoned that the existence of such a stage should result in significant MMN differences.

The MMN difference across the conditions, which was significant with French listeners, was absent (or shorter and weaker) with Japanese listeners, suggesting Japanese listeners' insensitivity to the mismatch and supporting the expectation from one-step models. Thus the MMN results from the first window were interpreted as suggesting that "phonotactics play a very early role that probably goes back to the coding of phonetic properties" (p. 643).

However, the components in the second window exhibited a significant effect of the mismatch not only with French but also with Japanese listeners, suggesting Japanese listeners' sensitivity to the mismatch. The component in the third window, again, failed to exhibit a significant difference across the conditions with Japanese listeners. If the first-window results reflect Japanese listeners' post-epenthetic perception, and if second-window results reflect a later stage of perceptual process than first-window results, the significant second-window results would sound rather 'paradoxical' (p. 643), because it would seem to suggest epenthesis-free perception (as phonological mismatch, according to the non-decisive characterization above of the second window) following epenthesis (as acoustic non-mismatch, according to the above

characterization of the first window). Dehaene-Lambertz et al. (2000:644) suggest two possible interpretations of the discrepancy between the first- and the second-window results with Japanese listeners:[53] (i) the results from the first window (44 ms) failed to reach significance because deviance detection by Japanese listeners are not as 'specific' (to this window timing) or as 'automatic' (to be early enough for this window) as French listeners and hence the cross-conditional brain activity differences were lost when averaged, while the results from the second window reached significance because it reflects accumulated results from a longer window (128 ms); (ii) the first- and the second-window results are due to distinct subsystems.[54] Unfortunately, neither suggestion seems to support their conclusion; according to (i), Japanese listeners would go through epenthesis-free perception before phonotactic repairs, which failed to be observed in the first window, and according to (ii), epenthesis-free perception would coexist with automatic phonotactic repairs. Thus a fair verdict would be that their results still lack a cohesive interpretation.[55]

Dupoux et al. (2011)

Dupoux et al. (2011) note that, according to two-step models, phonotactic repair should not exhibit coarticulation sensitivity, because subphonemic details such as coarticulation traces should not be available in the initially constructed phonemic strings on which phonotactic

[53] As for the third-window results, they only remark methodological difficulties inherent in ERP data analysis (and hence the difficulty in interpreting ERP data).

[54] As candidates for the subsystem responsible for the second-window results, Dehaene-Lambertz et al. suggest 'a prototypicality system' and 'a phonetic system that keeps track of the phonemes presented'.

[55] Dehaene-Lambertz et al. (2000) were followed by Jacquemot et al. (2003), who used fMRI to examine localization of brain activity when listeners are engaged in perceptual tasks. They noted that the discrimination between VCCV and VC V̲CV (consonant clusters vs. CVC) should be possible phonologically for French listeners but only acoustically for Japanese listeners on the one hand, and the discrimination between VC V̲CV and VCV̲V̲CV (vowel length contrasts) should be possible phonologically for Japanese listeners but only acoustically for French listeners on the other. They conflated French listeners' CC/CVC discriminations and Japanese listeners' vowel length discriminations on the one hand (phonological discriminations), and French listeners' vowel length discriminations and Japanese listeners' CC/CVC discriminations on the other (acoustic discriminations). They then compared the brain localizations of French and Japanese listeners (combined) in the conflated phonological discriminations, to the localizations of French and Japanese listeners (comibined) in the conflated acoustic discriminations.

However, the 'phonological' conflation presupposes that the CC/CVC contrasts for French listeners and the vowel length contrasts for Japanese listeners are the same kinds of contrasts, a view dictated by the **the suprasegmental matching** version of one-step models but denied by two-step models. Thus, as an attempt to choose from those two views, their design presupposes the conclusion.

repair should operate; in contrast, according to one-step models, phonotactic repair should operate on the pre-categorical input and should be able to exhibit sensitivity to coarticulation details available within the input. Thus they reasoned that the predictions of two-step models and one-step models could be teased apart by seeing whether perceptual epenthesis exhibits coarticulation sensitivity.

Dupoux et al. (2011) created two kinds of stimuli from French speakers' natural productions: two-consonant clusters produced without a vowel in between ("natural clusters") on the one hand, and two-consonant clusters created by removing /i/ or /u/ originally produced within the clusters ("coarticulated clusters") on the other.[56] The stimuli were presented to listeners of Japanese, European and Brazilian Portuguese in identification and discrimination experiments. The results can be summarized as:

(A) Japanese listeners tended to perceive /u/ within natural clusters (a replication of Dupoux, Dupoux, Kakehi, Hirose, Pallier, and Mehler' 1999 results).

(B) Brazilian Portuguese listeners tended to perceive /i/ within natural clusters (as expected from /i/ epenthesis observed in loanwords).

(C) Japanese listeners tended to perceive /i/ within /i/-coartciualted clusters.

(D) Brazilian Portuguese still tended to perceive /i/ even within /u/-coarticulated clusters.

(E) European Portuguese listeners did not epenthesize either within natural or coarticulated clusters.

(A) (=/u/ perception by Japanese listeners) and (B) (=/i/ perception by Brazilian Portuguese listeners) are observations of "default" epenthetic vowels, default in the sense that no coarticulation cues dictate the specific choice of epenthetic vowel. It is (C) (= /i/ perception by Japanese

[56] As was done by Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999), Dupoux et al. (2011) removed the vowels successively. We will only focus on the results with the extremest stimuli in which the vowels were completely removed.

listeners) that Dupoux et al. (2011) interpret as evidence for one-step models, because, in their interpretation, it is an observation of Japanese listeners' default-overriding epenthesis, and the source of the default-overriding identity of the epenthesized vowel must be coarticulation traces left by /i/, which should be available in the pre-categorical input but not in categorized phoneme strings.

However, their interpretations not only of (C) (= Japanese listeners' /i/ perception) but also of (D) (= Brazilian Portuguese listeners' /i/ perception) and of (E) (= European Portuguese listeners' non-epenthetic perception) are problematic.[57] Let us consider (E), (D) and (C) (in this order).

Dupoux et al. (2011) attribute (E) (= European Portuguese listeners' non-epenthetic behavior) to unstressed vowel deletion in European (as opposed to Brazilian) Portuguese, as a result of which, according to Dupoux et al., the consonant clusters in Dupoux et al.'s stimuli arise in natural productions in European Portuguese. For Dupoux et al., (E) suggests that epenthesis-free perception is enabled by "surface form" phonotactics; the consonant clusters are illegal phonemically but fine at "surface." To the extent that those clusters are claimed to be illegal phonemically, then, Dupoux et al.'s interpretation implies that, betraying the traditional conception of 'phonotactics' as constraints on phoneme strings, a single language has two distinct phonotactics: phonemic phonotactics and "surface" phonotactics.

However, they note that, according to some researchers, vowel devoicing in Japanese sometimes results in 'complete vowel deletion', which Dupoux et al. (more or less reluctantly) con-

[57] In fact, their account of the choice of the epenthetic vowel in Japanese and in Brazilian Portuguese (A)–(B) could also be problematic. They appeal to (i) the assumption that both /i/ and /u/ tend to be devoiced both in Japanese and in Portuguese, and (ii) the assumption that /u/ and /i/ are the shortest vowels respectively in Japanese and Portuguese. According to Dupoux et al. (2011), those two assumptions account for the choice of the default epenthewtic vowel in each langugae because the results of epenthetizing /i/ (in Brazilian Portuguese) or /u/ (in Japanese) should constitute better matches to [CC] than the results of epenthetizing the other vowels.

However, devoicing and physical duration in question here are both sub-phonemic details. Thus to attribute the choice of the default vowel to such sub-phonemic details is to presuppose the core idea of one-step models that sub-phonemic details affect epenthesis.

Indeed, this potential problem is not serious. If one-step models are indeed correct, such an account is coherent; it is just that the account of the choice of the default epenthetic vowels is presented before the superiority of one-step models over two-step models is concluded, thereby presupposing the conclusion.

cede. If so, we would have to assume that /-C_1VC_2-/ could be realized as [-C_1C_2-] in Japanese too, which would mean that their account of the European Portuguese results would imply that Japanese listeners should *not* epenthesize, contrary to their own observations.[58]

In fact, their comparison between European Portuguese and Japanese results seems rather misguided. Dupoux et al. (2011) cite Varden (1998) as claiming the existence of vowel deletion in Japanese, but, as seen in Chapter 2 above, for Varden (1998), as well as for Kondo (1997), 'vowel deletion' means the loss of a vowel that leaves no spectral trace and should be distinguished from "deletion in a broad sense, leaving coarticulation traces." To remind the readers of the distinction, if a [-C_1C_2-] realization of /-C_1VC_2-/ has some coloring of /V/ on [C_1], that is an instance of vowel deletion in a broad sense; in order for the [-C_1C_2-] realization to count as an instance of vowel deletion in the strict sense, no such coloring should be present in [C_1]. Varden does claim to have observed Japanese speakers' vowel deletions in this strict sense, but he also notes that they were very rare. This observation, together with Kondo's (1997) and Varden's (1998) observations that vowel devoicing in Japanese is a rather gradient process, ranging from partial devoicing to "deletion in a broad sense" and (very rare) strict deletion, suggests that 'vowel devoicing' in Japanese usually retains vowel gestures, which in turn suggests that phonologically or phonemically vowels do exist; it is only that their physical (or allophonic) realizations are sometimes nothing more than 'coarticulation traces within the preceding consonant' rather than pure vocalic durations or 'whispered vowels'.[59]

Such a distinction between strict deletion and 'deletion in a broad sense' suggests that, if phonotactics are understood as constraints on phoneme strings in the conventional manner,

[58] Dupoux et al. themselves were aware of this problem (and possibly that is why they mention vowel deletion in Japanese only in the General Discussion section, not in the Introduction section). In footnote 7 they suggest that the effects of 'complete vowel deletion' can be examined through a comparison between fricative and stop contexts, assuming that a vowel /V/ in /C_1VC_2/ tends to devoice more when /C_1/ is a fricative than when it is a stop. However, they do not specify why such a comparison would help examine possible effects of vowel 'deletion' (as opposed to mere devoicing); the assumption is simply that vowel devoicing rates, not vowel 'deletion' rates, differ across fricative and stop contexts.

[59] Faber & Vance (2000) also argue, based on a production study, that, even when a vowel gets devoiced, phonemically the vowel does exist, in Japanese; see also Nakamura (2003), already mentioned, who observed electropalatographical evidence of vowel gestures.

/-C_1C_2-/ is phonotactically fine in a language that 'deletes' /V/ in /-C_1VC_2-/, while it is phonotactically bad in a language that can only make /V/ 'deleted in a broad sense'. If European Portuguese deletes vowels in the strict sense, but Japanese can only delete them in a broad sense at best, /C_1C_2/ is phonotactically fine in European Portuguese but not in Japanese, in which case [-C_1C_2-] would count as a legitimate realization of /-C_1C_2-/, and hence should sound distinct from [-C_1VC_2-], for the listeners of European Portuguese, but not for listeners of Japanese. Thus [-C_1C_2-] occurrences in natural production due to vowel deletion, where 'vowel deletion' is understood to encompass both the strict and the broad sense, do not automatically guarantee that they will be perceived as /-C_1C_2-/ sequences.

Dupoux et al.'s (2011) problem here is that they fail to distinguish vowel deletion in the strict sense on the one hand, and vowel deletion in the broad sense on the other; if the /V/ in /-C_1VC_2-/ undergoes strict deletion, the result would be an underlying /-C_1C_2-/, which would surface as [-C_1C_2-] with no coloring of V in C_1; if it undergoes vowel deletion in the broad sense, the result would still be /-C_1VC_2-/ but would surface as [-$C_1^VC_2$-] with V's coloring in C_1. In fact, for their interpretation of the European Portuguese results to make sense, we would have to assume that [-C_1C_2-] and [-C_1VC_2-] are somehow distinguished by ordinary native speakers of European Portuguese. If so, it would be possible (and probably more natural) to say that /-C_1C_2-/ and /-C_1VC_2-/ are both phonemically possible in European Portuguese; it is simply that those words which accept vowel deletion have two alternative phonemic forms.[60] On the other hand, this does not apply to Japanese; only /-C_1VC_2-/ is possible, but V is allophonically realized either as a voiced or voiceless "vowel" duration or as some coloring within /C_1/. If so, C_1C_2 clusters would be phonotactically fine in European Portuguese, even if phonotactics is conceived in terms of phoneme strings (the traditional conception), but not in Japanese. This would explain the observed contrast between European Portuguese and Japanese.

[60]The fact that ordinary listeners seem to have no difficulty in lexical access in spite of massive 'surface' variability in phonetic forms suggest that the phonological parts of the entries in the mental lexicon are more or less underspecified. The 'two alternative phonemic forms' idea is only an extension of that suggestion.

This 'two phonemic forms' interpretation accounts for Dupoux et al.'s (2011) observation at least equally as Dupoux et al.'s own interpretation in terms of 'surface phonotactics'; note that the 'two phonemic forms' interpretation is neutral with respect to the choice between two-step models and one-step models. Indeed, under Dupoux et al.'s own interpretation, the results with European Portuguese listeners would suggest 'phonotactic' effects sensitive to subphonemic details, in the sense that non-epenthetic perception is claimed to be enabled by the specific ([C_1C_2]) realization patterns of the alleged /-C_1VC_2-/ sequences, in which case they could (in addition to the results with Japanese listeners) be taken as evidence in favor of one-step models (although Dupoux et al.'s do not propose to see their results with European Portuguese as evidence for one-step models). However, to the extent that such a 'two phonemic forms' interpretation is possible, then, Dupoux et al.' European Portuguese results do not decide the choice between one- and two-step models.

Next consider (D) (= Brazilian Portuguese listeners' /i/ perception). The observation was that Brazilian Portuguese listeners' epenthesis was insensitive to coarticulation, failing to override the default /i/ epenthesis based on coarticulation cues. This observation is just as expected from two-step models, rather than one-step models, and hence could rather be taken as evidence against Dupoux et al.'s (2011) own conclusion that one-step models should be employed. Dupoux et al. interpreted (D) as suggesting that "the exploitation of coarticulation cues is not due to universal auditory processes, but rather, is integrated within the language-specific process of segmental categorization" (p. 207). This is presented as an account of why Brazilian Portuguese listeners failed to exhibit coarticulation sensitivity in contrast to Japanese listeners (= (C)), but it remains totally unexplained what differences in 'segmental categorization' in the two languages exist and how they explain the observation. Indeed one-step models only predict the *possibility* of coarticulation sensitivity, not that listeners will *always* exhibit coarticulation sensitivity, and hence the Brazilian Portuguese results do not necessarily argue against one-step models. However, in the absence of a specific account of why Japanese and Brazilian

Portuguese listeners differed in the observed way, proponents of two-step models could appeal to the Brazilian Portuguese results, questioning the validity of (the interpretation of) the Japanese results.

However, Dupoux et al.'s (2011) characterization of Brazilian Portuguese seems to be too simplistic. According to Mateus and d'Andrade (2000:sections 2.2.2.1–2.2.2.2),[61] Brazilian Portuguese has seven vowel phonemes in stressed positions, and [u] in pre-stressed non-word-final positions is a realization of /u/, but [u] in post-stressed non-word-final positions is a reduced form not only of /u/ but also of /ɔ/ and /o/. If so, coarticulation sensitivity may or may not result in successful perception of /i/ depending on the stress pattern of the stimuli; if the epenthesis site in the middle of each stimulus was a pre-stressed position, coarticulation sensitivity should lead to /u/ perception, but if it was a post-stressed position, Brazilian Portuguese listeners could have had trouble determining that the vowel was /u/, rather than /ɔ/ or /o/, even if they were coarticulation sensitive. However, Dupoux et al. do not specify the stress pattern of their stimuli. Thus it is not clear what should be concluded from their results; their seeming coarticulation insensitivity could be in fact due to their coarticulation insensitivity, but could alternatively be due to their difficulty in choosing /u/, rather than /ɔ/ or /o/.

In fact, if the epenthesis sites were post-stressed positions, one possible account for the observed results would be something like the following. First, if we accept Dupoux et al.'s (2011) characterization of the difference between European and Brazilian Portuguese, vowels only devoice, but do not delete, in Brazilian Portuguese. If so, Brazilian Portuguese listeners are used to hearing [$C^{V}\underset{\circ}{V}$] realizations of /CV/, but, in contrast to Japanese listeners, who are used to [C^{V}] realizations of /CV/ not accompanied by [$\underset{\circ}{V}$], Brazilian Portuguese listeners do not necessarily have to resort to the coarticulation traces (i.e., the superscripted V) to perceive the identity of the V; [$\underset{\circ}{V}$] should offer enough cues. Then it would be natural that they have not developed as much coarticulation sensitivity as Japanese listeners have. (Recall the results

[61] I only have a Kindle version, which does not give the information of the page numbers of the paper edition.

reported by Kakehi et al., 1996; Japanese listers' sensitivity to pre-closure formant transition cues for the place of a stop is much weaker than Dutch listeners', presumably because Dutch listeners have to perceive the place from the pre-closure cues in their daily linguistic activity, whereas Japanese listeners do not. Sensitivity to a cue will develop more if that cue is the only available cue than when other cues are also available.) Then Brazilian Portuguese listeners' rate of coarticulation-sensitive responses to /u/-coarticulated stimuli should be low. Furthermore, if the epenthesis sites were post-stressed positions, they would experience further difficulty in deciding that the perceived [u] should be interpreted as /u/, rather than /ɔ/ or /o/, further reducing the rate of /u/ responses. (For example, if the perceived [u] is categorized as /u/, /ɔ/ or /o/ by chance, the 72 % rate of perceiving [u] should manifest only as 26 % /u/ responses.) In such a situation, seemingly coarticulation-insensitive results would be expected for coarticulation-sensitive Brazilian Portuguese listeners.

Of course, the above scenario is only a possibility. As stated above, Dupoux et al. (2011) do not specify the stress pattern of the stimuli, and hence it is not clear at all whether the epenthesis sites were pre- or post-stressed positions. Furthermore, what vowel allophones in what positions are reduced forms of the /u/ phoneme is not totally clear (at least to me; for example, Azevedo's 2005 characterization differs from that by Mateus and d'Andrade, 2000, in that both [o] and [u] are claimed to occur in post-stressed positions as reduced forms of /o/). Thus the above scenario is only a possible solution; crucial information (a convincing detailed phonological analysis of Brazilian Portuguese, as well as the stress pattern of the stimuli) is lacking and hence the implications of the Brazilian Portuguese results cannot be decisively evaluated.

Finally consider (C) (= Japanese listeners' default-overriding /i/ perception). Part of the possible solution suggested above for the problem of the difference between Japanese vs. Brazilian Portuguese assumed the devoicing-based phonemic categorization story for Japanese; Japanese listeners have to cope with [C^V] realizations of /CV/, but Brazilian Portuguese listen-

ers only have to deal with [$C^{V}\underset{\circ}{V}$] realizations, and hence Japanese listeners' more sensitivity to coarticulation cues within C (the superscripted V) than Brazilian Portuguese listeners. However, the possibility of such an account in turn seems to cast some doubt on the validity of Dupoux et al.'s (2011) overall conclusion that their results with Japanese listeners constitute evidence against 'two-step models'.

Recall Dupoux et al.'s (2011) overall logic. According to two-step models, phonotactic constraints operate on the phoneme strings constructed as a result of first-step phonemic categorization; the second-step parsing operation scans this phoneme string, and if a phonotactic violation is discovered, the parsing operation repairs it. Since subphonemic distinctions such as coarticulation traces should not be available in this phoneme string, the phonotactic repair made by the second-step parsing operation should be insensitive to coarticulation. Thus an examination of whether epenthesis is sensitive to coarticulation should tell us the (in)validity of two-step models.

However, according to the possible solution suggested for the difference between Japanese and Brazilian Portuguese listeners, Japanese listeners' sensitivity to /i/ coarticulation arises at least partially from phonemic categorization, rather than by phonotactic repairs; [C^{i}] is phonemically categorized as /Ci/, rather than /C/, in which case there should be no phonotactic violation in the first place, and their /i/ perception should not be an instance of epenthesis. If Japanese listeners' observed /i/ perception is mostly due to such coarticulation-sensitive *phonemic categorization*, the Japanese listeners' observed perceptual behavior would not constitute evidence for coarticulation-sensitive *phonotactic repairs*; the observed coarticulation sensitivity could be interpreted, within two-step models, as coarticulation sensitivity at the first-step phonemic categorization stage, rather than coarticulation sensitivity at the second-step phonotactic repair stage.

The (in)validity of this alternative interpretation cannot be decided from the results reported by Dupoux et al.'s (2011). The alternative account in terms of coarticulation-sensitive

phonemic categorization could apply only when the [C_1^i C_2] stimuli count as legitimate vowel-devoiced realizations of /C_1i C_2/ on the one hand, and C_1 is an appropriate consonant that carries enough /i/-coarticulation cues on the other. It has already been suggested in Chapter 2 that (i) /C_1/ should be voiceless in order for /i/ to be devoiced, and (ii) among voiceless consonants, /k/ should typically offer enough /i/-coarticulation cues. In other words, if the Japanese listeners' /i/-coarticulation sensitivity observed by Dupoux et al. is mostly due to /ki C_2/ stimuli, the phonemic categorization account would suffice to explain their results, in which case Japanese listeners' observed behavior should not be considered as instances of perceptual epenthesis; in contrast, if they exhibited /i/-coarticulation sensitivity also with other /C_1/'s (particularly voiced ones), such /C_1C_2/ stimuli must have constituted a phonotactic violation and hence Japanese listeners' observed coarticulation sensitivity should be seen as coarticulation-sensitive perceptual epenthesis, as Dupoux et al. claim.

Unfortunately, in Dupoux et al.'s (2011) 13 stimuli, /b, p, g, k, d/ were employed as /C_1/, and no comparison between voiced and voiceless /C_1/ (or among different places of articulation) is reported. Thus whether the phonemic categorization account would suffice to explain their observation with Japanese listeners is not known at this point.[62] Thus Dupoux et al.'s (2011) results with Japanese listeners do not decide the choice between two-step models and one-step models.

[62]Ogasawara and Warner (2009), reviewed above, compared /i/ perception based on (a) voiceless consonants followed by a devoiced /i/ on the one hand and (b) voiced consonants followed by a devoiced /i/ on the other, and observed that /i/ perception from (a) was easier than that from (b). Such a result accords with the alternative account in terms of coarticulation-sensitive phonemic categorization, rather than Dupoux et al.'s own account in terms of coarticulation-sensitive phonotactic repair.

However, the observed difference between (a)–(b) could be due to the amount or quality or /i/-coarticulation traces. As already noted, Ogasawara & Warner report the difficulty Ogasawara (as the speaker for their stimuli) in producing the (b) stimuli, which suggests that coarticulation cues were not natural with the (b) stimuli.

Furthermore, even if the relative difference was attributed to sensitivity to /i/-coarticulation within consonants, *less* sensitivity does not necessarily mean *no* sensitivity.

Thus Ogasawara & Warner's (2011) results do not decide the choice between the two accounts of Dupoux et al. (2011) results with Japanese listeners.

Conclusion from the subsection

Dehaene-Lambertz et al. (2000) and Dupoux et al. (2011) argued for one-step models and against two-step models, but their arguments were inconclusive. Dehaene-Lambertz et al.'s argument was inconclusive because, while the first-window EEG results could be taken as evidence for one-step models, the second- and the third-window results would rather conflict with natural expectations from one-step models, as far as their interpretations go; they do not provide an account of all the results from the three windows (which this thesis cannot provide, either). Dupoux et al.'s argument is inconclusive because, putting aside Portuguese results, which are rather hard to interpret without further details of Portuguese linguistic facts as well as experimental stimuli, their Japanese results do allow an alternative interpretation in terms of coarticulation sensitivity in the first-step phonemic categorization stage, rather than coarticulation sensitivity in the second-step phonotactic repair stage. Note that the suggested coarticulation sensitivity in the first-step phonemic categorization stage is the same as what I am calling **DB-sensitivity**. Thus, if the perceptual *un*reality of **DB-sensitivity** is confirmed, the suggested alternative interpretation of Dupoux et al.'s Japanese results would have to be rejected, an issue left open at this point.

3.5.4 Conclusion from the section

This section first pointed out that two versions of one-step models are possible, **suprasegmental matching** and **slot filling**, and then examined previous alleged evidence concerning the choice between one- and two-step models.

Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) saw Pallier et al.'s (1993) results as potential evidence against one-step models, and Dupoux et al. (2011) saw Kabak & Idsardi's (2007), Moreton's (2002) and Berent et al.'s (2007) results as such evidence; however, it was pointed out that they are all compatible with (both versions of) one-step models. However,

Matthews & Brown's (2004) results were argued to constitute evidence against the **suprasegmental matching** version, but not against the **slot filling** version of one-step models. Thus the previous evidence does not dictate our choice in favor of two-step models, because both two-step models and one version of one-step models (**slot filling**) are compatible with the experimental results.

On the other hand, Dehaene-Lambertz et al.'s (2000) and Dupoux et al.'s (2011) arguments for one-step models do not decide the choice between two-step models and one-step models; a cohesive interpretation of all of Dehaene-Lambertz et al.'s results is not yet available on the one hand, and **DB-sensitivity** would enable an alternative interpretation of Dupoux et al.'s Japanese results in favor of two-step models on the other.

3.6 Chapter Summary

Section 3.2 reviewed classical literature on phonotactic sensitivity; indeed phonotactics seem to affect perception. Section 3.2 reviewed arguments for and against lexical reductions of phonotactic sensitivity (Dupoux et al., 2001; Fais et al., 2005; Mazuka et al., 2011). Fais et al.'s and some of Dupoux et al.'s results superficially seem to argue against the sublexical reality of phonotactic sensitivity. However, it was suggested that, if **DB-sensitivity** is (either lexically or sublexically) real, those observations would not constitute instances of phonotactic repairs and hence Mazuka et al.'s and Dupoux et al.'s arguments for the sublexical reality of phonotactic sensitivity and against lexicalist models would be defended.

Section 3.3 reviewed evidence (mostly) concerned with the reality of **DB-sensitivity**. Beckman & Shoji's (1984), Tsuchida's (1994) and Ogasawara & Warner's (2009) results allow some room for an interpretation in terms of **DI-sensitivity**, rather than **DB-sensitivity**, but it was suggested that their results would best be interpreted by assuming that **DB-sensitivity** is real both sublexically and lexically. To the extent that it is real sublexically, this interpretation ar-

gues against lexicalist models, which claim that coarticulation sensitivity can be *completely* accounted for in terms of lexical sensitivity, and Cutler et al.'s (2009) claim of such accounts based on their own results as well as their interpretation of Kingston et al.'s (2011) results, was argued against.

Finally, after illustrating two possible versions of one-step models (**suprasegmental matching** and **slot filling**), alleged evidence for two-step models and attempted arguments for one-step models were examined; it was argued that the alleged evidence for two-step models (Berent et al., 2007; Kabak & Idsardi, 2007; Matthews & Brown, 2004; Moreton, 2002; Pallier et al., 1993) do not in fact constitute evidence for two-step models, at least if the **slot filling** version is considered as a possible implementation of one-step models, and also that the attempted arguments for one-step models (Dehaene-Lambertz et al., 2000; Dupoux et al., 2011) are not conclusive. The reason for Dupoux et al.'s (2011) argument to be inconclusive is that their Japanese results could well be interpreted in terms of **DB-sensitivity**, in which case the observed coarticulation sensitivity could be interpreted in favor of two-step models as a result of the first-step phonemic categorization stage, rather than phonotactic repair.

Thus the discussion in this chapter leads us to ask whether **DB-sensitivity** is indeed real, particularly as sublexical coarticulation sensitivity. If it turns out to be indeed real, that would argue for the irreducibility of phonotactic sensitivity to lexical sensitivity (by defending Dupoux et al.'s 2001 and Mazuka et al.'s 2011 arguments), as well as, of course, the irreducibility of coarticulation sensitivity to lexical sensitivity, in which case lexicalist models should be rejected and only two-step models and one-step models would remain as candidates. At the same time, however, its reality would also constitute evidence that Dupoux et al.'s (2011) Japanese results do not decide the choice between two-step models and one-step models, by enabling (but not forcing) an alternative interpretation of those results in favor of two-step models.

On the other hand, if the reality of **DB-sensitivity** is confirmed, the choice between the remaining candidates, two-step models and one-step models, would have to be made by exam-

ining the (un)reality of Japanese listeners' **DI-sensitivity**. If **DB-sensitivity** is real, Dupoux et al.'s (2011) Japanese results would fail to establish the superiority of one-step models over two-step models because **DB-sensitivity** could well be taken as coarticulation sensitivity in the first-step phonemic categorization stage within two-step models. Thus, in order to demonstrate coarticulation sensitivity in phonotactic repair, coarticulation sensitivity that could not be attributed to the first-step phonemic categorization stage should be examined, and **DI-sensitivity** is just such sensitivity.

Thus we are led to ask two questions:

- Is **DB-sensitivity** (sublexically) real? (If it is real, lexicalist models should be rejected.)
- Is **DI-coarticulation sensitivity** real? (If it is real, one-step models should be chosen over two-step models; if it is not real, two-step models should be chosen over one-step models.)

They are empirical questions, that could be answered only through experimental investigations. The next step is to convert them to specific questions amenable to experimental examinations.

Chapter 4

The Questions

4.1 Introduction

The goal of this thesis to distinguish among one-step models, two-step models and lexicalist models, and according to the discussion in the previous chapter, the answers to the following two questions should provide crucial pieces of information for that goal:

- Is **DB-sensitivity** (sublexically) real? (The answer will enable us to distinguish lexicalist vs. non-lexicalist models.)

- Is **DI-sensitivity** real? (The answer will enable us to distinguish one- and two-step models.)

This chapter formulates them in forms that can be examined experimentally.

Section 4.2 deals with the first experimental question, i.e., the (sublexical) reality of **DB-sensitivity**. Its (either sublexical or lexical) reality would validate the defense, suggested in Chapter 3, of Dupoux et al.'s (2001) and Mazuka et al.'s (2011) argument against a lexicalist reduction of phonotactic sensitivity, in the face of Fais et al.'s (2005) results and some of Dupoux et al.'s (2001) results. Of course, its *sublexical* reality would argue against a lexicalist reduction of coarticulation sensitivity. Thus its sublexical reality would argue against lexicalist

models, which claim that coarticulation and/or phonotactic sensitivity can be accounted for by lexical sensitivity.

In addition, we have the subsidiary question of what affects the relative ease of allegedly devoicing-based /…CV…/ perception from (what non-Japanese listeners would perceive as) […C…] as compared to /…CV…/ perception from [CV]. As seen in the previous chapter, even when the experimental task is not explicitly lexical, /CV/ perception from vowel-devoiced stimuli is sometimes observed to be harder (Cutler et al., 2009; Beckman & Shoji, 1984), sometimes easier (Tsuchida, 1994), and sometimes not easier or harder (Ogasawara & Warner, 2009), than /CV/ perception from non-vowel-devoiced stimuli. The difference of the amount or quality of coarticulation traces within the consonants before the devoiced vowels was suggested as a possible cause of such conflicting results, but it is only a conjecture. Although the (in)validity of that conjecture would not directly determine the choice between lexicalist models vs. sublexical models (one- and two-step models), its validity would make the discussion in the previous chapter more convincing; so this conjecture will also be examined when the reality of **DB-sensitivity** is examined.

Section 4.3 deals with the question of whether **DI-sensitivity** is real. If the reality of **DB-sensitivity** is supported, only one- and two-step models would be left as candidates. However, at the same time, the reality of **DB-sensitivity** would also cast doubt on the validity of Dupoux et al.'s (2011) argument, based on their results with Japanese listeners, for the superiority of one-step models over two-step models; their results would then become interpretable in terms of coarticulation sensitivity in the first-step phonemic categorization stage under two-step models. Thus a support for the reality of **DB-sensitivity** not only would exclude lexicalist models as a candidate but also would tell us that the choice between one- and two-step models is still open; the choice between one- and two-step models would require an examination of **DI-sensitivity**.

Section 4.4 summarizes the chapter, with rough illustrations of the plans for the experiments to be conducted, as well as of the predictions of the hypotheses to be compared.

In the remainder of this thesis, the following **production presuppositions** are assumed concerning coarticulation within /CV/ in Japanese (Vance, 1987; 2008):

- Both /i/ and /e/ are fronted, but /i/ is more fronted than /e/.
- Velars exhibit rich place variations and hence carry rich coarticulation cues.
- Bilabials do not exhibit rich place variations and hence carry few or weak coarticulation cues.

4.2 DB-Sensitivity

In their discrimination experiment, Monahan et al. (2009) employed stimuli of the form /eCVma/ (the V-set) on the one hand, and those obtained from the V-set by deleting the medial /V/ (the C-set) on the other.[1] Such stimuli constitutions can be conveniently borrowed; they enable us to examine the sublexical reality of **DB-sensitivity**. For one thing, with /C/ being a velar or bilabial stop, as in the experiments to be conducted, /eCVma/ does not constitute a real word in Japanese, whatever /V/ is. For another, the voiced consonant /m/ helps us accomplish our goal, for the following reason.

First note that the C-set stimuli are stimuli with 'coarticulation traces of vowels' of the form [eC^{V}ma]. They are unlikely to be perceived as phonotactically illicit /eCma/; rather, they would be perceived as /eCVma/.[2] The question is what exactly leads the listeners to /eCVma/ perception. Devoicing-based categorization of [C^{V}] as /CV/ on the one hand, and /V/ epenthesis due to phonotactic repair on the other, both could lead them to such perception. Thus mere observations of /eCVma/ perception would not tell us whether such perception was

[1] Monahan et al.'s (2009) goal was to examine whether /o/ epenthesis after coronal obstruents, observed in Japanese loanwords, is perceptual in its origin, just as /u/ epenthesis is perceptual as argued by Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999). The results of their discrimination experiment suggests a negative answer, but it is not directly relevant to this thesis.

[2] However, whether they are indeed perceived as /eCVma/ rather than /eCma/ will be examined in Experiments 9–10.

due to devoicing-based phonemic categorization or to phonotactic repair. One way to examine which is responsible is to focus on the identity of the perceived /V/.

When the coarticulation traces are those of a non-front vowel, devoicing-based phonemic categorization exhibiting **DB-sensitivity** leads us to expect /Cu/ perception. Since /Cu/ perception in such a case is also expected from phonotactically induced default /u/ epenthesis, the identity of the /V/ perceived from non-front coarticulation would not help examine which is responsible.

In the case of /Ci/ perception from /i/-coarticulation ([eC^{i}ma]), an obvious interpretation would be in terms of devoicing-based phonemic categorization of [eC^{i}ma] as /Ci/ (**DB-sensitivity**), but alternatively it could also be interpreted as a result of coarticulation-sensitive phonotactic repair as claimed by Dupoux et al. (2011) (**DI-sensitivity**); thus /Ci/ perception from /i/-coarticulation could be interpreted as a result of coarticulation-sensitive *phonemic categorization* or of coarticulation-sensitive *phonotactic repair* (and that is the primary reason why the previous results by Beckman & Shoji, 1984, Ogasawara, 2013, Ogasawara & Warner, 2009, and Tsuchida, 1994, fell short of establishing the reality of **DB-sensitivity**).

However, /CV/ phonemic categorization exhibiting **DB-sensitivity** on the one hand, and phonotactically induced epenthesis exhibiting **DI-sensitivity** on the other, predict different effects of /e/-coarticulation ([C^{e}]); when /e/-coarticulation is strong enough to be exploited, **DB-sensitivity** should bias listeners toward /Ci/ perception (the 'non-veridical' /i/ perception from /e/-coarticulation), because /i/ is the only vowel that could result from devoicing-based phonemic categorization of coarticulation cues of a front vowel such as /e/, whereas **DI-sensitivity** should bias listeners toward /Ce/ perception (the 'veridical' recovery of the original /e/).[3] Of

[3]Speaking more precisely, /e/ perception, rather than /i/ perception, is expected from **DI-sensitivity** because (a) if high/low coarticulation is not at work, sensitivity to fronting should lead to /i/ and /e/ perception, with /i/ and /e/ perception divided by chance, (b) however, it is rather unlikely that high/low coarticulation is totally absent, and there is no obvious reason for **DI-sensitivity** not to exploit high/low coarticulation while exploiting front/back coarticulation; hence non-high coarticulation resulting from /e/ could only bias perception toward /e/, and furthermore (c) /e/ is not as fronted as /i/, and the moderate amount of fronting would presumably function as a cue for /e/, rather than /i/, again biasing perception toward /e/.

Although this thesis is not meant to be a study of an acoustic investigation, just in case, see Figure 2.5 on page 25. If /ki/- and /ke/-bursts are regarded as similar, that would only account for non-/ku/-perception from those bursts;

course, given the above **production presuppositions**, according to which phonetic realizations of /e/ are not as fronted as realizations of /i/ (Vance, 1987; 2008), exploitation of /e/-coarticulation is more likely to fail than exploitation of /i/-coarticulation; if /e/-coarticulation cues failed to be exploited in phonemic categorization, [C^{e}] would be phonemically categorized as /C/, and exploitation failure of /e/-coarticulation cues in phonotactic repair would result in the default /u/ epenthesis after /C/. Thus some tendency for the default /u/ epenthesis is expected anyway; the question is which non-/u/ vowel would be perceived in the case of default overrides. Exploitation of /e/-coarticulation in phonemic categorization exhibiting **DB-sensitivity** leads us to expect that default-overrides should be by /i/, whereas exploitation of /e/-coarticulation in phonotactic repair exhibiting **DI-sensitivity** leads us to expect that default-overrides should be by /e/.

Thus, by seeing which vowels function as default-overrides for [C^{e}] stimuli, we can examine effects of devoicing-based phonemic categorization exhibiting **DB-sensitivity** on the one hand, and of phonotactically induced epenthesis exhibiting **DI-sensitivity** on the other; the former predicts that the percepts should be mostly /i/ or /u/, whereas the latter predicts that the percepts should be mostly /e/ or /u/. (Of course, phonotactically induced phonotactic repair *in*sensitive to coarticulation predicts that the percepts should be mostly /u/.) See Table 4.1 for a comparison of the predictions. (Table 4.1 itself is neutral with respect to whether predicted perception is sublexical or lexical.)

there is no obvious reason for **DI-sensitivity** to lead to the perception of /ki/, rather than /ke/, from those bursts, while **DB-sensitivity** leads us to expect that the categorization should be biased toward /ki/, rather than /ke/.

Table 4.1: Vowel perception for [C^V] as predicted by **(a)** devoicing-based phonemic categorization exhibiting **DB-sensitivity**, **(b)** phonotactically induced epenthesis exhibiting **DI-sensitivity**, and **(c)** coarticulation-*in*sensitive epenthesis (where C is voiceless).

	[C^u] (/u/-coarticulation)	[C^i] (/i/-coarticulation)	[C^e] (/e/-coarticulation)
(a) devoicing-based phonemic categorization exhibiting **DB-sensitivity**	/u/	/i/, /u/	/i/, /u/
(b) phonotactically induced epenthesis exhibiting **DI-sensitivity**	/u/	/i/, /u/	/e/, /u/
(c) coarticulation-*in*sensitive epenthesis	/u/	/u/	/u/

However, the predicted effect of **DB-sensitivity** should be observed with [eC^{V}ma] stimuli only if the sensitivity is *sublexical*. Note that, in such C-set stimuli, the 'devoicing' site is followed by a voiced consonant [m]; thus such stimuli correspond to Ogasawara & Warner's (2009) 'post-voiced stimuli'. In actual productions, devoicing of /V/ is not expected in such a context, so devoicing-based phonemic categorization of /e/-coarticulation as /i/ could only be interpreted as categorization of the sublexical [C^{e}] portion as /Ci/, rather than as categorization of the whole [eC^{e}ma] as /ekima/. Thus the confirmation of **DB-sensitivity** with C-set [eC^{e}ma] stimuli would not only confirm the mere reality of **DB-sensitivity** as opposed to **DI-sensitivity** but also its *sublexical* reality. This result would validate the interpretation suggested in the previous chapter of the discrepancy between the /i/ monitoring results vs. the lexical decision results by Ogasawara & Warner, an interpretation according to which 'post-voiced stimuli' function as legitimate 'devoiced environment' stimuli with a sublexical task (/i/ monitoring) because the [C^{i}] portion is sublexically categorized as /Ci/, whereas they function as 'voiced environment' stimuli with a lexical task (lexical decision) because post-voiced forms are atypical for attested sound patterns of the words with which the stimuli should be matched. Thus the C-set stimuli [eC^{V}ma] will enable us to examine the *sublexical* reality of **DB-sensitivity**.

Of course, if C-set [eC^{e}ma] stimuli resulted mostly in /ekema/ and /ekuma/ perception, rather than /ekima/ and /ekuma/, that could be interpreted as suggesting either that **DB-sensitivity** is not real in the first place, or that it is real but only lexically. In such a case, in order to tease apart those two interpretations, additional experiments should be conducted with such stimuli as [eC^{e}ta], in which the consonant after the 'devoicing' site is voiceless so that *lexical* **DB-sensitivity** could be exerted.

An additional advantage of [m] as the consonant after the 'devoicing' site is that it is not a fricative. As seen in Chapter 3, Dupoux et al. (2001) attributed /i/ perception against their *reksi* and *riksi* stimuli to voiceless stop-fricative sequences, an attribution argued against in Chapter 3. If a fricative should follow the 'devoicing' site in order for /i/ to be perceived,

[eC^{V}ma] should not lead to /i/ perception.

We also have the subsidiary question of what affects the effectiveness of **DB-sensitivity** (if it is real). It was suggested above that coarticulation-sensitive vowel perception from vowel-devoiced stimuli may or may not be harder than vowel perception from non-devoiced stimuli depending on the amount or quality of coarticulation traces in the devoiced stimuli. With the C- and the V-set stimuli, the relative ease of vowel perception from devoiced and non-devoiced stimuli translates to comparisons between vowel perception from C- and V-set stimuli. The suggestion can be tested in two different ways. First, as already noted, velars are expected to exhibit rich place variation and hence to provide rich coarticulation cues, whereas bilabials are not. Thus that suggestion leads us to expect that **DB-sensitivity** should be exerted well with the C-set /k/ stimuli ([ek^{V}ma]), but not so with the C-set /p/ stimuli ([ep^{V}ma]). Second, artificial shortening of the durations of the [k] and the [p] bursts would presumably destroy some of the spectral properties carrying coarticulation information, which would enable us to examine the effect of the amount or quality of coarticulation cues *within the same consonants* (rather than across different consonants); if the above suggestion is on the right track, artificial shortening of the durations of the bursts should make coarticulation-sensitive vowel categorization more difficult. This expectation can be examined by conducting identification experiments with full [k] and [p] stimuli on the one hand, and with shortened [k] and [p] stimuli on the other.

4.3 DI-Sensitivity

Note that, if **DB-sensitivity** turns out to be real, that does not necessarily deny the reality of **DI-sensitivity**; they could both be real, with their effects competing with each other (in the case of conflicts). For example, when /e/-coarticulation in [k^{e}] stimuli is successfully exploited, **DB-sensitivity** (if real) would dictate /i/ perception, while **DI-sensitivity** (if real) would dictate /e/ perception. An observation of a significant bias toward /i/ perception could be interpreted as

suggesting either (i) that **DB-sensitivity** is real but **DI-sensitivity** is not, or (ii) that both are real, but the effect of the former won over the effect of the latter. Whichever interpretation is adopted, the reality of **DB-sensitivity** would be implicated, but **DI-sensitivity** may or may not be real depending on which interpretation should be adopted. Thus, if a significant bias toward /i/ perception is observed for [k^e], supporting the prediction from the reality of **DB-sensitivity**, the (un)reality of **DI-sensitivity** would have to be examined in a setting in which **DB-sensitivity** is inapplicable in the first place.

As already noted, vowels (/V/) in /C_1VC_2/ do not (easily) devoice unless both /C_1/ and /C_2/ are voiceless. Devoicing-based V perception from [$C_1^VC_2$], where /C_1/ is voiceless but /C_2/ is voiced (as in [eC^Vma]), would be possible if **DB-sensitivity** is real sublexically so that the [C_1^V] portion would be phonemically categorized as /C_1V/, but that would not enable a similar phonemic categorization of [C_1^V] as /C_1V/ when /C_1/ itself is voiced; the alleged sublexical devoicing-based phonemic categorization is phonemic categorization of a *voiceless* [C_1^V] as /C_1V/, which should be inapplicable to a *voiced* [C_1^V]. Thus possible effects of **DB-sensitivity** would be blocked if C-set stimuli with voiced [C_1^V]'s (e.g., [eg^ima] or [eg^ita]) are employed as stimuli.[4]

In fact, the biggest reason that Dupoux, Kakehi, Hirose, Pallier, & Mehler's (1999) results, as well as most of the subsequent results, have been interpreted as instances of perceptual epenthesis unrelated to vowel devoicing, is that most of the stimuli involved voiced consonants, which would block vowel devoicing (or deletion with coarticulation traces being left). For example, as noted above, in Maekawa & Kikuchi's (2005) corpus-based survey, the rates of /i, u/ devoicing in /-C_1VC_2-/ are about 89 % if both /C_1/ and /C_2/ are voiceless (voiceless-voiceless contexts), from 17 to 20 % if /C_1/ is voiceless and /C_2/ is voiced (voiceless-voiced contexts),

[4]Recall Maekawa & Kikuchi's (2005) corpus results, according to which the devoicing rate is about 1 % if /C_1/ is voiced whether or not /C_2/ is voiceless or voiced, in contrast to the devoicing rates when /C_1/ is voiceless (see immediately below); thus, when /C_1/ is voiced, the voicing of /C_2/ is very unlikely to affect the applicability of **DB-sensitivity** and hence both [eg^Vma] and [eg^Vta] stimuli are expected to be exempt from possible effects of **DB-sensitivity**.

and around 1 % if /C_1/ is voiced (whether /C_2/ is voiceless or voiced; pre-voiced contexts). However, in Dupoux et al.'s (2011) experiments, voiceless-voiceless context stimuli and pre-voiced context stimuli are both employed, as seen in the previous chapter. Because Dupoux et al. conflated both kinds of stimuli, it is not clear whether Japanese listeners' observed coarticulation sensitivity should be interpreted as mostly being due to voiceless-voiceless stimuli and hence as a result of phonemic categorization exhibiting **DB-sensitivity** or as being also due to pre-voiced stimuli and hence as a result of phonotactic repair exhibiting **DI-sensitivity**.

Thus, by examining whether Japanese listeners' vowel perception from pre-voiced stimuli exhibits coarticulation sensitivity, we can hope to determine whether Dupoux et al.'s (2011) results should be seen as reflecting coarticulation-sensitive phonotactic repair. Coarticulation sensitivity with pre-voiced stimuli of the form [e$C_1^V C_2$], where [C_1] is voiced, could only be conceived of as **DI-sensitivity**. If **DI-sensitivity** is observed with such stimuli, that could not be interpreted in terms of coarticulation-sensitive phonemic categorization and hence would argue for the coarticulation sensitivity of phonotactic repair, in accord with one-step models but not with two-step models. In contrast, if **DI-sensitivity** fails to be observed with such stimuli, the coarticulation sensitivity as observed by Dupoux et al. (2011) could be attributed to coarticulation-sensitive *phonemic categorization*, rather than coarticulation-sensitive *phonotactic repair*, which is compatible with two-step models. In short, an examination of **DI-sensitivity** with such pre-voiced stimuli will enable an evaluation of the claimed superiority (Dupoux, et al., 2011) of one-step models over two-step models.

In contrast to **DB-sensitivity**, **DI-sensitivity** itself dictates no 'non-veridical' bias and is the most compatible with 'veridical' restorations of the underlying vowels; thus, while **DB-sensitivity** leads us to expect that successful exploitation of /i/- and /e/-coarticulation within a voiceless consonant should both results in a bias toward /i/-perception, **DI-sensitivity** (with no additional 'non-veridical' bias) leads us to expect that successful exploitation of /i/- and /e/-coarticulation within a voiced consonant should result in biases toward /i/- and /e/-perception

respectively. Note that, in the case of [eC^{V}ma] stimuli where [C] is voiced, an observation of the tendency for /i/ perception from /i/-coarticulation would suffice for our purpose; such a tendency could only be interpreted as an effect of **DI-sensitivity**.

Speaking more precisely, **DI-sensitivity** both in voiced cases and in voiceless cases would suggest coarticulation-sensitive *phonotactic repair*, as opposed to coarticulation-sensitive *phonemic categorization*. Thus the validity of Dupoux et al.'s (2011) claim will be examined doubly; if the results with voiceless cases turn out to support the reality of **DI-sensitivity**, as opposed to **DB-sensitivity**, that would constitute evidence for coarticulation-sensitive *phonotactic repair*. Of course, as already noted, if the results with voiceless cases turn out to support the reality of **DB-sensitivity**, but not the reality of **DI-sensitivity**, that would not necessarily tell us that **DI-sensitivity** is unreal; both kinds of sensitivity could be real, with the effects of the former winning those of the latter when they are in conflict. Thus, in such a case, examinations with voiced stimuli become crucial.

4.4 Summary, Research Plan, and Predictions

In order to distinguish among lexicalist vs. non-lexicalist models, the sublexical reality of **DB-sensitivity** should be examined. When C is voiceless, /Ci/ perception from /i/-coarticulation could be interpreted either as a result of **DB-sensitivity** or as a result of **DI-sensitivity**, but /Ci/ perception from /e/-coarticulation could only be interpreted as **DB-sensitivity**. Thus (a) **DB-sensitivity**, (b) **DI-sensitivity** and (c) coarticulation insensitivity make different effects of /i/- or /e/-coarticulation traces, as illustrated in the two 'voiceless [C^{V}] columns in Figure 4.2. They will be examined in four identification experiments (Experiments 1–2 and Experiments 9–10) and in two discrimination experiments (Experiments 3–4), where the crucial stimuli are of the form [eC^{V}ma]; given that vowels do not devoice before a voiced consonant (in this case, [m]), /Ci/ perception from such stimuli could only be interpreted as a result of *sublexical* **DB-**

sensitivity.

The effects of **DB-sensitivity**, if real, should depend on the strength of coarticulation traces. This prediction will be examined (i) through a comparison between [ek^{V}ma] and [ep^{V}ma] stimuli (Experiments 1–4), and (ii) through a comparison between full burst stimuli (Experiment 1) and shortened burst stimuli (Experiment 2); the effects of **DB-sensitivity** should be observed with velar stimuli rather than with bilabial stimuli on the one hand, and should be observed more with full burst stimuli than with shortened burst stimuli.

On the other hand, in order to distinguish among two non-lexicalist models (i.e., one- and two-step models), the reality of **DI-sensitivity** should be examined. When C is voiced, Japanese listeners' /Ci/ perception from /i/-coarticulation could only be interpreted as a result of **DI-sensitivity**; in such a case, **DB-sensitivity** should be inapplicable. Thus the perceptual reality of **DI-sensitivity** leads us to expect Japanese listeners' tendency to perceive /Ci/, whereas its unreality leads us to expect the 'default epenthesis' /Cu/, as illustrated in the rightmost 'voiced [C]' columns in Figure 4.2. This prediction will be examined in four identification experiments (Experiments 5 and 8–10) and two discrimination experiments (Experiments 6–7).

Experiments 1–4 employing voiceless consonant stimuli are reported in Chapter 5, while Experiments 5–8 employing (additional) voiced consonant stimuli are reported in Chapter 6; Experiments 9–10, which are supplementary identification experiments, are reported in Chapter 7.

Table 4.2: Vowel perception for [C^V] as predicted by **(a)** devoicing-based phonemic categorization exhibiting **DB-sensitivity**, **(b)** phonotactically induced epenthesis exhibiting **DI-sensitivity**, and **(c)** coarticulation-*in*sensitive epenthesis.

		voiceless [C^i] (/i/-coarticulation)	voiceless [C^e] (/e/-coarticulation)	voiced [C^i] (/i/-coarticulation)
(a)	devoicing-based phonemic categorization exhibiting **DB-sensitivity**	/i/, /u/	/i/, /u/	(N/A)
(b)	phonotactically induced epenthesis exhibiting **DI-sensitivity**	/i/, /u/	/e/, /u/	/i/, /u/
(c)	coarticulation-*in*sensitive epenthesis	/u/	/u/	/u/

Chapter 5

Experiments 1–4

This chapter reports Experiments 1–4, which aimed at confirming the sublexical reality of **DB-sensitivity**. To repeat, the stimuli are natural utterances of /eCVma/ with a removal of three pitch periods of the medial /V/ (the V-set) on the one hand, and those derived from the /eCVma/ utterances by deleting the medial vowel (the voiced portion between the end of the /C/ burst and the beginning of the /m/ constriction; the C-set) on the other; the specific hypotheses to be compared are:

When the perception of a vowel /V/ is induced by $[C_1^V C_2]$,

coarticulation insensitivity: Subphonemic details of C_1 due to coarticulation traces of V have no effect and hence /V/ should be uniformly the default /u/ (when $[C_1]$ is either a velar or bilabial stop).

DB-sensitivity: $[C_1]$ with front coarticulation traces (either due to /i/ or /e/) are phonemically categorized as /C_1i/ and hence /V/ should be /i/.

DI-sensitivity: Coarticulation traces left by a specific vowel within $[C_1]$ enable the listeners to perceptually recover the vowel.

In the stimuli, V ranges over the five Japanese vowels /a, e, i, o, u/. However, the analyses employ only those results from the stimuli in which V (either elided or not) is either /e/, /i/ or /u/; those involving /a/ or /o/ are treated as fillers and not employed in the analyses. This decision is based on the following considerations.

For one thing, while there is an agreement in the literature that /i, u/ tend to devoice but /e/ does not, it is not empirically clear whether /a, o/ have a tendency for devoicing; as already noted in Chapter 2, Nihon Hoosoo Kyookai (1998:227) claims that /a, o/ in /ka, ko/ morae tend to devoice when the /ka, ko/ morae are word-initial low-pitched morae immediately followed by high-pitched morae. This is simply a claim, not accompanied by any empirical justification, and unfortunately, Maekawa & Kikuchi (2005) did not report the devoicing rates of /a, o/ in such a phonological environment. However, if it is not clear whether /a, o/ tend to devoice in productions, the empirical predictions of the **DB-sensitivity** hypothesis is not clear for /a, o/. For another, the minimum **production presuppositions** are the only assumptions made concerning the strength of coarticulation cues, which means that we do not know about the strength of coarticulation cues left by /a/ or /o/. That in turn means that the predictions from the **DI-sensitivity** hypothesis are not clear either with respect to stimuli involving /a/ or /o/. Finally, we are specifically interested in whether Japanese listeners will exhibit coarticulation sensitivity to front vowels, and if they do, whether /e/-coarticulation is perceived as 'front coarticulation' to induce /i/ perception (**DB-sensitivity**) or as '/e/-coarticulation' to induce /e/ perception (**DI-sensitivity**). It is the results with /i, e/ stimuli that would give an answer to such questions. In contrast, whether listeners are sensitive to coarticulation or not, results with /u/ stimuli would function as a baseline, because the two kinds of coarticulation sensitivity on the one hand, and coarticulation insensitivity on the other, are expected to induce /u/ perception. Thus results with stimuli involving /e, i, u/ would suffice for our purpose.

Experiments 1–2 are identification experiments. Recall the **production presuppositions**, according to which coarticulation sensitivity would be expected with velars but probably not

with bilabials. In the case of the C-set stimuli of the form [eC^{V}ma], the **coarticulation insensitivity** hypothesis predicts uniform /u/ identification between /C/ and /m/; the **DB-sensitivity** hypothesis (coupled with its assumed sublexical reality of the sensitivity) leads us to expect the possibility of default-overriding /i/ identifications not only for [eC^{i}ma] but also for [eC^{e}ma] stimuli; the **DI-sensitivity** hypothesis predicts the possibility of default-overriding /i/ identification for [eC^{i}ma] stimuli on the one hand, and of default-overriding /e/ identification for [eC^{e}ma] stimuli on the other. Given the default status of /u/ epenthesis, then, the three hypotheses can be teased apart by seeing whether the tendency for /i/ identification differs among C-set stimuli on the one hand, and if it differs, by seeing the specific effect of the coarticulation cues (particularly within [eC^{e}ma] stimuli) on the other.

At the same time, recall that the conflicting observations in the literature (Beckman & Shoji, 1984; Cutler et al., 2009; Ogasawara & Warner, 2009; Tsuchida, 1994) were noted in Chapter 3 with respect to whether /i/ perception is easier or harder from [C$\underset{\circ}{\text{i}}$] than from [Ci], for which the amount or strength of coarticulation cues was suggested as a possible cause. That suggestion leads us to expect, under the **production presuppositions**, coarticulation sensitivity (either devoicing-based or devoicing-independent) with C-set velar stimuli ([ek^{V}ma]) but not with C-set bilabial stimuli ([ep^{V}ma]). That means that the default-overriding /i/ or /e/ perception expected from coarticulation sensitivity should be observed with C-set velar stimuli but probably not with C-set bilabial stimuli. Furthermore, the conflicting observations in the literature concern the relative ease of /Ci/ perception from vowel-devoiced (C-set) and non-devoiced (V-set) stimuli; the above suggestion leads us to expect the possibility of comparable success in identifying /ki/ in C-set [ek^{i}ma] stimuli and in their corresponding V-set stimuli on the one hand, and the definitely more difficult /pi/ identification in C-set [ep^{i}ma] stimuli than in their corresponding V-set stimuli on the other, given the **production presuppositions** dictating that coarticulation cues within bilabials are poor.

In addition to this across-consonant comparison, Experiments 1–2 also attempt a within-consonant comparison, by manipulating the durations of [k] or [p] bursts. Experiment 2 mimics Experiment 1 except that the durations of [k] and [p] bursts are shortened. A comparison between the results of Experiment 1 and of Experiment 2 will constitute a within-consonant comparison. The above suggestion concerning the relative ease with which /CV/ is perceived from vowel-devoiced or non-devoiced stimuli predicts that whatever ease with /CV/ perception from vowel-devoiced (i.e., C-set) stimuli as compared to /CV/ perception from non-devoiced (i.e., V-set) stimuli should be diminished with burst shortening. That is, the potentially negative effect of vowel-devoicing on /CV/ perception should be more severe in Experiment 2 than in Experiment 1. A confirmation of this expectation would support the suggested account of the discrepancy among Beckman & Shoji's (1984), Cutler et al.'s (2009), Ogasawara & Warner's (2009) and Tsuchida's (1994) results concerning the effect of vowel devoicing on /CV/ perception.

Experiments 3–4 are AX discrimination experiments, in which listeners are presented pairs of C-set stimuli and instructed to discriminate between the medial vowels. The discrimination performance would depend on the percepts of both members of the stimulus pairs. The predictions of the three hypotheses for each of the two AX discrimination experiments, as well as for a comparison between the two experiments, are somewhat complicated and will be described in a separate section (Section 5.3).

5.1 Experiment 1

Experiment 1 examined Japanese listeners' ability to identify the missing vowel based on velar and bilabial bursts excised from /CV/, where V is one of the Japanese vowels /e, i, u/, and the /CV/ is immediately followed by /m/ (a voiced consonant).

As stated above, the three hypotheses in question make different predictions for /C/ bursts

Table 5.1: Vowel identifications predicted by the three hypotheses

	/i/-coarticulation	/e/-coarticulation	/u/-coarticulation
coarticulation insensitivity	/u/	/u/	/u/
DB-sensitivity	/i/	/i/	/u/
DI-sensitivity	/i/	/e/	/u/

excised from /CV/, as illustrated in Table 5.1. (For space reasons, the /k/–/p/ differences predicted are omitted.)

5.1.1 Method

Stimuli

Three native speakers of Japanese (two females and one male) produced three-mora nonwords of the form /eCVma/, where C is either /p/ or /k/, V is one of the Japanese five vowels /a, i, u, e, o/, and an accent is placed on the first mora. Their utterances were recorded in a sound-attenuated room with Marantz Solid State Recorder PMD660 with a sampling frequency of 44,100 Hz. For each /eCVma/, at least two utterances were recorded, and among them, those utterances whose accent placement sounded most precise to the experimenter's ears were selected. Two sets of stimuli were produced from the original utterances: **the C-set** stimuli were created by deleting the V's (the voiced portions between the end of the /p/ or /k/ bursts and the onsets of the /m/ constrictions), leaving the bursts of the C's, and **the V-set** stimuli were created by deleting three pitch periods (beginning in a zero-crossing and ending in another zero-crossing) arbitrarily chosen from the midst of the [V] portions, such as the shaded portion in Figure 5.1, so that both the C-set and the V-set stimuli are edited versions of the original recordings.[1] The intensity (RMS intensity averaged over the whole token) of each of the stimuli was rescaled to the same level before presentation. Three repetitions of the C- and the V-set stimuli were intermixed in a computer-generated random order and presented to the listen-

[1] Again, this is an imitation of Monahan et al. (2009).

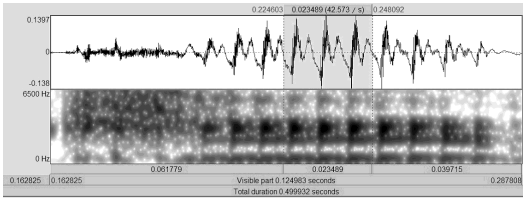


Figure 5.1: An illustration of the deletion of three pitch periods in creating V-set stimuli; the figure as a whole is the /ki/ portion in the original recording of the male speaker's utterance of /ekima/, and the shaded portion constitutes three pitch periods to be deleted.

ers through circumaural closed-back headphones (SONY MDR-ZX700) in a sound-attenuated room. Example waveforms and spectrograms are shown in Figure 5.2.

Participants

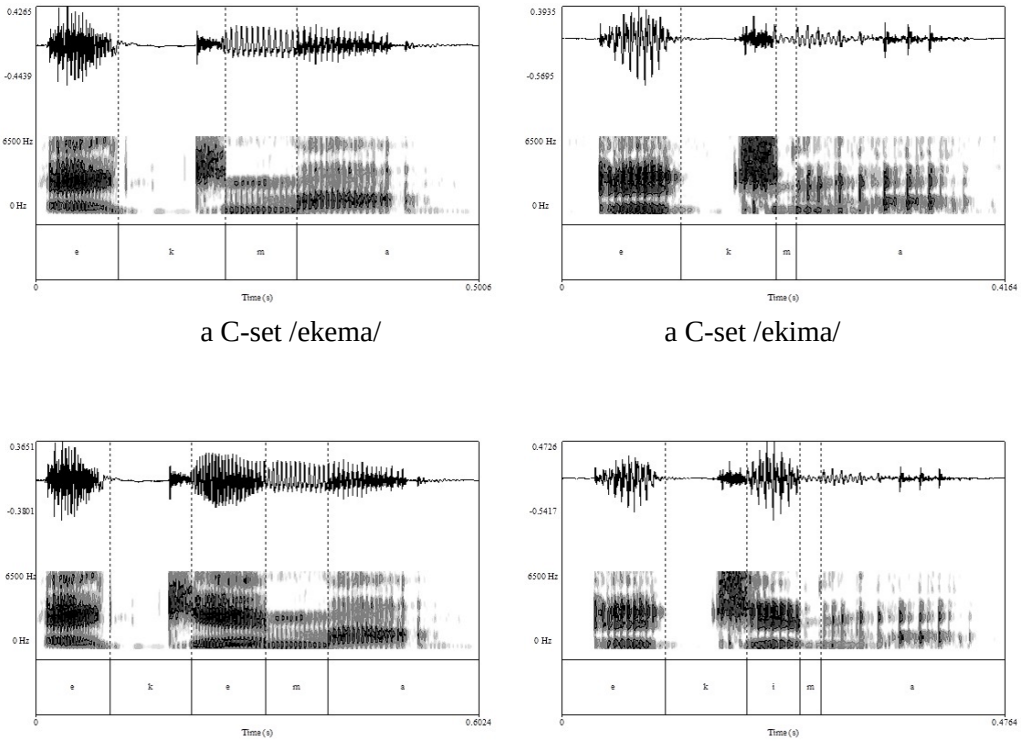
Nine native Japanese listeners with no known hearing disability participated in the experiment for course credit at undergraduate programs at Hosei University, Tokyo (seven males and two females; mean age $= 19$, S.D. $= .27$).

Procedure

The experiment was controlled by E-prime (Psychology Software Tools, Inc.) run on a Windows XP machine (Panasonic Let's note CF-S9KYKBDU). To examine vowel identification by the listeners, the notion of "dan" was conveniently employed. Japanese has convenient terms for morae with specific vowels, "V-dan", where "V" is replaced by either one of the five vowels /a, i, u, e, o/, which, together with its order /a, i, u, e, o/, everybody is familiar with since the days spent at compulsory elementary school.

Listeners were told that the stimuli were of the form /eCVma/, and they were asked to identify what "dan" the medial /CV/ was by pressing one of the response keys on a response box, where the response keys were assigned to the "dan"s respecting the order they are familiar with. Thus, in effect, the task was vowel identification.

Listeners were first required to familiarize themselves with the task by going through a



a C-set /ekema/ a C-set /ekima/

a V-set /ekema/ a V-set /ekima/

Figure 5.2: Example waveforms and spectrograms of the stimuli for Experiment 1; the /ekema/ stimuli are by one of the female speakers and the /ekima/ stimuli are by the male speaker.

practice session (with /ebVma/ and /egVma/ stimuli, where V ranges over the five Japanese vowels, spoken by the experimenter, with no significant editing), and only after scoring more than 80 % accuracy were they allowed to proceed to the main experimental session. As stated above, each stimulus was presented three times, intermixed with each other in a random order generated by E-prime.

Statistical Analyses

As noted above, results with medial /a/ and /o/ are treated as fillers; only those with medial /e, i, u/ are to be employed in the analyses.

The primary dependent variable is the listener-averaged /i/ identification rates, with identification rates for the other vowels being used to supplement the interpretations when necessary. When the responses are conceived of as being classified into /i/ and non-/i/, the results will constitute binomial distributions, which are, strictly speaking, not appropriate for parametric tests such as ANOVA or *t* tests, which assume underlying normal distributions. The usual technique to cope with such a situation is to arcsin square root transform the obtained rates to approximate normal distributions (McNicol, 1972). Employing such a technique, the transformed identification rates were submitted first to two-sided one-sample *t* tests, with the expected value being the transform of the chance level for a five-choice task ($= .2$). Since the responses to V-set stimuli on their own are uninformative with respect to which hypothesis should be adopted, only the C-set results were analyzed with such one-sample *t* tests; yet, six tests (for C-set /ekema/, /ekima/, /ekuma/, /epema/, /epima/, and /epuma/) were conducted, so the significance values are corrected with the Holm method (also known as the Sequentially Rejective Bonferroni method). Results of these one-sample *t* tests will tell us which stimuli do or do not tend to induce the default-overriding /i/ responses so that the three hypotheses in question will be teased apart.

Furthermore, (the arcsin square root transforms of) the listener-averaged /i/ identification

Table 5.2: The numbers of responses for the V-set /k/ stimuli in Experiment 1.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	79	2	0	0	81
	/i/	0	1	80	0	0	81
	/u/	3	0	1	0	77	81
Total		3	80	83	0	77	243

Table 5.3: The numbers of responses for the V-set /p/ stimuli in Experiment 1.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	79	0	0	2	81
	/i/	0	0	81	0	0	81
	/u/	0	1	0	0	80	81
Total		0	80	81	0	82	243

rates with the C- and the V-set /ki/ and /pi/ stimuli were submitted to a two-way between-subject ANOVA, with 'consonant' (/k/ vs. /p/) and the 'set' (the C- vs. the V-set) as independent factors, in order to examine whether /Ci/ perception is easier or harder from 'devoiced' stimuli (C-set) than from 'non-devoiced' stimuli (V-set), and whether the difference across the C- and the V-set depends on the consonant (with a different amount of coarticulation cues, according to the **production presuppositions**).

5.1.2 Results and Discussion

The results were divided according to the following three variables:

CV: whether the stimulus was from the C- or the V-set.

consonant: whether the medial consonant was /k/ or /p/.

vowel: the identity of the medial vowel (/i/, /e/, /u/).

Naturally, listeners exhibited good identifications of the original vowels with the V-set stimuli irrespective of the consonant or the vowel, as shown in Table 5.2 and Table 5.3.

Table 5.4: The numbers of responses for the C-set /k/ stimuli in Experiment 1.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	1	58	1	21	81
	/i/	0	0	80	0	1	81
	/u/	0	1	1	0	79	81
Total		0	2	139	1	101	243

Table 5.5: The numbers of responses for the C-set /p/ stimuli in Experiment 1.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	4	8	0	69	81
	/i/	1	2	18	0	60	81
	/u/	0	1	0	0	80	81
Total		1	7	26	0	209	243

In contrast, in the case of the C-set stimuli, listeners' identification of the original vowel was poor with /ke/, /pe/ and /pi/, as shown in Table 5.4, and Table 5.5,

The crucial measures for an examination of the prediction of **DB-sensitivity** are the observed /i/ response rates, which are as described in Table 5.6, Table 5.7, Figure 5.3, and Figure 5.4. Recall that, under the **production presuppositions**, it leads us to expect a high rate of /i/ identification with /ki/ and (to a lesser extent) /ke/ stimuli, in contrast to the other stimuli, a prediction that seems to be born out.

The listener-averaged /i/ identification rates with the C-set /ke/, /ki/, /ku/, /pe/, /pi/, and /pu/ were arcsin square root transformed and submitted to two-sided one-sample t tests with the expected value of the arcsin square root transform of .5. Naturally, the /i/ identification rate with the C-set /ku/ stimuli was significantly below chance [$t(8) = -11.343,\ p < .001$]; the /i/ identification rate for the C-set /pu/ reached the floor (i.e., no /i/ identification). The /i/ identification rate with the C-set /pe/ stimuli was also significantly below chance [$t(8) = -2.837,\ p = .044$; $p = .022$ if uncorrected], but the /i/ identification rate with the C-set /pi/ stimuli did not differ from chance [$t(8) = -.526,\ p = .613$ uncorrected]. Thus none of the

Table 5.6: /i/ response rates within the C-set in Experiment 1, with the vowel factor (horizontal) and the consonant factor (vertical); within each cell, the means are on the top line in the bold, the standard deviations and *N*'s are on the second and the third lines respectively.

	/e/	/i/	/u/	total
/k/	**.72** .46 81	**.99** .11 81	**.01** .11 81	**.57** .50 243
/p/	**.10** .30 81	**.22** .42 81	**.00** .00 81	**.07** .25 243
total	**.41** .49 162	**.60** .49 162	**.01** .08 162	**.34** .47 486

Table 5.7: /i/ response rates within the V-set in Experiment 1, with the vowel factor (horizontal) and the consonant factor (vertical); within each cell, the means are on the top line in the bold, the standard deviations and *N*'s are on the second and the third lines respectively.

	/e/	/i/	/u/	total
/k/	**.25** .16 81	**.99** .11 81	**.01** .11 81	**.34** .48 243
/p/	**.00** .00 81	**1.00** .00 81	**.00** .00 81	**.33** .47 243
total	**.01** .11 162	**.99** .08 162	**.01** .08 162	**.34** .47 486

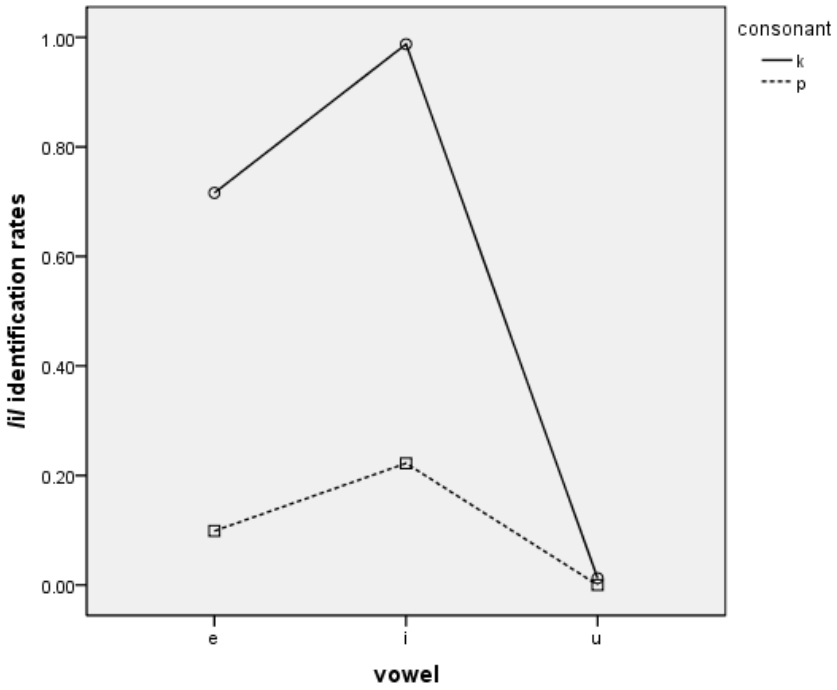


Figure 5.3: /i/ response rates within the C-set in Experiment 1.

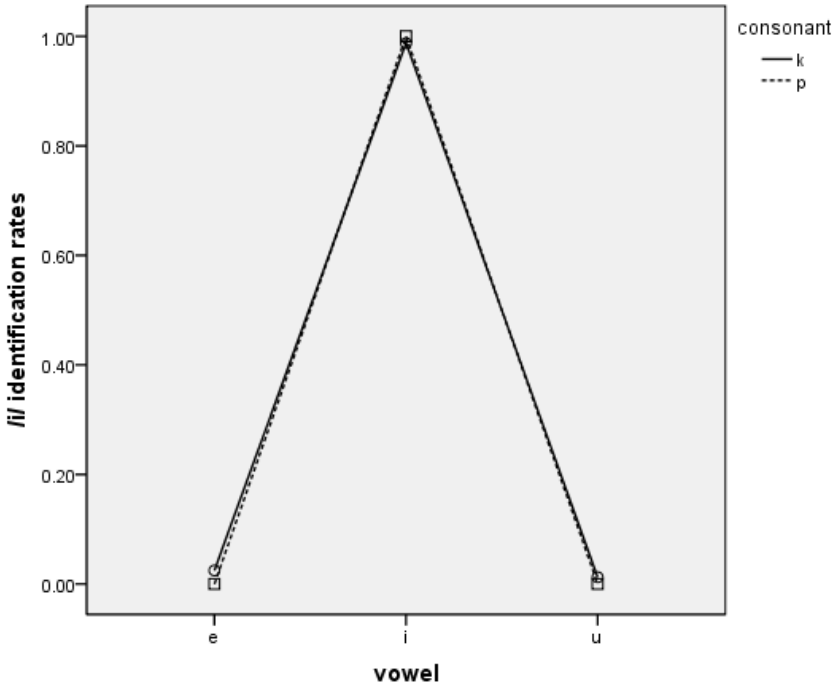


Figure 5.4: /i/ response rates within the V-set in Experiment 1.

C-set /p/ stimuli exhibited a significant tendency to induce /i/ identification, which is naturally expected if coarticulation cues within bilabials are too poor to be successfully exploited by coarticulation sensitivity.

Our primary interests are in the results with the C-set /ki/ and /ke/ stimuli. Confirming the common prediction by the **DB-sensitivity** and the **DI-sensitivity** hypothesis, the /i/ response rate for the C-set /ki/ stimuli was significantly higher than chance [$t(8) = 28.475,\ p < .001$]. Furthermore, the /i/ identification result with the C-set /ke/ stimuli seem to support the prediction of **DB-sensitivity** rather than of **DI-sensitivity**; it was highly significantly higher than chance [$t(8) = 5.232,\ p = .002;\ p = .001$ if uncorrected]. To further confirm this conclusion, the rate of /e/ identification, expected from **DI-sensitivity**, was also examined; it was significantly below chance [$t(8) = -2.7579,\ p = .025$, uncorrected]. Thus the results are best interpreted in terms of **DB-sensitivity**.

As a supplementary analysis, listener-averaged transforms of /i/ response rates against the C- and the V-set /ki/ and /pi/ stimuli were submitted to a two-way between-subject ANOVA, with 'consonant' (/k/ vs. /p/) and 'set' (the C- vs. the V-set) as independent factors. The main effect of 'consonant' was highly significant [$F(1,\ 8) = 58.460,\ p < .001$], suggesting /k/ stimuli elicited more /i/ identifications than /p/ stimuli; the main effect of 'set' was also highly significant [$F(1,\ 8) = 47.762,\ p < .001$], suggesting that V-set stimuli elicited more /i/ identifications than C-set stimuli. Their interaction was also highly significant [$F(1,8) = 162.639, p < .001$], and examinations of simple main effects revealed that the /i/ identification rate was highly significantly higher against /k/ stimuli than against /p/ stimuli in the C-set [$F(1,\ 8) = 162.639,\ p < .001$] but not in the V-set [$F(1,8) = 1.000, p = .347$], while the /i/ identification rate was significantly lower against C-set stimuli than against V-set stimuli in the case of /p/ stimuli [$F(1,\ 8) = 89.017,\ p < .001$] but not at all in the case of /k/ stimuli [$F(1,\ 8) = .000,\ p = 1.000$]. In short, the deletion of /i/ (the voiced portion between the end of the burst and the onset of the following nasal closure) in creating the C-set stimuli had a

negative effect on /i/ identification in the case of /pi/, but not in the case of /ki/; thus the /ki/ results replicated Ogasawara & Warner's (2009) /i/ monitoring results, while the /pi/ results replicated Beckman & Shoji's (1984) (and Cutler et al.'s, 2009) results. This is just as expected from the assumption that rich coarticulation cues within velars lead to successful coarticulation-sensitive phonemic categorization while poor coarticulation cues within bilabials do not.

Thus the results of Experiment 1 support the sublexical reality of **DB-sensitivity** hypothesis; the tendency to perceive [ek^{e}ma] as /ekima/, rather than /ekema/ or /ekuma/, should not have been observed unless the [k^{e}] portion is phonemically categorized as /ki/, even though /ekima/ as a potential word would not be produced as [ek^{e}ma] due to the voiced /m/.

Furthermore, the comparison between the C- and the V-set /ki, pi/ stimuli suggests that, with rich enough coarticulation cues (as are found in /k/ bursts), /i/ perception is no less difficult from voiceless consonants than from voiceless consonants followed by [i], replicating Ogasawara & Warner's (2009) /i/ monitoring results.

5.2 Experiment 2

It was observed in Experiment 1 that /i/ identification was no less difficult with C-set /ki/ stimuli than with V-set /ki/ stimuli on the one hand, and that /i/ identification was indeed more difficult with C-set /pi/ stimuli than with V-set /pi/ stimuli. The /k/–/p/ difference was interpreted in terms of rich coarticulation cues within /k/ bursts vs. poor coarticulation cues within /p/ bursts. Such an interpretation suggests that devoicing did not result in more difficulty in /i/ perception in Ogasawara & Warner's (2009) /i/ monitoring experiment because their stimuli may have been coarticulation-rich, whereas devoicing did result in more difficult /i/ perception in Beckman & Shoji's (1984) experiments because their stimuli may have been rather coarticulation-poor; similarly, /CV/ perception was more difficult with 'devoiced' stimuli than with 'non-devoiced' stimuli in Cutler et al.'s (2009) experiments because coarticulation cues

within their their consonantal stimuli may have been rather poor so that the consonantal stimuli tended to be categorized as /C/, rather than /CV/.

However, the /ki/–/pi/ difference could be interpreted in terms of the place difference per se, rather than the different amounts or strength of coarticulation cues. Experiment 2 will help us further evaluate the above account for the discrepancy among Ogasawara & Warner's (2009), Beckman & Shoji's (1984) and Cutler et al.'s (2009) results.

In Experiment 2, the stimuli within the C-set were created anew so that the durations of the burst portions were equated across the /k/ and the /p/ condition. It has long been noted that velar bursts are far longer than bilabial bursts in natural productions (Fant, 1969; Winitz et al., 1972), and indeed, the durations of the /k/ bursts in the stimuli in Experiment 1 [$M = 26.78$ ms, SD $= 10.19$] were significantly longer than the /p/ bursts [$M = 11.04$ ms, SD $= 5.94$; $t(23) = 5.17,\ p < .001$ (two-tailed)]. Thus, durational normalization across velar and bilabial bursts would mean considerable shortening for velar bursts, and such shortening is likely to result in some spectral modifications, which in turn is likely to result in the degradation of coarticulation cues. Presumably, the magnitude of the degradation will be larger for /k/ bursts than for /p/ bursts, because the durational shortening is larger for /k/ bursts than for /p/ bursts. However, the original strength of coarticulation cues within /k/ bursts is assumed to be larger than those within /p/ bursts, and hence it is not clear whether the amounts of coarticulation cues within the shortened /k/ bursts are larger or smaller than those within the shortened /p/; thus a comparison across /k/ and /p/ stimuli would not be informative with respect to the validity of the above account. Rather, comparisons between the /ki/ results of Experiment 2 with those of Experiment 1 will be informative. In Experiment 1, /i/ identification rates were comparable for the C-set /ki/ stimuli and for the V-set /ki/ stimuli; this translates to 'no negative effect of devoicing on /i/ perception', replicating Ogasawara & Warner's (2009) /i/ monitoring results. If the amount or quality of coarticulation cues, rather than the particular place of articulation, is responsible for the ease with which /i/ was identified from C-set /ki/ stimuli

in Experiment 1, the degradation of spectral properties carrying coarticulation information in Experiment 2 should make such perception harder, resulting in a lower /i/ identification rates with C-set /ki/ stimuli than with V-set /ki/ stimuli, as is expected from the results by Beckman & Shoji (1984), as well as from Cutler et al. (2009).

5.2.1 Method

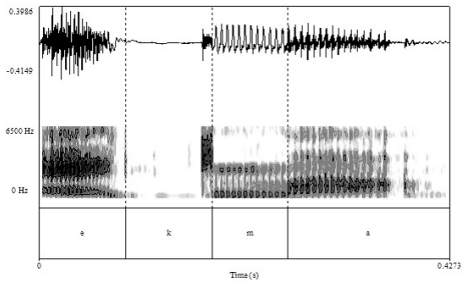
Stimuli

The original utterances employed in Experiment 1 were re-used. If the durations of the bursts were shortened to 5 ms, most of the C-set stimuli sounded like two-mora /ema/, rather than three-mora /eCVma/, to the experimenter's ears, so it was decided to scale the bursts to 10 ms. However, the bilabial bursts in some of the utterances employed in Experiment 1 were shorter than 10 ms. Thus, from the pooled utterances unemployed in Experiment 1, those utterances in which the burst portions were longer than 10 ms were chosen to replace such utterances.

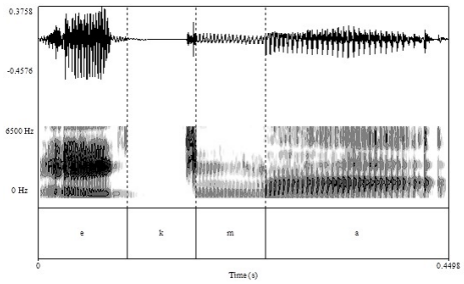
Utterances of all of the three speakers were presented to the listeners, but for the reason to be described below, the male speaker's utterances were completely excluded from the analysis. Among the female speakers' utterances (i.e., those employed for the analysis), /epama/ by both and /epema/, /epima/ and /epuma/ by the second female speaker were those chosen from the pool.

The C-set stimuli were created by deleting all the bursts and the medial vowels (the portions between the bursts and the onset of the following nasal closures) except for the initial 10 ms portions of the bursts. The V-set stimuli were created by deleting 3 pitch periods arbitrarily chosen from the medial vowels. As with Experiment 1, each utterance was presented three times, and the /k/ and the /p/ stimuli within the C- and the V-set were intermixed and presented in a computer-generated random order.

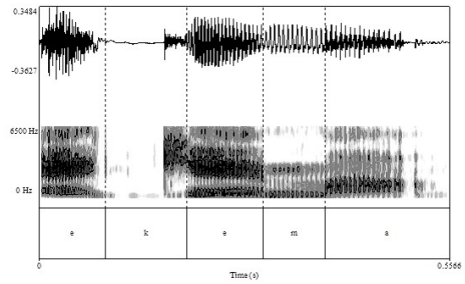
Note that, in the unedited stimuli by the two female speakers (i.e., those stimuli employed



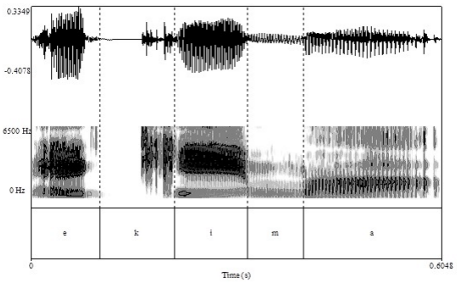
a C-set /ekema/



a C-set /ekima/



a V-set /ekema/



a V-set /ekima/

Figure 5.5: Example waveforms and spectrograms of the stimuli for Experiment 2; the /ekema/ stimuli are by one of the female speakers and the /ekima/ stimuli are by the other female speaker.

for analysis), the mean durations of the /k/ and /p/ bursts were 29 ms (SD = 9.79) and 12 ms (SD = 2.28) respectively, which differed significantly [$t(10) = 2.29,\ p < .001$, two-sided]. Thus the shortening to 10 ms deleted most of the latter halves of the /k/ bursts from the presented stimuli, while most of the /p/ bursts were kept intact.

Some example waveforms and spectrograms of the stimuli are shown in Figure 5.5.

Participants

Nine native Japanese listeners with no known hearing disability participated in the experiment for course credit at undergraduate programs at Hosei University, Tokyo (two males and seven females; mean age = 20, S.D. = .85). None of them had participated in the previous experiment.

Procedure

The same as Experiment 1.

Statistical Analyses

The same as Experiment 1.

5.2.2 Results and Discussion

The original vowel restoration is expected to be almost perfect with the V-set stimuli, which was indeed the case in Experiment 1. However, one of the listeners in Experiment 2 scored only 16 % accuracy within the V-set, where all the other listeners exhibited at least 95 % accuracy. Furthermore, this listener had to repeat the practice session 10 times before reaching the 80 % threshold to proceed to the main session, where all the other listeners only had to repeat the practice session at most twice. Thus this listener was regarded as not having understood the experimental task and hence excluded from the analyses.

Also, it turned out after the experiment that the V-set version of the /ekima/ utterance by the male speaker was erroneously used as his C-set /ekima/ stimulus too, and hence, the results with the male speaker were completely excluded from the analyses.

Generally the pattern observed in Experiment 1 is replicated except that the /i/ identification rate with C-set /ki/ stimuli has now become significantly lower than that with V-set /ki/ stimuli, in line with observations by Beckman & Shoji (1984) (and by Cutler et al., 2009).

Naturally, listeners' identification of the original vowels were pretty good with the V-set stimuli, as can be seen in the total numbers of the responses described in Tables 5.8–5.9. On the other hand, /i/ responses were the most dominant for the C-set /ki/ and /ke/ stimuli, as seen in Tables 5.10–5.11.

The /i/ response rates, the crucial measures, were as described in Table 5.12; the /i/ response rates with the C-set /ki/ and /ke/ seem larger than the chance level ($= .2$), but not as large as

Table 5.8: The numbers of responses for the V-set /k/ stimuli in Experiment 2.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	47	0	0	1	48
	/i/	0	0	47	0	1	48
	/u/	0	0	0	0	48	48
Total		0	47	47	0	50	144

Table 5.9: The numbers of responses for the V-set /p/ stimuli in Experiment 2.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	48	0	0	0	48
	/i/	1	0	46	0	0	48
	/u/	0	2	1	0	46	48
Total		1	50	47	0	46	144

Table 5.10: The numbers of responses for the C-set /k/ stimuli in Experiment 2.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	0	28	2	18	48
	/i/	0	2	32	2	12	48
	/u/	0	0	3	0	45	48
Total		0	2	63	4	75	144

Table 5.11: The numbers of responses for the C-set /p/ stimuli in Experiment 2.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	1	2	0	45	48
	/i/	0	0	6	0	42	48
	/u/	0	1	0	0	47	48
Total		0	2	8	0	134	144

Table 5.12: /i/ response rates with the C and the V-set /ki/ and /ke/ in Experiment 2.

stimulus	mean	S.D.	N
C-set /ke/	**.50**	.51	48
C-set /ki/	**.69**	.47	48
V-set /ke/	**.00**	.00	48
V-set /ki/	**.81**	.39	48

the /i/ response rates in Experiment 1 (99 % for /ki/; 72 % for /ke/), just as expected from the assumed effect of burst shortening.

As with Experiment 1, listener-averaged arcsin square root transformed /i/ identification rates were submitted to two-sided one-sample *t* tests, again with Holm corrections. Again, the /i/ identification rate for the C-set /pu/ stimuli reached the floor (i.e., no /i/ identification), and the rate for the C-set /pe/ was significantly below chance [$t(7) = -6.5561,\ p = 002$], while the rate for the C-set /pi/ did not differ significantly from chance after the Holm correction [$t(7) = -2.3060$, $p = .109$; $p = .055$ if uncorrected]; thus no C-set /p/ stimuli resulted in an above-chance /i/ identification rate.

Similarly, the /i/ identification rate with the C-set /ku/ stimuli was significantly below chance [$t(7) = -5.4067,\ p = .005$]. However, the /i/ identification rate with the C-set /ki/ was significantly above chance [$t(7) = 4.6468$, $p = .009$; $p = .002$ if uncorrected], as well as that with the C-set /ke/ [$t(7) = 4.5783$, $p = .008$; $p = .003$ if uncorrected]. The significant tendency for the C-set /ke/ stimuli to induce /i/ identifications again support the sublexical reality of **DB-sensitivity**; in other words, **DB-sensitivity** persisted after the burst duration shortening.

On the other hand, recall that the /i/ identification rates did not differ significantly in Experiment 1 across the C- and the V-set /ki/ stimuli. In order to see the effects of burst duration shortening, which would presumably result in spectral changes and degradations of the coarticulation cues, listener-averaged transforms of the /i/ identification rates against the C- and the V-set /ki, pi/ stimuli were again submitted to a two-way between-subject ANOVA with

'consonant' and 'set' as independent factors. The main effect of 'consonant' was again highly significant [$F(1,\ 7) = 22.427,\ p = .002$], suggesting that /i/ identification rate was higher for /ki/ stimuli than for /pi/ stimuli; the main effect of 'set' was also highly significant [$F(1,\ 7) = 7.725,\ p < .001$], suggesting that /i/ identification rate was higher for the V-set than for the C-set stimuli. Their interaction was again highly significant [$F(1,\ 7) = 35.166,\ p = .001$], and examinations of simple main effects revealed that /i/ identification rate was higher for /k/ stimuli than for /p/ stimuli within the C-set [$F(1,\ 7) = 32.677,\ p = .001$] but the /k/–/p/ difference was far from being significant within the V-set [$F(1,\ 7) = .000,\ p = 1.000$]; the /i/ identification rate was highly significantly lower for the C-set stimuli than for the V-set stimuli not only in the case of /p/ [$F(1,\ 7) = 108.961,\ p < .001$] but also in the case of /k/ [$F(1,\ 7) = 12.690,\ p = .009$].

Thus the generally the pattern observed in Experiment 1 is replicated except that the /i/ identification rate with C-set /ki/ stimuli has now become significantly lower than that with V-set /ki/ stimuli, in line with observations by Beckman & Shoji (1984) (and by Cutler et al., 2009). It was suggested above that the negative effect of 'devoicing' of /i/ on /i/ perception comes and goes depending on the amount or quality of coarticulation cues, a suggestion supported by across-consonant comparison in Experiment 1. It is now also supported by a within-consonant comparison; a negative effect of /i/ 'devoicing' not observed with non-shortened /k/ bursts in Experiment 1 is indeed observed with shortened /k/ bursts in Experiment 2.[2]

To summarize the findings of Experiment 2, with burst duration of velar bursts, indeed /i/ perception has become more difficult with the C-set /ki/ stimuli than with the V-set stimuli, in line with Beckman & Shoji's (1984) (and Cutler et al.'s 2009) results. This contrasts with the observation in Experiment 1, in which the C- and the V-set /ki/ stimuli were far from being significantly different, which replicated Ogasawara & Warner's (2009) /i/ monitoring

[2]Recall that Experiment 1 employed the results from nine listeners in the analyses, while Experiment 2 employed those from eight listeners; the decreased number of listeners could have made it more difficult to reach significance. Yet the non-significant difference between the C- and the V-set in Experiment 1 has become significant in Experiment 2.

result. Presumably burst duration shortening has resulted in spectral changes and degraded coarticulation cues, an interpretation which accords with the results in Experiment 1 in which **DB-sensitivity** is exerted more with consonants with richer coarticulation cues (velars) than with consonants with poorer coarticulation cues (bilabials). Thus whether vowel perception from vowel-devoiced stimuli is more difficult than vowel perception from non-devoiced stimuli Japanese listeners depend on the amount or quality of coarticulation cues in the consonantal stimuli; this conclusion gives a support to the interpretation suggested in Chapter 3 of the discrepancy among Beckman & Shoji's (1984), Cutler et al.'s (2009) and Ogasawara & Warner's (2009) results.

5.3 Preliminary Considerations for Experiment 3–4

Thus far the experimental task has been, in effect, vowel identification, which involves some sort of meta-linguistic categorization. The observed **DB-sensitivity** effects are further examined below through two AX discrimination experiments., which examine Japanese listeners' discriminations among C-set stimuli.

Generally speaking, discrimination experiments have two advantages over identification experiments: (i) listeners' perception involving less meta-linguistic categorization could be observed, and (ii) a manipulation of interstimulus intervals (ISI's) will help us tease apart the effects of different kinds of processing (Dupoux et al., 1997; Kawasaki et al., 2012; Pisoni, 1973; Werker & Logan, 1985). In the two AX discriminations reported below, particularly the second advantage is exploited in order to examine the reality of **DB-sensitivity**, as opposed to **DI-sensitivity**. That is, the ISI is manipulated across the two experiments; the three hypotheses make different predictions concerning what effects such a manipulation should have. However, each of the two discrimination experiments on its own will also help us compare the **coarticulation insensitivity** hypothesis on the one hand, and the two sensitivity hypotheses on

the other. (The two discrimination experiments, seen separately, are rather unlikely to offer much help for the choice between the two sensitivity hypotheses.) First let us see what would be expected under each hypothesis as results of such an AX discrimination experiment (Subsection 5.3.1), after which we will see what effect of a manipulation of the ISI is expected from each hypothesis (Subsection 5.3.2).

Before we proceed, a note on the nature of **DB-sensitivity** on the one hand, and of **DI-sensitivity** on the other, is needed.

Clearly, **DB-sensitivity** is strongly phonological, because it would put heavy weight on front coarticulation cues but not on non-high coarticulation cues, so that /i/- and /e/-coarticulation would have similar effects. Thus the effects of **DB-sensitivity** are expected only in phonological processing. Furthermore, since it is something that should lead to certain kinds of phonemic categorization, the percepts it would induce are presumably categorical (and could be approximated with Japanese phonemes).

In contrast, presumably **DI-sensitivity** per se is rather auditory, rather than phonological; it claims no weighting differences on coarticulation cues. Given its auditory, rather than phonological, nature, presumably the percepts it would induce are not categorical (at least not categorical in the way the percepts to be induced by **DB-sensitivity** are).

With such differences between the two kinds of coarticulation sensitivity in mind, let us examine what results would be expected from the three hypotheses.

5.3.1 The Predictions for Each Discrimination Experiment

The two discrimination experiments reported below examine Japanese listeners' discriminations among C-set stimuli, and the analyses employ only those results from the C-set stimuli in which the medial vowels are either /e/, /i/ or /o/. Thus the AX pairs employed in the analyses are /e/–/i/, /e/–/u/ and /i/–/u/ (ignoring the order). Let us call them /ei/ pairs, /eu/ pairs, and /iu/

pairs respectively for short.

Clearly, **coarticulation insensitivity** leads us to expect that discriminations should result in uniform failure. In contrast, both kinds of coarticulation sensitivity lead us to expect discrimination successes due to coarticulation-sensitive percepts. However, the predictions of the two sensitivity hypotheses differ from each other.

First consider the **DB-sensitivity** hypothesis. Note that **DB-sensitivity** should induce categorical percepts. It should induce /i/-like percepts for both /i/- and /e/-members within the pairs on the one hand, and /u/-like percepts for /u/-members within the pairs on the other. Thus, clearly, it should contribute to successful discriminations with /iu/ and /eu/ pairs. In the case of /ei/ pairs, however, the prediction is somewhat different. Under **DB-sensitivity**, discriminations with /ei/ pairs should fail in the following two cases:

(a) failed exploitation of both /i/- and /e/-coarticulation cues (/u/–/u/ percepts).

(b) successful exploitation of both (/i/–/i/ percepts).

and should succeed in the following two cases:[3]

(c) successful exploitation of /i/-coarticulation cues but failed exploitation of /e/-coarticulatuion cues (/i/–/u/ percepts).

(d) successful exploitation of /e/-coarticulation cues but failed exploitation of /i/-coarticlatuion cues (/u/–/i/ percepts).

Thus not only extremely unsuccessful exploitation (a) but also extremely successful exploitation (b) of coarticulation cues should result in discrimination failures with /ei/ pairs. Discriminations should succeed with /iu/ pairs in the (a) and (c) cases, and with /eu/ pairs in the (a) and (d) cases. Thus the expected success rates can be approximated as:

[3]Given the **production presuppositions**, (d) is rather unlikely. However, 'more chances for /i/-like percepts with /i/-coarticulation cues than with /e/-coarticulation cues' does not logically exclude the possibility that the few cases of /i/-like percepts with /e/-coarticulation cues sometimes coincide with the few cases of non-/i/-like percepts with /i/-coarticulation cues, and hence (d) cannot be excluded a priori.

(b) + (c) for /iu/ discriminations

(b) + (d) for /eu/ discriminations

(c) + (d) for /ei/ discriminations

Without a clue for the expected rates of (a)–(d), any of those could be higher or lower than the others. The only clue we have is the prediction of the **production presuppositions**, according to which /i/-like percepts should be more likely for /i/-members than for /e/-members within the pairs. This prediction leads us to assume that the rate of the (d) cases (i.e., successful exploitation of /e/-coarticulation without successful exploitation of /i/-coarticulation) is close to zero. If we substitute (d) with zero, then the approximations of the expected success rates will reduce to:

(b) + (c) for /iu/ discriminations

(b) for /eu/ discriminations

(c) for /ei/ discriminations

We have no clue for the expected rate of (b) or (c) except that the lowest possible value is zero. Thus the only prediction from the **DB-sensitivity** will be:

/iu/ discrimination successes $\geq$ /eu/ discrimination successes

/iu/ discrimination successes $\geq$ /ei/ discrimination successes

(The **production presuppositions** state that coarticulation cues are poor within bilabials, and hence whatever kind of coarticulation sensitivity is not expected to lead to successful discriminations anyway. The above predictions are for stimuli with rich coarticulation cues, such as velar pairs, but incidentally, the equality signs appear above, which can cover bilabial cases too.)

Next consider the **DI-sensitivity** hypothesis. Note that it is presumably auditory in nature and hence the percepts it would induce are not categorical in the way the percepts expected

from **DB-sensitivity** are. Successful exploitations of whatever coarticulation should contribute to successful discriminations with all the three pairs. However, we do not know which pairs should be more similar than which in their auditory quality. Thus it is not clear which pairs should benefit more than which pairs from **DI-sensitivity**.[4]

Thus, when seen separately, the results of the two discrimination experiments are likely to be informative only with respect to whether the **coarticulation insensitivity** hypothesis should be adopted or rejected. Given the **production presuppositions**, discriminations with all the bilabial pairs are likely to fail; thus only the results with the velar pairs could be expected to help the choice from the three hypotheses. An observation of significantly successful discriminations with any velar pair would argue against the **coarticulation insensitivity** hypothesis, while an observation of significantly unsuccessful discriminations with all the velar pairs would argue for the **coarticulation insensitivity** hypothesis. However, when it comes to the choice between the **DB-sensitivity** hypothesis vs. the **DI-sensitivity** hypothesis, even the results with velar pairs are unlikely to be informative, because (i) only **DB-sensitivity** makes specific predictions about the relative success rate differences among the /ei/, /eu/, and /iu/ pairs, and hence, unless those specific predictions of **DB-sensitivity** were betrayed by the results, they would be compatible with both, and (ii) even when the specific predictions of **DB-sensitivity** are betrayed, the results could be interpreted as mixtures of the predicted effects of **DB-sensitivity** and the unknown effects of **DI-sensitivity**.[5]

[4]Intuitively, /e/–/i/ may be more similar than /e/–/u/ or /i/–/u/, but this thesis has no empirical evidence to offer for that intuition. Note that, if that intuition is correct, the prediction would not markedly different from what would be expected from **DB-sensitivity**.

[5]Upon reflection, this shortcoming could have been remedied if discrimination across C- and V-set stimuli, rather than among C-set stimuli, were examined.

If discriminations between C- and V-set stimuli were examined, the **coarticulation insensitivity** hypothesis would predict a tendency for the 'different' responses between V-set /i/ stimuli and C-set /i/ stimuli, for the latter of which coarticulation-insensitive default epenthesis should result in /u/-like percepts; in contrast, the two sensitivity hypotheses would predict a tendency for the 'same' responses for such discriminations (given enough coarticulation traces in the C-set stimuli), because the C-set stimuli should sound like /eCima/. The predictions of the two sensitivity hypotheses would differ with respect to whether V-set /i/ stimuli and C-set /e/ stimuli should be heard as the same or different; given enough coarticulation traces in the C-set stimuli, the **DB-sensitivity** hypothesis would predict a tendency for the 'same' responses because C-set /e/ stimuli should sound like /eCima/, whereas the **DI-sensitivity** hypothesis would predict a tendency for the 'different' responses because C-set /e/ stimuli should not sound like /eCima/.

Unfortunately, examinations of discriminations between C- and V-set stimuli have to be left for future research.

However, note that the (ii) interpretation is an interpretation according which both **DB-sensitivity** and **DI-sensitivity** are real. Thus, if some coarticulation sensitivity is observed, and if the specific predictions by the **DB-sensitivity** hypothesis are betrayed, that would mean that the observed coarticulation sensitivity should be attributed to **DI-sensitivity** alone, or to the mixtures of **DB-sensitivity** and **DI-sensitivity**, but not to **DB-sensitivity** alone. Indeed, if the specific predictions by the **DB-sensitivity** are borne out, we would have no clue to the determination of which kind of coarticulation sensitivity is or is not operative, but if some coarticulation sensitivity is observed but the relative discrimination successes do not conform to the pattern expected from **DB-sensitivity** alone, we would have to conclude that at least **DI-sensitivity** is real; thus the specific predictions should be examined, by comparing relative differences among the discrimination success rates with the /ei/, /eu/ and /iu/ pairs.

5.3.2 The Predictions for Cross-experimental Comparisons

Now let us return to the second advantage of discrimination experiments, i.e., the possibility of teasing apart the effects of different processing stages through a manipulation of ISI's. A comparison between the two discrimination experiments will certainly be informative with respect to which kind of coarticulation sensitivity is operative.

It has already been noted in the literature that listeners' responses tend to be based on phonological, rather than acoustic, representations, when (i) the experimental task is memory-demanding (e.g., ABX rather than AX), (ii) the interstimulus interval (ISI) is long (say, 1,500 ms), and (iii) the stimuli to be compared are uttered by different speakers and hence cannot be compared without resorting to phonological representations (Dupoux et al., 1997; Kawasaki et al., 2012; Matthews & Brown, 2004; Pisoni, 1973; Werker & Logan, 1985). Unfortunately, calculation of the number of stimulus for Experiments 3–4 suggested that the estimated total amount of time for a single ABX session based on the current stimuli for Experiments 3–

However, the simultaneous evaluation of the two kinds of coarticulation sensitivity enabled by the (ii) interpretation would not be possible with such a design. (See the discussion immediately below.)

4 would be too long,[6] and hence an ABX design was avoided; instead, an AX design was employed throughout. Instead, in Experiments 3–4, the ISI is manipulated.

Pisoni (1973) attributed the effect of ISI manipulation to auditory memory decay.[7] However, Kawasaki et al. (2012) obtained evidence that auditory representations are 'erased' by the interference by phonological processing completed with a long enough ISI, while auditory memory decay in time in the absence of the interference by phonological processing is also admitted. So let us assume that, with a shorter ISI, the percepts will depend *less* on phonological processing and *more* on auditory processing. Thus ISI shortening is assumed to result in decreased effects of **DB-sensitivity** (if it is real) on the one hand, and increased effects of **DI-sensitivity** (if it is real) on the other.[8]

If so, the **DB-sensitivity** and the **DI-sensitivity** hypothesis lead us to expect different effects of ISI shortening on discrimination performance. The **DB-sensitivity** hypothesis leads us to expect that ISI shortening should result in some loss of discrimination successes, while the **DI-sensitivity** hypothesis leads us to expect that ISI shortening should result in some gain

[6]The participants were recruited from the undergraduate population at Hosei University, Japan. The estimated single session duration sounded to an informal consultant (who graduated from Hosei) to be too long for the additional credit to be given so that ordinary Hosei undergraduates would be rather unlikely to participate.

[7]Matthews & Brown (2004) attribute this claim to Werker & Logan (1985). However, Werker & Logan simply cited Pisoni (1973), without committing themselves to that claim.

On the other hand, Werker & Logan manipulated the ISI and claimed to have found evidence for three (rather than two) independent factors behind speech perception, namely, 'auditory', 'phonetic' and 'phonological', which manifest their effects on perception to different degrees depending on the ISI. While it is clear that phonological factors were claimed to manifest their effects most when the ISI was 1,500 ms, it is not clear which of phonetic factors and auditory factors were claimed to manifest their effects at which ISI (250 ms, or 500 ms); their own writing is extremely unclear. Werker's (1991:98) interpretation claims the 250 ms–phonetic and 500 ms–auditory combination, while Matthews & Brown's (2004) interpretation claims the 250 ms–auditory and 500 ms–phonetic combination. Conceptually, the Werker combination is very mysterious; phonetic factors at the shortest, phonological the longest, and auditory in between. Furthermore, it is not clear to me why the statistical analyses reported by Werker & Logan (19854) could be interpreted a la Werker (1991).

However, for the purpose of this thesis, the auditory vs. phonetic distinction is not crucial, where 'phonetic' refers to those distinctions found only in non-native languages; none of the discrimination experiments conducted for this thesis involves a distinction found in non-Japanese languages. Thus only a two-way classification of 'factors' is made: auditory vs. phonological.

[8]Throughout this thesis it is assumed that one-step models claim that **DI-sensitivity**, if real, should be exerted in phonotactic repair, while phonotactic repair is clearly phonological. If long ISI's would be required for phonological processing (phonotactic repair) on the one hand, and would discourage effects of auditory processing (**DI-sensitivity**) on the other, the chances for observing coarticulation-sensitivity in phonotactic repair should indeed be small; it could be observed only when phonotactic repair is completed before the decay of coarticulation-sensitive percepts in auditory memory. This consequence is compatible with the possibility that **DI-sensitivity** is in fact real but its effects were too weak to win the effects of **DB-sensitivity** in Experiments 1–2; due to such conflicting temporal requirements for phonotactic repair and for **DI-sensitivity**, the effect of **DI-sensitivity** is rather hard to observe.

of discrimination successes. Of course, the **coarticulation insensitivity** hypothesis leads us to expect that ISI shortening should have no effect, because discrimination performance should be at the floor irrespective of the ISI.

Indeed, there is the possibility that both **DB-sensitivity** and **DI-sensitivity** are operative, and the effects of the two kinds of coarticulation sensitivity would compete. In such a case, worse discriminations resulting from ISI shortening could be interpreted as suggesting that the magnitude of the reduction of the effects of **DI-sensitivity** was larger than the gain of the effects of **DB-sensitivity**; similarly, better discriminations resulting from ISI shortening could be interpreted as suggesting that the magnitude of the gain of the effects of **DB-sensitivity** was larger than the reduction of the effects of **DI-sensitivity**. However, worse discriminations should not be observed unless *at least* **DI-sensitivity** is real and its effects decreased with ISI shortening, and better discriminations should not be observed unless *at least* **DB-sensitivity** is real and its effects decreased with ISI shortening. Thus, unless the lack of a significant effect of ISI shortening is observed (which could be interpreted either in terms of the absence of both kinds of sensitivity, or in terms of the equal magnitudes of the loss and the gain), a cross-experimental comparison would tell us *at least* which kind of coarticulation sensitivity is operative.

5.4 Experiment 3

Experiment 3 is the first of the two discrimination experiments reported here. As noted above, it has already been noted in the literature that listeners' responses tend to be based on phonological, rather than acoustic, representations, when (i) the experimental task is memory-demanding (e.g., ABX rather than AX), (ii) the interstimulus interval (ISI) is long (say, 1,500 ms), and (iii) the stimuli to be compared are uttered by different speakers and hence cannot be compared without resorting to phonological representations (Dupoux et al., 1997; Dupoux et al.,

2001; Matthews & Brown, 2004; Werker & Logan, 1985). Experiment 3 attempted to encourage the listeners to focus on phonological representations by manipulating (ii) (= ISI) and (iii) (= speaker variations); the stimuli employed in Experiment 2 (i.e., utterances by two female speakers) were employed, as a result of which the A and the X stimuli were always uttered by different speakers; the ISI was set to 1,500 ms.

5.4.1 Method

Stimuli

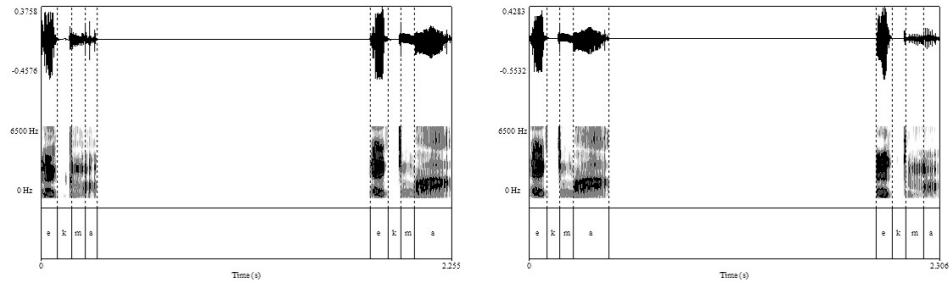
The twenty C-set stimuli from Experiment 2 (two consonant $\times$ five vowels $\times$ two speakers) were first divided into those in which the medial consonant was /k/ (the /k/ set) and those in which the medial consonant was /p/ (the /p/ set). Each stimulus in the /k/ set by either one of the two speakers was concatenated with each stimulus in the /k/ set by the other speaker, with an ISI of 1,500 ms. Since there were five stimuli by each speaker in the /k/ set, 50 files (5 original files by one speaker $\times$ 5 original files by the other speaker $\times$ 2 speaker order) were obtained from the /k/ set. Similarly for the /p/ set. Thus the intended discrimination was among /k/ bursts or among /p/ bursts, not across /k/-/p/ bursts.

The /k/ and the /p/ pairs were intermixed and presented in a computer-generated random order through circumaural closed-back headphones (SONY MDR-ZX700).

The waveforms and spectrograms of example trails are shown in Figure 5.6.

Participants

Ten native Japanese listeners with no hearing disability participated in the experiment for course credit at undergraduate programs at Hosei University, Tokyo (5 males and 5 females; mean age = 20, S.D. = .29). None had participated in the previous experiments.



/ekuma/–/ekima/ /ekima/–/ekema/

Figure 5.6: Example waveforms and spectrograms of the stimuli for Experiment 3; the left panel shows /ekuma/–/ekima/, while the right panel shows /ekima/–/ekema/; /ekima/ is by the first female speaker, while /ekuma/ and /ekema/ are by the second female speaker.

Procedure

The experiment was controlled by E-prime and conducted in a sound-attenuated room. The participants were told that the experiment was meant to examine how non-Japanese sounds are perceived by Japanese listeners (instructions imitating Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999), and they were instructed to judge whether the portions between /e/ and /ma/ were the same. The responses were made by pressing either the "1" or the "2" keys on the computer ("1 = same" and "2 = different").

Again, only after scoring more than 80 % accuracy in the practice session were they allowed to proceed to the experimental session; the stimuli in the practice session were 12 files created by concatenating the experimenters' and another female speaker's productions of /t͡ɕu/, /t͡ɕi/, /ɕu/, /ɕi/, for half of which the correct response was 'the same' and for another half of which the correct response was 'different'.[9]

Statistical Analyses

As with Experiments 1–2, the analyses employ only the results from the stimuli in which the medial vowels are either /e/, /i/ or /u/. All the analyses are conducted separately for the /k/ and

[9]Speaking more precisely, they were: /t͡ɕu/-/t͡ɕi/-03, /ɕu/-/t͡ɕi/-30, /t͡ɕu/-/t͡ɕu/-03, /t͡ɕu/-/t͡ɕu/-30, /ɕi/-/t͡ɕu/-30, /t͡ɕu/-/ɕu/-03, /ɕi/-/t͡ɕi/-30, /ɕi/-/ɕi/-03, /ɕi/-/ɕi/-30, /ɕi/-/ɕu/-03, /ɕu/-/ɕu/-03, /ɕu/-/ɕu/-30, where '03' refers to the oder 'the experimenter's voice – the female voice' and '30' refers to the reverse order.

the /p/ set. Three different measures were employed.

The first is the arcsin square root transformed listener-averaged accuracy rates. However, responses are presumably joint products of (a) listeners' perceptual sensitivity to differences between the stimuli and (b) response decision processes, rather than pure reflections of their sensitivity; since they conflate the results of listeners' sensitivity on the one hand, and response biases in the decision processes on the other, accuracy rates might not be the best measure. Thus the accuracy rate analysis is supplemented by analyses employing the second and the third measure.

The second measure is the d' value of Signal Detection Theory (SDT) (Macmillan & Creelman, 2005). Assume that four kinds of pairs, AB, BA, AA, and BB are presented to the participants; AB and BA pairs are pairs of different stimuli and hence 'different' responses are correct (called 'hits') and 'same' responses are wrong (called 'misses'); AA and BB pairs are pairs of the same stimuli and hence 'different' responses are wrong (called 'false alarms') and 'same' responses are correct (called 'correct rejections'). The SDT measure d' is intended to be a measure of response-bias-independent sensitivity, calculated based on the hit rates (the proportion of the hits given AB or BA stimuli) and the false alarm rates (the proportion of the false alarms given AA or BA stimuli), while the β value is intended to measure the participants' response bias in the form of the likelihood ratio of hits vs. false alarms.

In fact, d' computations are rather complicated. The first complication comes from the fact that there are several ways to compute d', partially depending on the experimental design. For AX (same-different) discriminations, Macmillan & Creelman propose that participants cope with the discrimination task either with the Independent Observations (IO) strategy or with the Differencing strategy; if the IO strategy is adopted, the participants' response rule is to decide, separately for each member of a given pair, whether it is A or B, and then report whether the results of the subdecisions are the same or different, whereas if the Differencing strategy is adopted, their response rule is to compare the two members of a given pair and

report whether the difference between them exceeds some threshold (a 'different' response) or not (a 'same' response). The d' calculations differ depending on which strategy is assumed to have been adopted by the participants. Both IO computations and Differencing computations have been adopted in the previous speech research. For example, Francis & Ciocca (2003) adopted the Differencing computations in analyzing AX discriminations with synthetic stimuli, while Kabak (2003) adopted the IO computations in analyzing AX discriminations with natural stimuli. The second complication comes from the SDT presupposition that the false alarm rate should never be larger than the hit rate; however, in actual experimental results, the false alarm rates occasionally exceed the hit rates, and some decision has to be made concerning how to deal with such data. The third complication comes from the SDT presupposition that neither the hit or the false alarm rate should be 1.0 or .0; again, in actual experimental data, the hit or the alarm rates are occasionally 1.0 or .0, and some decision has to be made concerning how to deal with such data.

In this thesis, Kabak's (2003) procedure is imitated. That is, IO computations of d' are adopted, rather than Differencing computations; when the false alarm rate is larger than the hit rate, d' is assumed to be zero; when either the hit or the false alarm rate is observed to be 1.0 or .0, the hit or the false alarm count (count, not rate) is reduced or increased by .5 so that the corrected hit or alarm rate will not be 1.0 or .0 (Macmillan & Creelman, 2005:8).[10]

The third measure is the 'hit rate minus false alarm rate' (H–FA) scores for each listener-pair combination. Since the d' value computation (whether based on the IO or the Differencing strategy) involves "corrections" for 0 and 1 H or FA rates, Francis & Ciocca (2003) followed Maddox & Estes (1997) and employed H–FA scores as alternative measures to d' values; this

[10]Kabak (2003:104–105) argue that the IO computations should be better than Differencing computations, while Francis & Ciocca (2003) simply employ Differencing computations with no argument.

In this thesis, the IO computations of the d' values are executed with Kenneth Knoblauch's 2013 'psyphy' package for the R language (http://cran.r-project.org/web/packages/psyphy/index.html). The IO computations of the β values are executed with Macmillan & Creelman's (2005:220) following formula,

$$\beta = \phi(\text{hit rate})/\phi(\text{false alarm rate})$$

Table 5.13: The mean accuracies for each pair in Experiment 3.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	.48	.64	.73	.51	.51	.54
S.D.	.17	.15	.20	.15	.20	.12
N	10	10	10	10	10	10

Table 5.14: The d' values for each pair in Experiment 3.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	1.02	1.83	2.77	.66	.66	1.02
S.D.	1.02	1.93	2.01	1.40	1.43	1.35
N	10	10	10	10	10	10

thesis also employs the H–FA scores as the third measure.

To examine if coarticulation sensitivity is exhibited in the first place, the measures are submitted to two-sided one-sample *t* tests, separately for the six pairs (/k/ vs. /p/; /ei/ vs. /eu/ vs. /iu/), with the expected value being the arcsin square root transform of .5, (for accuracy rates) and 0 (for d' and H–FA scores), with Holm corrections separately for the /k/ and the /p/ set.

Next, the listener-averaged dependent measures with /iu/, /eu/ and /ei/ pairs are submitted to two-sided repeated-measures *t* tests for pairwise comparisons among pairs, to examine the specific predictions by the **DB-sensitivity** hypothesis.

5.4.2 Results and Discussion

The raw results are summarized in Tables 5.13–5.16; discriminations seem to be good at least with /ke/–/ku/, which is rather unexpected from the **coarticulation insensitivity** hypothesis.

Descriptively, for /ke/-/ki/, the averaged accuracy score is below chance (Table 5.13) and the H–F value was negative (Table 5.16), which should be observed only when listeners responded 'different' more often to identical vowel pairs than to distinct vowel pairs (the larger false alarm rate, .475, than the hit rate, .425). Such a situation is indeed a possibility in those

Table 5.15: The β values for each pair in Experiment 3.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	1.06	1.13	1..28	1.06	1.08	1.08
S.D.	.08	.17	.24	.12	.18	.12
N	10	10	10	10	10	10

Table 5.16: The mean H–FA scores for each pair in Experiment 3.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	−2.00	1.10	1.80	.10	.10	.30
S.D.	1.40	1.20	1.62	1.20	1.52	.95
N	10	10	10	10	10	10

AX experiments conducted for this thesis (including Experiment 3) in which the two members of each pair are uttered by different speakers, if the listeners' strategy is the Differencing strategy rather than the IO strategy (against our decision to imitate Kabak 2003). Assume that something like the coordinate in Figure 5.7 describes the degrees to which the percept from the members of /e/-/i/ pairs, produced by speakers A and B, are perceptually /i/-like. Assume that two members of a pair are judged to be different when their distance is three units or more in Figure 5.7, in which case we should expect:

- /e/-/e/ and /i/-/i/ pairs should be judged different, because they are four units apart;
- /e/-/i/ pairs, where /e/ and /i/ were produced by A and B respectively, should be judged different, because they are 10 units apart;
- /e/-/i/ pairs, where /e/ and /i/ were produced by B and A respectively, should be judged the same, because they are only two units apart.

Thus, although the two speakers' /e/ utterances do not overlap perceptually with their /i/ utterances, the false alarm rate and the hit rate could be close to 100 % and 50 % respectively, resulting in more false alarms than hits. Such a situation arises when inter-speaker distances between A and B are larger than the distance between the closer endpoints '/e/ by B' and '/i/

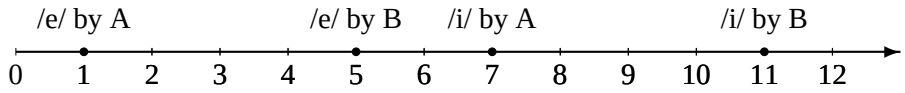


Figure 5.7: A hypothetical perceptual dimension against speaker A's and speaker B's productions of /e/- and /i/-stimuli.

by A'.[11] Under this scenario, /i/- and /e/-members are indeed distinguished, but the effect of the listeners sensitivity to the unintended dimension (inter-speaker variation) is mixed with the effect of their sensitivity to the intended dimension (inter-vowel variation). Of course, when they are indeed *in*sensitive (to both dimensions) and the *population* hit and false alarm rates are equal, the *sample* differences between them could be either positive or negative as a result of random variability. Whether the above scenario is real,[12] or the observed larger false alarm rates than hit rates are simply a result of random variation, is not immediately clear. However, if the 'unintended dimension' scenario is real when the false alarm rates exceeded the hit rates, probably it was also real when the false alarm rates did not exceed the hit rates, and that scenario should only lower the accuracy rates (or the 'H–FA' values), and hence, if listeners' sensitivity turn out to be significantly above chance in the statistical tests below, probably such results would be trustworthy.

To examine the listeners' sensitivity to coarticulation, the arcsin square root transformed lister-averaged accuracy scores with each pair were first submitted to two-sided one-sample *t* tests, with the expected value of the arcsin square root transform of .5, with the significance values being Holm corrected separately for /k/ and /p/. The discrimination accuracy scores did

[11] The chances for such a situation also depend on the 'threshold' distance (in SDT terms, the *k* value; Macmillan & Creelman, 2005:223–225); for example, in the above example, if the two members of a pair are judged to be different if they are only two units apart, all the pairs would then be judged to be different; alternatively, if the required minimum distance is six units, then all the pairs would then be judged to be the same.

[12] As noted, the above scenario assumes that the listeners adopted the Differencing strategy. However, not only the IO comutations but also the Differencing computations of the d' values presuppose that the false alarm rates should not exceed the hit rates, and hence, if the above scenario is indeed real, SDT analyses will be rather inappropriate in the first place.

not differ from chance in the case of /p/ pairs [with /ei/, $t(9) = .274$, $p = .790$, uncorrected; with /eu/, $t(9) = .249$, $p = .809$, uncorrected; $t(9) = 1.009$, $p = .340$, uncorrected]. However, in the case of /k/ pairs, discrimination was significantly above chance for /eu/ [$t(9) = 2.864$, $p = .037$; $p = .019$ if uncorrected] and for /iu/ [$t(9) = 3.041$, $p = .042$; $p = .014$ if uncorrected], while descriptively below-chance discrimination for /ei/ did not differ significantly from chance [$t(9) = .458$, $p = .658$, uncorrected].[13]

A similar pattern of results was observed when d' values were submitted to two-sided one-sample t tests with the expected value being 0. Again, all the /p/ pairs failed to reach significance (after Holm corrections for /iu/) [with /ei/, $t(9) = 1.494$, $p = .169$ uncorrected; with /eu/, $t(9) = 1.458$, $p = .179$ uncorrected; with /iu/, $t(9) = 2.388$, $p = .122$ if corrected, $p = .041$ if uncorrected]. However, in the case of /k/ pairs, d' was significantly above 0 for /iu/ [$t(9) = 4.350$, $p = .006$ if corrected, $p = .002$ if uncorrected], for /eu/ [$t(9) = 2.994$, $p = .030$ if corrected, $p = .015$ if uncorrected], and for /ei/ [$t(9) = 2.284$, $p = .048$]. However, for /ei/, this result should be seen with caution; descriptively, the false alarm rates did exceed the hit rates with five listeners out of ten, for whom the d' values were assumed to be zero, a considerable 'correction' to raise the mean d' value.

A similar pattern of results was also observed when two-sided one-sample t tests were conducted on the H–FA scores, with the expected value being 0. Again, all the /p/ pairs failed to reach significance [with /ei/, $t(9) = .264$, $p = .798$, uncorrected; with /eu/, $t(9) = .208$, $p = .840$, uncorrected; with /iu/, $t(9) = 1.000$, $p = .343$]. However, in the case of /k/ pairs, discrimination was significantly above chance for /eu/ [$t(9) = 2.905$, $p = .035$; $p = .017$ if uncorrected] and for /iu/ [$t(9) = 3.515$, $p = .020$; $p = .007$ if uncorrected], while descriptively below-chance discrimination for /ei/ did not differ significantly from chance [$t(9) = -.452$, $p = .662$, uncorrected].

Given the assumed poor amount of coarticulation traces within bilabials (the **production**

[13] Thus the 'unintended discrimination' scenario was not statistically supported.

presuppositions), the non-significance with /p/ pairs is not surprising, irrespective of which kind of coarticulation sensitivity is real or not. On the other hand, the /ke/–/ki/ discrimination was significantly successful when the measure is d', but not when the measure is the accuracy or the H–FA scores; note that, when the false alarm rates exceed the hit rates, only the latter two measures reflect the magnitudes of the differences; the d' computations simply regard the differences to be zero; thus the 'significantly above chance' result with d' is very weak evidence for asserting listeners' successful discriminations with /ke/–/ki/. However, the /ke/–/ku/ and the /ki/–/ku/ discriminations were significantly successful whatever the measure is, and hence it seems reasonable to conclude that /ke/–/ku/ and /ki/–/ku/ discriminations were indeed successful. Such results are hard to interpreted unless some sort of coarticulation sensitivity is indeed real. Thus the results support coarticulation sensitivity.

Next, the arcsin square root transforms of the listener-averaged success rates with /ei/, /eu/ and /iu/ pairs were compared through two-sided repeated-measures t tests with Holm corrections, separately for the /k/ and the /p/ set results. Recall that the specific predictions by the **DB-sensitivity** hypothesis are: (i) /iu/ discriminations should not be less successful than /eu/ discriminations, and (ii) /iu/ discriminations should not be less successful than /ei/ discriminations. Among the /k/ pairs, the /iu/ discrimination was significantly better than the /ei/ discrimination [$t(9) = 3.356$, $p = .025$; $p = .008$ if uncorrected]; the /eu/ discrimination showed only a tendency towards being significantly better than the /ei/ discrimination after a Holm correction [$t(9) = 2.373$, $p = .083$; $p = .042$ if uncorrected].; the /iu/ discrimination did not differ significantly from the /ke/–/ku/ discrimination after a Holm correction [$t(9) = 2.330$, $p = .045$ if uncorrected, but the Holm procedure declares that this is non-significant].[14] On the other

[14]Suppose that we have four comparisons, and the smallest uncorrected p value was .01. The Holm correction procedure multiplies this value by four (the number of the remaining comparisons), which results in .04. Since this corrected p value is smaller than the conventional significance value of .05, the procedure tells us that this first comparison should be regarded as significant and prompts us to examine the next comparison. Suppose that the next smallest uncorrected value was .02. Since we have three more comparisons left, the procedure dictates that this value should be corrected by multiplying it by three, which results in .06. Since this corrected value is larger than .05, the procedure dictates that this comparison, as well as all the remaining comparisons, should be regarded as non-significant.

What if we ignore this dictation and continue the procedure? Suppose that the next smallest uncorrected value

hand, no comparison among /p/ pairs approached significance even before Holm corrections.

Similar comparisons were made with the d' values. Among the /k/ pairs, the /iu/ discrimination was significantly better than the /ei/ discrimination before a Holm correction but only showed a tendency towards significance after a Holm correction [$t(9) = 2.603$, $p = .086$ if corrected, $p = .029$ if uncorrected]; similarly, the /iu/ discrimination was significantly better than the /eu/ discrimination before a Holm correction but only showed a tendency towards significance after a Holm correction [$t(9) = 2.494$, $p = .068$ if corrected, $p = .034$ if uncorrected]; the /ei/–/eu/ comparison simply failed to reach significance [$t(9) = 1.129$, $p = .288$ uncorrected]. No comparison reached significance in the case of /p/ pairs [for /eu/–/iu/, $t(9) = .863$, $p = .411$ uncorrected; for /ei/–/iu/, $t(9) = .570$, $p = .582$ uncorrected; for /ei/–/eu/, $t(9) = .010$, $p = .992$].

Similar comparisons were also made with the H–FA scores. Among the /k/ pairs, the /iu/ discrimination was significantly better than the /ei/ discrimination [$t(9) = 3162$, $p = .035$; $p = .012$ if uncorrected]; the /eu/ discrimination only showed a tendency towards being significantly better than the /ei/ discrimination after a Holm correction [$t(9) = 2.333$, $p = .089$; $p = .045$ if uncorrected]; the /iu/ discrimination did not differ significantly from the /eu/ discrimination [$t(9) = 2.327$, $p = .045$ if uncorrected, but the Holm procedure again declares that this is non-significant]. On the other hand, no comparison among /p/ pairs approached significance even before Holm corrections.

Thus the only solid result, if any, from the comparisons among the pairs is the better /ki/–/ku/ discrimination than the /ke/–/ki/ discrimination, which is consistent with the expectation from **DB-sensitivity** that /iu/ discriminations should be at least as successful as /eu/ or /ei/ discriminations. It is also consistent with **DI-sensitivity**, which makes no specific prediction about the relative discrimination successes among the pairs. Thus, while the better /ki/–/ku/

was .023. Since we have two more comparisons left, the procedure would correct this value by multiplying it by two, which would result in .046, which is smaller than .05. Thus, if we do not stop the procedure when it dictates that we should, we would end up regarding a comparison with the uncorrected value of .023 as significant while regarding a comparison with the uncorrected value of .02 as non-significant.

discrimination than the /ke/–/ki/ discrimination would indeed argue against **coarticulation insensitivity**, the comparisons among the pairs do not tell us whether the coarticulation sensitivity operative in Experiment 3 is **DB-sensitivity** or **DI-sensitivity**.

5.5 Experiment 4

As stated above, a shorter ISI is expected to discourage the effects of phonological processing and encourage the effects of auditory processing. Thus, if the successful discriminations in Experiment 3 are at least partially due to **DB-sensitivity**, ISI shortening should result in some loss of its effect; if they are at least partially due to **DI-sensitivity**, ISI shortening should result in some gain of its effects. Thus an observation of *less* successful discrimination would constitute evidence that at least **DB-sensitivity** is real, while an observation of *more* successful discriminations would constitute evidence that at least **DI-sensitivity** is real.

Furthermore, while we failed to observe clear results concerning the specific predictions by the **DB-sensitivity** hypothesis concerning success rate differences among the /ei/, /eu/ and /iu/ pairs, the gain or the loss expected from the reality of the two kinds of sensitivity might result in clearer differences among the three pairs.

Thus, by changing the ISI, we might be able to obtain discrimination-based evidence for the nature of the coarticulation sensitivity observed in Experiment 3. In Experiment 4, which is the second of the discrimination experiments reported here, the ISI is shortened; in Experiment 3, it was 1,500 ms, but in Experiment 4, it is 250 ms.

5.5.1 Method

Stimuli

The stimuli of Experiment 3 were modified so that the ISI's were 250 ms, rather than 1,500 ms.

Table 5.17: The mean accuracies for each pair in Experiment 4.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	.44	.47	.75	.56	.52	.42
S.D.	.13	.09	.19	.09	.14	.09
N	8	8	8	8	8	8

Participants

Eight native Japanese listeners with no known hearing disability participated in the experiment for course credit at undergraduate programs at Hosei University, Tokyo (three males and five females; mean age = 19, S.D. = .87). None of them had participated in the previous experiments.

Procedure

The same as that for Experiment 3.

Statistical Analyses

The analyses conducted for Experiment 3 will be repeated. In addition, the /eu/ and /iu/ results of Experiments 4 are compared with those of Experiment 3 by conducting an independent t test for each pair (with the three dependent measures being the arcsin square root transformed accuracy scores, d' values, and the H–FA scores), with Holm corrections (separately for the /k/ and the /p/ set) for the significance values.

5.5.2 Results and Discussion

The results of Experiment 4 are first examined, after which the results of Experiment 3–4 are compared.

The raw results are summarized in Tables 5.17–5.20.

The /ki/–/ku/ discriminations seem pretty good, which is unexpected from the **coarticu-**

Table 5.18: The d' values for each pair in Experiment 4.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	.42	.20	3.12	.78	.95	.00
S.D.	.78	.56	2.08	1.11	1.02	.00
N	8	8	8	8	8	8

Table 5.19: The β values for each pair in Experiment 4.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	1.03	1.02	1.26	1.08	1.08	1.00
S.D.	.06	.05	.23	.11	.08	.00
N	8	8	8	8	8	8

lation insensitivity hypothesis. On the other hand, this time, the accuracy score was below chance (the false alarm rate was greater than the hit rate) not only with /ke/–/ki/ but also with /ke/–/ku/ and with /pi/–/pu/. Such results would make sense under the 'unintended dimension' scenario described in the discussion of the results of Experiment 3, in which the sensitivity to non-phonemic inter-speaker variation could result in such an observation, because ISI shortening should increase listeners' sensitivity to non-phonemic auditory variations; alternatively, they were simply due to random variations.

The listener-averaged accuracy scores for each pair were arcsin square root transformed and submitted to two-sided one-sample t tests, with the expected value being the arcsin square root transform of .5 and the significance values being Holm corrected separately for /k/ and /p/ pairs. The /ki/–/ku/ discrimination was significantly better than chance [$t(7) = 3.258$, $p = .042$; $p = .014$ if uncorrected], while no other pair reached significance (even if uncorrected). Sim-

Table 5.20: The mean H–FA scores for each pair in Experiment 4.

pair	ke-ki	ke-ku	ki-ku	pe-pi	pe-pu	pi-pu
mean	−.50	−.25	2.00	.50	.13	−.63
S.D.	1.07	.71	1.51	.76	1.13	.74
N	8	8	8	8	8	8

ilar results were obtained with two-sided one-sample t tests conducted for d' values; the /ki/–/ku/discrimination was significantly good [$t(7) = 4.252$, $p = .011$; $p = .004$ if uncorrected], while no other pair reached significance even if uncorrected, except /pe/–/pu/, for which the uncorrected value was significant [$t(7) = 2.613$, $p = .070$ if corrected, $p = .035$ if uncorrected] and /pi/–/pu/, for which the test could not be conducted because the value reached the floor. A similar pattern was observed when the H–FA scores were employed as the dependent measure; the /ki/–/ku/ discrimination was significantly better than chance [$t(7) = 3.742$, $p = .021$; $= .007$ if uncorrected], but no other pair reached significance even before Holm corrections, except /pi/–/pu/, which was significantly worse than chance only before a Holm correction [$t(7) = -2.376$, $p = .148$ if corrected; $p = .049$ if uncorrected]. Again, **coarticulation insensitivity** is incompatible with the significant success with the /ki/–/ku/ pairs.

Next the scores were compared among /ei/, /eu/ and /iu/ pairs. When the dependent measure is (the arcsin square root transform of) the listener-averaged accuracy score, naturally the /ki/–/ku/ discrimination was significantly better than the /ke/–/ku/ discrimination [$t(7) = 4.584$, $p = .008$; $p = .003$ if uncorrected] and than the /ke/–/ki/ discrimination [$t(7) = 4.216$, $p = .008$; $p = .004$ if uncorrected], while no other comparison reached significance even before Holm corrections, except that the /pi/–/pu/ discrimination was significantly less successful than the /pe/–/pu/ discrimination only before a Holm correction [$t(7) = -2.393$, $p = .144$; $p = .048$ if uncorrected]. The results were similar when the d' values are employed as the dependent measure. The /ki/–/ku/ discrimination was significantly better than the /ke/–/ku/ discrimination [$t(7) = 4.42$, $p = .009$ if corrected, $p = .003$ if uncorrected] and than the /ke/–/ki/ discrimination [$t(7) = 4.279$, $p = .007$ if corrected; $p = .004$ if uncorrected], while the /ke/–/ki/ and the /ke/–/ku/ discrimination did not differ significantly even before a Holm correction. The /pe/–/pi/ and the /pe/–/pu/ discrimination did not differ significantly even before a Holm correction, but the /pi/–/pu/ discrimination was significantly worse than the /pe/–/pu/ discrimination only before a Holm correction [$t(7) = 2.613$, $p = .104$ if corrected; $p = .035$ if uncorrected] and

than the /ei/ discrimination, again only before a Holm correction [$t(7) = 1.991$, $p = .174$ if corrected, $p = .087$ if uncorrected]. The results were similar when the dependent measure was the H–FA score. The /ki/–/ku/ discrimination was significantly better than the /ke/–/ku/ discrimination [$t(7) = 5.463$, $p = .003$; $p = .001$ if uncorrected] and than the /ke/–/ki/ discrimination [$t(7) = 4.410$, $p = .006$; $p = .003$ if uncorrected], while no other comparison reached significance even before Holm corrections, except that the /pi/–/pu/ discrimination was significantly less successful than the /pe/–/pu/ discrimination only before a Holm correction [$t(7) = -2.393$, $p = .144$; $p = .048$ if uncorrected]. Thus the only reliable results were that, among the /k/ pairs, the /iu/ discrimination was the most successful. This result is consistent with the expectation from **DB-sensitivity** that /eu/ and /ei/ discriminations should not be more successful than /iu/ discriminations, but is also consistent with **DI-sensitivity**.

Finally, the results of Experiment 4 were compared with those of Experiment 3. by conducting two-sided independent *t* tests (with Holm corrections separately for the /k/ and for the /p/ set). To help interpret the results, the discrimination accuracy rates across Experiments 3–4 are depicted in Figure 5.8 (the /k/ set stimuli) and Figure 5.9 (the /p/ set stimuli); the d' values across Experiments 3–4 are depicted in Figure 5.10 (the /k/ set stimuli) and Figure 5.11; and the H–FA scores across the two experiments are depicted in Figure 5.12 (the /k/ set stimuli) and Figure 5.13 (the /p/ set stimuli).

The ISI shortening in Experiment 4 (as opposed to Experiment 3) resulted in a significant loss of successes for /ke/–/ku/ pairs, with accuracy scores [$t(14.438) = -2.963$, $p = .030$; $p = .010$ if uncorrected], and with H–FA scores [$t(14.914) = 2.976$, $p = .028$; $p = .009$ if uncorrected], although the negative effect of ISI shortening on /ke/–/ku/ discrimination was a trend after a Holm correction when the measure is d' [$t(10.839) = 2.539$, $.p = .083$ if corrected, $p = .028$ if uncorrected]. The negative effect of the ISI shortening was significant for /pi/–/pu/ pairs only if uncorrected, with accuracy scores [$t(16) = -2.251$, $p = .116$ if corrected; $p = .039$ if uncorrected], with d' values [$t(9) = -2.388$, $p = .122$ if corrected; $p. = 041$ if uncorrected],

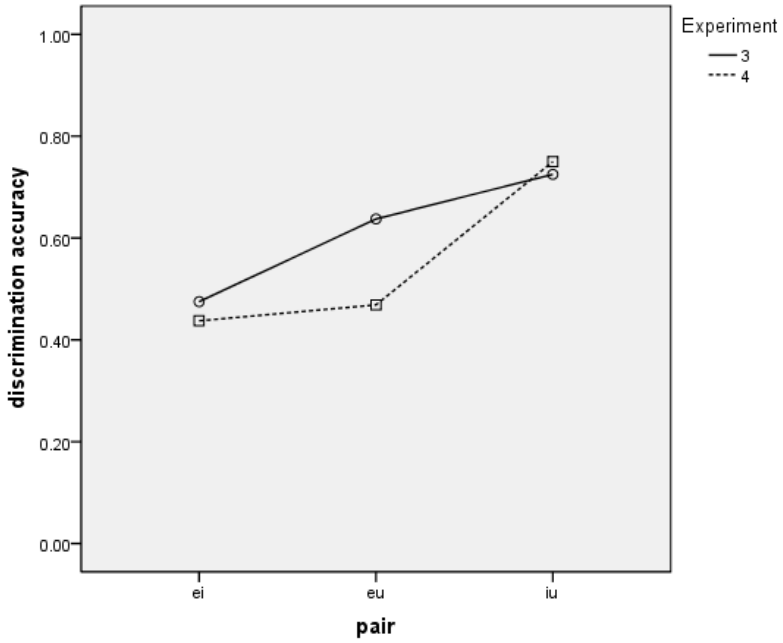


Figure 5.8: The discrimination accuracy rates with the C-set /k/ stimuli across Experiments 3–4

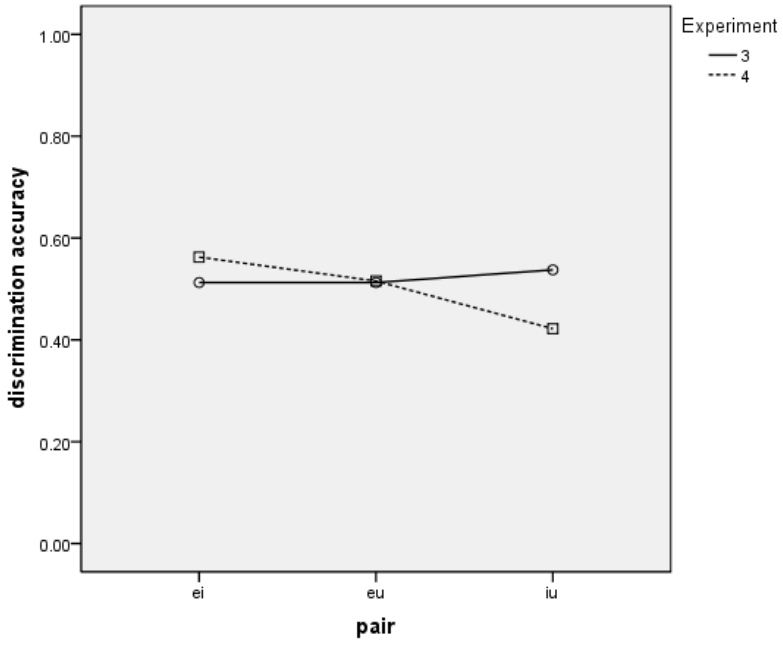


Figure 5.9: The discrimination accuracy rates with the C-set /p/ stimuli across Experiments 3–4

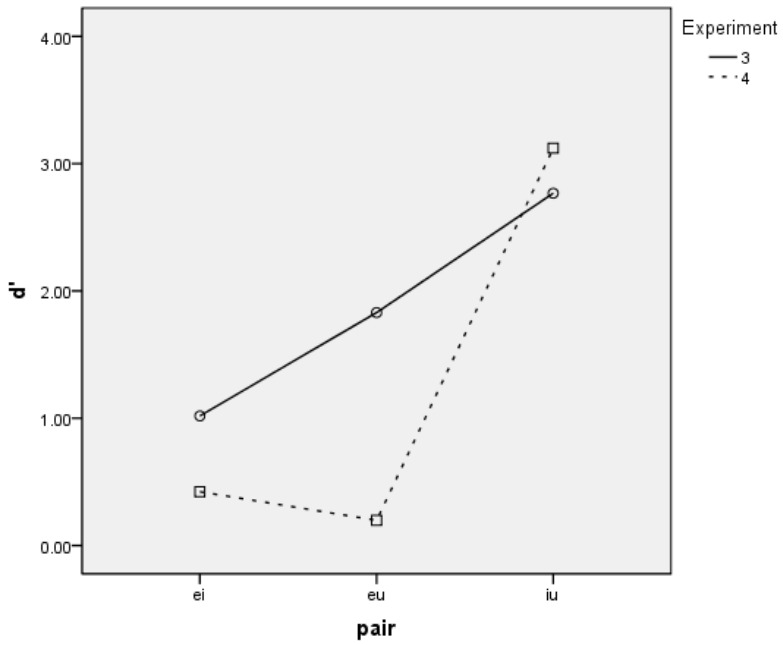


Figure 5.10: The averaged *d′* values with the C-set /k/ stimuli across Experiments 3–4

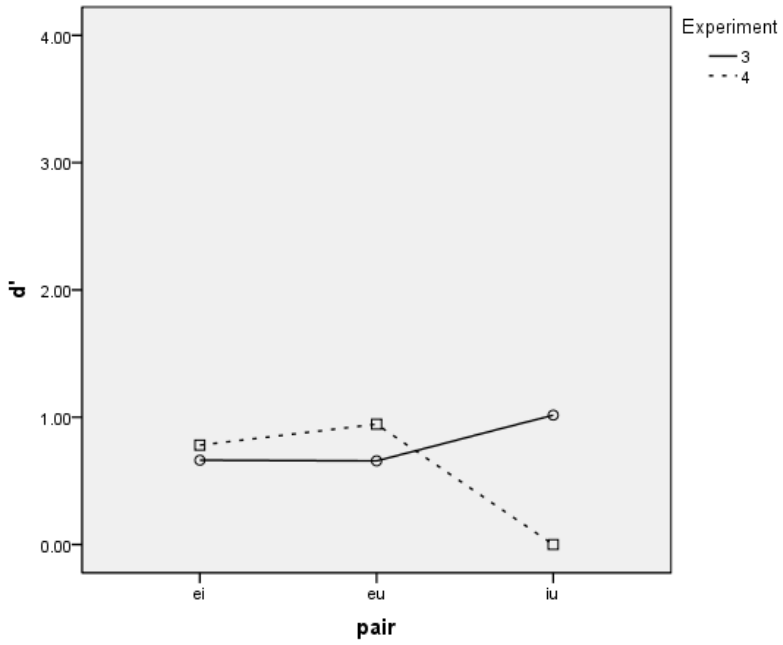


Figure 5.11: The averaged *d′* values with the C-set /p/ stimuli across Experiments 3–4

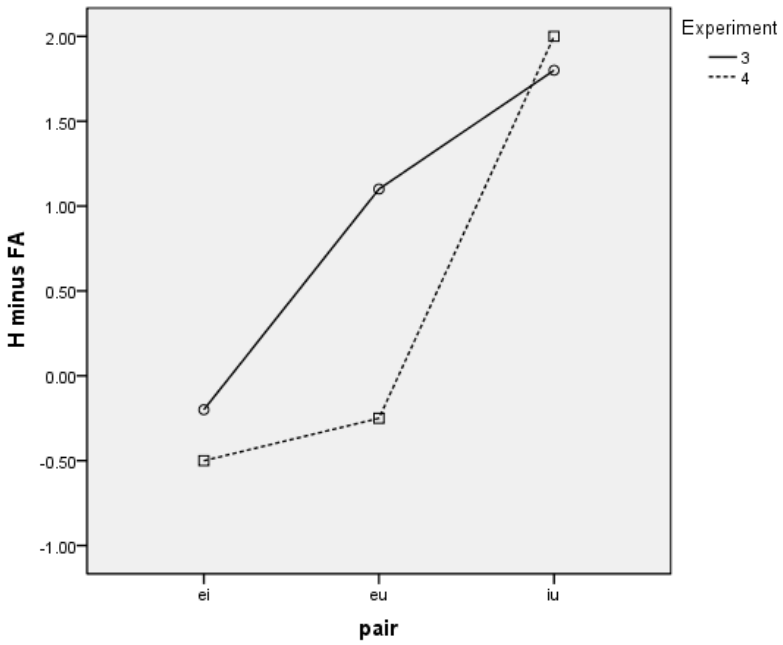


Figure 5.12: The H–FA scores with the C-set /k/ stimuli across Experiments 3–4

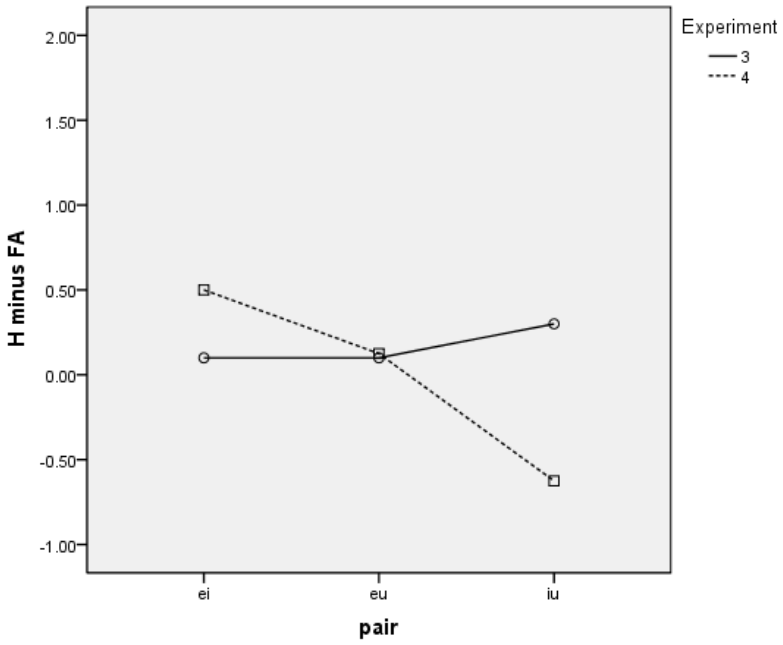


Figure 5.13: The H–FA scores with the C-set /p/ stimuli across Experiments 3–4

and with H–FA scores [$t(16) = -2.254$, $p = .116$ if corrected; $p = .039$ if uncorrected]. No other cross-experimental comparison reached significance (even before Holm corrections).

Thus the results with the accuracy rates and the H–FA scores clearly suggest that the ISI shortening has resulted in some loss of sensitivity for the /k/–/ku/ pairs, which support the prediction from **DB-sensitivity**, while the conclusion from the result with d' is not so clear (to the extent that the reduction of discrimination success remained a trend after a Holm correction). One difference between the accuracy rates and the H–FA scores on the one hand, and the d' values (as computed in this thesis) on the other, is, again, whether the magnitudes of the differences between the false alarm rates and the hit rates are taken into account. Thus the failure for the d' result to reach clear significance (in contrast to the results with accuracy rates or H–FA scores) could be attributed to the reduction of the effect size due to the analytic procedure of regarding the d' values to be zero when the false alarm rates exceeded the hit rates. According to this interpretation, the accuracy rate and H–FA score results should be seen as evidence for the reality of **DB-sensitivity**.

However, an alternative interpretation of the above results in terms of the 'unintended dimension' scenario would also be possible, according to which the ISI shortening has raised listeners' auditory sensitivity to non-phonemic inter-speaker differences with /ke/–/ke/ and /ku/–/ku/ stimuli. If this interpretation is adopted, the accuracy rate and the H–FA score results should be attributed to the raised sensitivity along the 'unintended dimension' scenario. Indeed, the mean H–FA score with /ke/–/ku/ discrimination was positive in Experiment 3 but negative in Experiment 4, which would be compatible with such an interpretation.

Those two interpretations differ with respect to whether the accuracy rate and the H–FA score results should be attributed to the change of the hit counts or the the change of the false alarm counts. In order to decide which interpretation is superior, the false alarms and the hits for /ke/–/ku/ were separately compared across the two experiments. The average false alarm count for /ke/–/ku/ indeed increased from 2.1 (SD=.88) in Experiment 3 to 2.25 (SD=1.04)

in Experiment 4, but the increase is far from being significant [$t(16) = .33$, $p = .743$]. In contrast, the average hit count decrease from 3.2 (SD=.79) in Experiment 3 to 2.0 (SD=.76) in Experiment 4 was highly significant [$t(16) = 3.266$, $p = .005$]. Thus the worse /ke/–/ku/ discrimination in Experiment 4 than in Experiment 3 should be attributed to *less* sensitivity to the /ke/–/ku/ distinction, rather than *more* sensitivity to inter-speaker variation, which argues for the **DB-sensitivity** reduction interpretation, rather than the 'increased sensitivity to inter-speaker variations' interpretation, of the accuracy rate and the H–FA score results. Thus the reasonable conclusion seems to be that the ISI shortening has indeed resulted in **DB-sensitivity**.

To summarize, the results of one-sample *t* tests in Experiment 3 and in Experiment 4 suggested that some sort of coarticulation sensitivity is real. The comparisons among pairs failed to offer firm evidence concerning whether the coarticulation sensitivity should be regarded as **DB-sensitivity** or **DI-sensitivity**, but significant reductions of successful discriminations due to ISI shortening were observed in cross-experimental comparisons, which suggest that (at least) **DB-sensitivity** is real.

5.6 General Discussion

Beckman & Shoji (1984), Ogasawara & Warner (2009) and Tsuchida (1998) observed Japanese listeners' coarticulation sensitivity and interpreted such an observation in terms of vowel devoicing. However, their results could be interpreted not only in terms of **DB-sensitivity** but also in terms of **DI-sensitivity**; the recovery of /i/ based on /i/-coarticulation could be interpreted in either terms. However, the reality of phonotactics-independent **DB-sensitivity**, as opposed to **DI-sensitivity** (to be exploited in phonotactic repair), is crucial for the validity of the defense in Chapter 3 of Dupoux et al.'s (2001) and Mazuka et al.'s (2011) argument for the sublexical nature of phonotactic sensitivity, because the argument was that Fais et al.'s (2005) results on the one hand, and Dupoux et al.'s (2001) *reksi* and *kiksi* results on the other, should

best be seen as results of phonemic categorization exhibiting **DB-sensitivity**, and hence do not directly bear on the nature of phonotactic repairs. Furthermore, for a full evaluation of lexicalist models, not only the (un)reality but also the *sublexical* (un)reality of **DB-sensitivity** has to be examined, because the crucial claim of lexicalist models would be that **DB-sensitivity** could be reduced to lexical sensitivity. Thus Experiments 1–4 aimed at examining the (un)reality of sublexical **DB-sensitivity**. The reality of **DB-sensitivity** would validate the defense in Chapter 3 of Dupoux et al.'s (2001) and Mazuka et al.'s (201) argument for the sublexical nature of phonotactic sensitivity and hence against the lexicalist reduction of phonotactic sensitivity, and the sublexical reality of **DB-sensitivity** would argue against the lexicalist reduction of coarticulation sensitivity.

In Experiment 1 it was observed that [k] bursts coarticulated with either /i/ or /e/ induce /i/ identifications. Such observations argue against **coarticulation insensitivity**, because coarticulation does affect the identity of the perceived vowel. The /e/ result further argues for **DB-sensitivity**, as opposed to **DI-sensitivity**, because it suggests that /e/-coarticulation is perceived not as /e/-coarticulation but rather as front coarticulation, to be phonemically categorized as /i/. Furthermore, this evidence for **DB-sensitivity** was obtained with [ek^{V}ma] stimuli, where the /e/-coarticulated consonant was followed by a *voiced* consonant [m]. Given that vowels do not devoice before a voiced consonant *in production*, such results constitute evidence that it is the [C^{V}] portion, not the whole [eC^{V}ma], that gives rise to devoicing-based /Ci/ categorization. This justifies the interpretation suggested in Chapter 3 of Ogasawara & Warner's (2009) observation that 'post-voiced' stimuli (i.e., those stimuli in which devoiced vowels are preceded by voiceless consonants but followed by voiced consonants) behave like 'voiceless environment' stimuli in /i/-monitoring but like 'voiced environment' stimuli in lexical decision; for a sublexical task such as /i/-monitoring, the following voicing consonant does not matter, but for a lexical task such as lexical decision, the production patterns are mirrored in perception. Thus, the sublexical reality of **DB-sensitivity** is supported by the results of Experiment 1, which ar-

gues against lexicalist models. Note that the suggested interpretation of Ogasawara & Warner's results admits both the sublexical and the lexical reality of **DB-sensitivity**; lexicalist models are argued against in the sense that (phonotactic and) coarticulation sensitivity cannot be *reduced* to lexical sensitivity.

In Experiment 2, the burst durations were shortened so that the velar and bilabial burst stimuli would be equal in durations. Velar bursts coarticulated with either /i/ or /e/ still induced /i/ identifications. Presumably, burst shortening has degraded the coarticulation cues. The fact that the effects expected from **DB-sensitivity** were still observed gives a further support to **DB-sensitivity**, as well as its sublexical reality. On the other hand, the /i/ identification rate with the C-set /ki/ stimuli was significantly below the /i/ identification rate with the V-set /ki/ stimuli in Experiment 2, in contrast to Experiment 1, in which they did not differ. This observation supports the interpretation, suggested in Chapter 3, of the contradictory results by Beckman & Shoji (1984), Cutler et al. (2009), and Ogasawara & Warner's (2009), with respect to whether vowel devoicing inhibits /CV/ perception. If **DB-sensitivity** leads Japanese listeners to phonemically categorize a consonant [C] with enough front coarticulation cues as /Ci/, the /Ci/ perception should be a function of the amount or quality of front coarticulation. The significant tendency for /ki/ perception as opposed to the lack of such a tendency for /pi/ perception in Experiment 1 bears out such a prediction. In addition, the results of Experiment 2, as compared to those in Experiment 1, constitute a within-consonant confirmation of that prediction. The lack of difference between the difficulty of /i/ identification from the C- and from the V-set /ki/ stimuli in Experiment 1 replicates Ogasawara & Warner's /i/ monitoring results, while the significant difficulty of /ki/ perception from the C-set stimuli than from the V-set stimuli in Experiment 2 replicates Beckman & Shoji's (and Cutler et al.'s 2009) results. Thus the results of Experiments 1–2 suggest that whether vowel devoicing makes /CV/ perception more difficult or not depends not only on the experimental task (Ogasawara & Warner, 2009) but also on the amount or strength of coarticulation cues. This conclusion accounts for the

conflict among the results by Beckman & Shoji, Cutler et al., and Ogasawara & Warner.

The sublexical reality of **DB-sensitivity** was further supported by the results of Experiments 3–4. For an AX discrimination experiment, (i) **coarticulation insensitivity** predicts uniform failure, (ii) **DB-sensitivity** predicts that /ei/ or /eu/ discriminations should not be more successful than /iu/ discriminations, and (iii) **DI-sensitivity** makes no specific predictions concerning which pairs should be discriminated better than which pairs.[15] The /ki/–/ku/ discrimination success was significant in both Experiments 3–4, and the /ke/–/ku/ discrimination success was significant in Experiment 3. Such results argue against the **coarticulation insensitivity** hypothesis. Although the pairwise comparisons among the pairs failed to give a clear indication of which kind of coarticulation sensitivity is real, the observations of *less* successful /ke/–/ku/ discriminations with a shorter ISI (250 ms) in Experiment 4 than with a longer ISI (1,500 ms) in Experiment 3 could be interpreted only in terms of the reduced effects of **DB-sensitivity**, which is clearly phonological.

Note that, with the [ek^{V}ma] stimuli, only the *sublexical* kind of **DB-sensitivity** could exert its effect. Thus the results of the comparisons between Experiment 3 and Experiment 4 support the sublexical reality of **DB-sensitivity**.

Thus the results of Experiments 1–4 support the following conclusions:

- **DB-sensitivity** is sublexically real.
- /CV/ perception may or may not be more difficult from vowel-devoiced stimuli than from non-devoiced stimuli depending on the amount or quality of coarticulation cues in the consonants.

Such findings explain the /i/ ‘epenthesis’ after /k/ in loanwords; it is due to devoicing-based phonemic categorization of fronted voiceless velar bursts. Furthermore, they justify the ac-

[15] Possibly, if /e/ and /i/ are acoustically more similar than /i/ and /u/ or than /e/ and /u/, **DI-sensitivity** would predict that /ei/ discrimination should be less successful than /iu/ or /eu/ discriminations. However, as noted already, this thesis has no empirical justification to offer for this assumption, no specific prediction is drawn from **DI-sensitivity**. However, if this assumption is adopted, the observed results were compatible not only with the prediction from **DB-sensitivity** but also with the resulting prediction from **DI-sensitivity**.

counts in Chapter 3 of the results of Fais et al.'s (2005) rating study on the one hand, and the /i/ transcriptions for 'reksi' and 'riksi' stimuli observed by Dupoux et al. (2001) on the other; both observations should naturally be attributed to **DB-sensitivity**, which induces /ki/ perception from the velar portions of the stimuli. Under this devoicing-based accounts, those observations are reflections of devoicing-based phonemic categorization, not of phonotactic repair, and hence not relevant to the issue of whether phonotactic sensitivity could or could not be reduced to lexical sensitivity. Thus the confirmed reality of **DB-sensitivity** defends the claim of the sublexical nature of phonotactic sensitivity by Dupoux et al.'s (2001) and Mazuka et al. (2011) against Fais et al.'s (2005) rating results and Dupoux et al.'s (2001) *reksi* and *riksi* results (by expelling them from the realm of phonotactic repairs). Furthermore, the confirmed *sublexical* reality of **DB-sensitivity** also argues against a lexicalist reduction of coarticulation sensitivity. With both phonotactic sensitivity and coarticulation sensitivity being shown not to be reducible to lexical sensitivity, lexicalist models should be rejected.

With this conclusion, the remaining candidates are one- and two-step models. Speaking more specifically, the discussion in Chapter 3 concluded that, if one-step models should be adopted at all, it should be the **slot filling** version, rather than the **suprasegmental matching** version, that should be adopted (because Matthews & Brown's 2004 results support the idea that listeners do have access to within-syllable elements). Thus the remaining candidates are the **slot filling** version of one-step models on the one hand, and two-step models on the other. Both kinds of models assume phonotactic parsing, in addition to phonemic parsing (categorization of the speech input as a sequence of phonemes). They differ with respect to whether the two parsing operations are sequential (two-step models) or parallel (the **slot filling** version of one-step models).

However, the support for the perceptual reality of **DB-sensitivity** also suggests that Dupoux et al.'s (2011) argument for one-step models should be questioned: Dupoux et al. (2011) compared two- and one-step models. According to two-step models, individual phonemes are first

perceived, and the resulting phoneme sequences are repaired according to phonotactics; thus phonotactic repair should be insensitive to subphonemic details including coarticulation traces. In contrast, according to one-step models, phonemic perception and phonotactic repair should both operate on pre-categorized speech inputs and hence phonotactic repair could be sensitive to subphonemic details including coarticulation traces. Based upon Japanese listeners' coarticulation-sensitive /i/ 'epenthesis', Dupoux et al. argued that one-step models are superior to two-step models. However, if **DB-sensitivity** is real, coarticulation-sensitive /Ci/ phonemic categorization of [C^i] is sometimes expected, in which case Dupoux et al.'s (2011) observations could be interpreted in terms of the coarticulation sensitivity in the 'first step' phonemic categorization stage (rather than in the 'second step' phonotactic repair stage) under two-step models.

Indeed, the perceptual reality of **DB-sensitivity** does not necessarily exclude the reality of **DI-sensitivity**. For example, in vowel identification based on [k^e], the former leads us to expect a tendency for /i/ identification while the latter would lead us to expect a tendency for /e/ identification. The observed tendency for /i/ identification could be interpreted either by assuming (i) that **DB-sensitivity** is the only coarticulation sensitivity, or (ii) that both kinds of sensitivity are real, but the effect of **DB-sensitivity** (a bias toward /i/) was larger and won the effect of **DI-sensitivity** (a bias toward /e/). Similarly, although no firm evidence for **DI-sensitivity** was obtained in Experiments 3–4 either, the lack of evidence for something does not establish the validity of its denial.

Thus, while the results of Experiments 1–4 argue for the reality of **DB-sensitivity**, they do not necessarily argue against the reality of **DI-sensitivity**.

Translated into our terms, coarticulation sensitivity that could constitute evidence against two-step models is **DI-sensitivity**. Thus, in order to determine whether one-step models should indeed be favored over two-step models, as Dupoux et al. (2011) claim, the reality of coarticulation sensitivity that could only be interpreted as **DI-sensitivity**, rather than as **DB-sensitivity**,

should be examined. The next set of experiments (Experiments 5–8) are meant to be such examinations.

Chapter 6

Experiments 5–8

The results of Experiments 1–4 suggest that front coloring within voiceless velar bursts are perceived as /ki/, whether the fronting was due to coarticulation with /i/ or with /e/, which could only be interpreted as a result of **DB-sensitivity**. However, if such /i/ perception is a result of **DB-sensitivity**, it should *not* be seen as an instance of epenthesis, because /ki/ perception would then be simply a result of phonemic categorization of a fronted voiceless velar burst as /ki/, which would not violate phonotactics and should not induce phonotactic repair. This in turn casts some doubt on Dupoux et al.'s (2011) argument against two-step models. Their argument hinges on the observation of Japanese listeners' tendency for /i/ perception from consonants coarticulated with [i] ([C^i]'s), which they interpreted as evidence for coarticulation-sensitive phonotactic repair. However, the [C^i]'s in their stimuli consisted of voiceless and voiced ones. The perceptual reality of **DB-sensitivity** suggests that the tendency for /Ci/ perception (as opposed to /Cu/ perception) should be attributed to coarticulation-sensitive phonemic categorization when the [C^i]'s are voiceless (and have enough coarticulation cues). Thus an observation of a tendency for /Ci/ perception from [C^i]'s, when voiceless and voiced consonant stimuli are mixed, could be interpreted as a combination of the tendency for /Ci/ perception with voiceless consonant stimuli on the one hand, and the lack of such a tendency with voiced consonant

stimuli on the other. While the tendency for /Ci/ perception with voiceless stimuli could well be attributed to phonotactics-independent devoicing-based phonemic categorization, the lack of such a tendency with voiced stimuli would be compatible with coarticulation-*in*sensitive phonotactic repair. In order to eliminate such an interpretation so as to argue for coarticulation sensitivity in phonotactic repair, a similar tendency for /i/ perception has to be shown when the [C^i]'s are voiced; only with voiced [C^i]'s could the tendency for /i/ perception be interpreted as coarticulation-sensitive *phonotactic repair* rather than *phonemic categorization*. Unfortunately, they mixed voiceless and voiced results; no separate analyses for voiceless and voiced cases are reported.

Thus the next set of experiments (Experiments 5–8), reported in this chapter, examined whether Japanese listeners exhibit coarticulation sensitivity in their vowel perception from *voiced* consonants. Given that devoicing-based phonemic categorization is the categorization of a *voiceless* consonant stimulus as /Ci/ (or /Cu/), coarticulation-sensitive vowel perception from *voiced* consonant stimuli could not be interpreted as results of such phonemic categorization. Thus, if coarticulation sensitivity is observed with voiced consonants, that would support Dupoux et al.'s (2011) argument for the reality of coarticulation-sensitive *phonotactic repair*, which would be incompatible with two-step models. Put in the terms employed so far, coarticulation sensitivity in the voiced cases is **DI-sensitivity**. In Experiments 1–4, evidence for **DI-sensitivity**, as opposed to **DB-sensitivity**, failed to be obtained, but, as already noted, such a result could either be interpreted either in terms of the absence of **DI-sensitivity**, or in terms of competitions between (larger effects of) **DI-sensitivity** and (smaller effects of) **DB-sensitivity**; thus the results of Experiments 1–4 do not necessarily argue against the reality of **DI-sensitivity**.

Before we proceed, some notes concerning voiced stops in Japanese are necessary. The following **production presuppositions** (Vance, 1987; 2008) have been assumed so far:

- Both /i/ and /e/ are fronted, but /i/ is more fronted than /e/.
- Velars exhibit rich place variations and hence velar bursts carry rich coarticulation cues.
- Bilabials do not exhibit large variations due to place of articulation[1] and hence bilabial bursts carry few or weak coarticulation cues.

However, one *additional* observation concerning Japanese production should be kept in mind; as noted in Chapter 2, voiced stop phonemes in Japanese are often subject to lenition, and their physical realizations are often voiced fricatives, rather than voiced stops with clear closures. This observation suggests the possibility that, even when they are realized as stops, rather than fricatives, the closures for /g/ and /b/ are rather weak, which in turn suggests the possibility that coarticulation traces are rather poor in voiced bursts produced by Japanese speakers, with bursts themselves being very weak. Thus the possibility should be kept in mind that coarticulation sensitivity with voiced consonants could fail to be observed with stimuli produced by Japanese speakers not because Japanese listeners are coarticulation-insensitive with voiced consonants but rather because the stimuli do not contain enough coarticulation cues. (Recall the comparison between the results in Experiment 1 and in Experiment 2; indeed degenerated bursts result in smaller effects of coarticulation sensitivity.)

Experiment 5 is an identification experiment, with stimuli produced by Japanese speakers; Experiments 6–7 are AX discrimination experiments, again with stimuli produced by Japanese speakers; Experiment 8 is another identification experiment, but with stimuli produced by an English speaker. Because Experiments 5–8 are meant to examine **DI-sensitivity**, the coarticulated vowels do not necessarily have to be limited to /e/ or /i/ (or /u/), as is done in the analyses for the results of Experiments 5–8; it is not relevant that non-high vowels would not easily devoice. However, in the analyses of the results of Experiments 5–8, the coarticulated vowels are again confined to /e/, /i/ and /u/, for the following reasons. For one thing, the results of

[1]The source of the turbulence noise is behind the bilabial constriction and consequently the noise in the bursts won't be affected by differences in tongue position.

Experiments 1–4 have already suggested that, even if it is real, the effects of **DI-sensitivity** are likely to be rather weak; with weak effects, multiple comparisons with five vowels are likely to lack enough statistical power, so the number of vowels to be analyzed should best be limited. For another, we are specifically interested in determining whether Japanese listeners' tendency for /i/ perception observed by Dupoux et al. (2011) would persist even when the consonants are limited to voiced ones. Finally, voiceless cases were examined with those three vowels, so analyzing those three vowels would enable us to compare voiceless and voiced cases.

6.1 Experiment 5

Experiment 5 imitates Experiment 2 (vowel identification), except that the main targets are those cases in which the crucial consonants are voiced.

If Japanese listeners' success of /i/ identification with consonants coarticulated with /i/ should be seen only in terms of **DB-sensitivity**, a straightforward expectation is that they should fail to identify /i/ with voiced stops, resulting in the default /u/ epenthesis. In contrast, their successful identification of /i/ with voiced stops coarticulated with /i/ would suggest their **DI-sensitivity** to /i/-coarticulation, in conformity to Dupoux et al.'s (2011) claim.

6.1.1 Method

Stimuli

Speakers 2–3 (one male and one female) of the previous experiments produced three-mora nonwords of the form /eCVma/, where C is either /b/ or /g/, V is one of the Japanese five vowels /a, i, u, e, o/, and an accent is placed on the first mora. Their utterances were recorded in a sound-attenuated room with Marantz Solid State Recorder PMD650 with the sampling frequency of 44,100 Hz. Again, following Monahan et al. (2009), two sets of stimuli were produced from the original utterances: **the C-set** stimuli were created by deleting the bursts

and the V's (the voiced portions between the bursts and the onsets of the following nasal closures), except the initial 10 ms portions of the bursts of the C's, and **the V-set** stimuli were created by removing three pitch periods arbitrarily chosen from the midst of the V's. The RMS average intensity of each of the stimuli was rescaled to the save level before presentation. Two repetitions of the sets, together with two repetitions of the stimuli in the the voiceless versions of the **C-** and the **V-set** employed in Experiment 2 (by Speakers 1–2, for a technical reason), were intermixed in a computer-generated random order and presented to the listeners through circumaural closed-back headphones (SONY MDR-ZX700) in a sound-attenuated room. The waveforms and spectrograms of example C-set stimuli are shown in Figure 6.1.

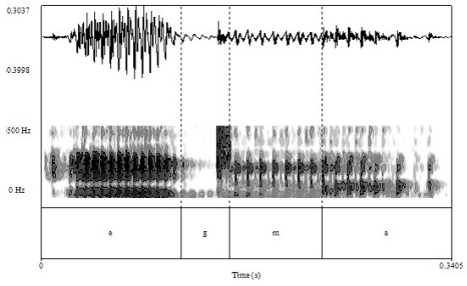
Participants

Eight native Japanese listeners with no known hearing disability participated in the experiment for course credit at undergraduate or postgraduate programs at Hosei University, Tokyo (four males and four females; mean age = 21, S.D. = 1.31). None of them had participated in the previous experiments.

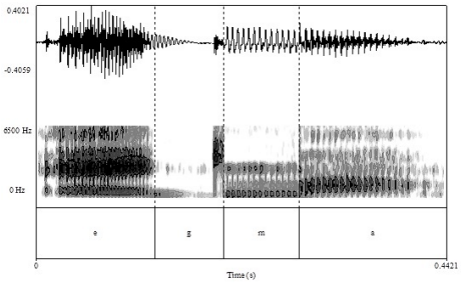
Procedure

The experiment was controlled by E-prime run on a Windows XP machine (Panasonic Let's note CF-S9KYKBDU). The specific instructions were the same as in Experiments 1–2; in short, the participants were asked to identify the medial vowel by pressing one of the five response keys on a response box.

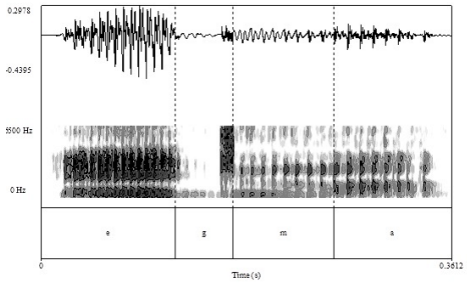
Listeners were first required to familiarize themselves with the task by going through a practice session (where the stimuli were the experimenter's productions of /edoma/, /ẹdima/, /enuma/, /esama/, and /etema/with no significant editing), and only after scoring more than 80 % accuracy were they allowed to proceed to the main experimental session. Each stimuli (a combination of a consonant, a vowel, and a speaker) was played twice, intermixed with the



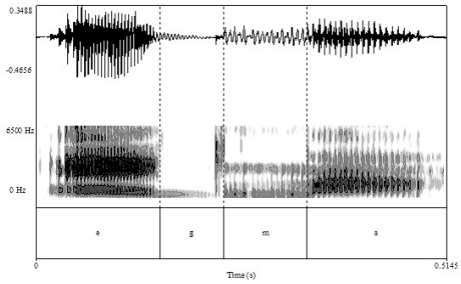
the male C-set /egema/



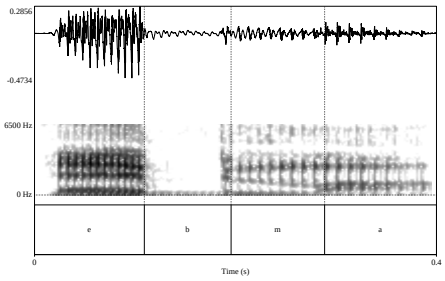
the female C-set /egema/



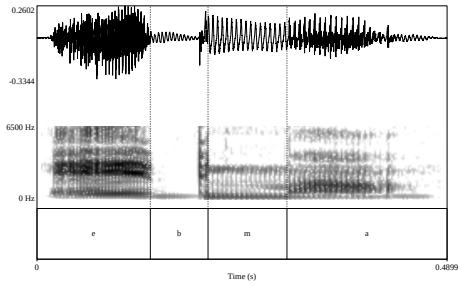
the male C-set /egima/



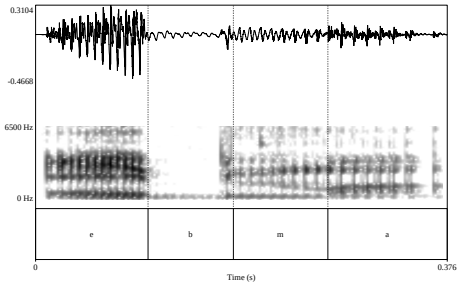
the female C-set /egima/



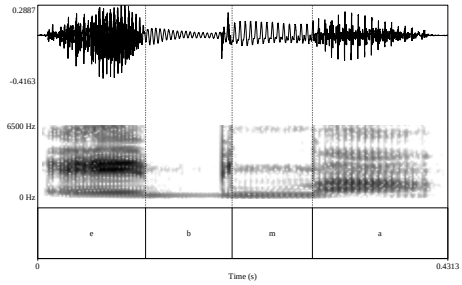
the male C-set /ebema/



the female C-set /ebema/

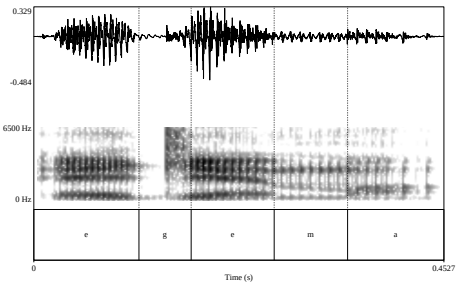


the male C-set /ebima/

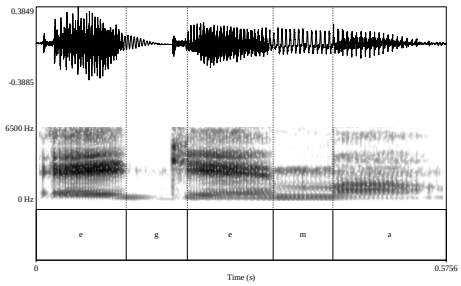


the female C-set /ebima/

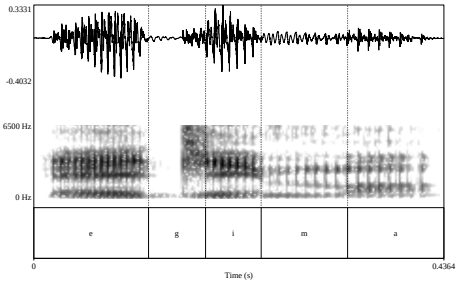
Figure 6.1: Example waveforms and spectrograms of the stimuli for Experiment 5; the stimuli by the male speaker on the left, and the stimuli by the female speaker are on the right.



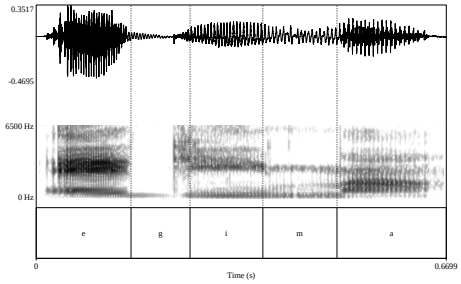
the male V-set /egema/



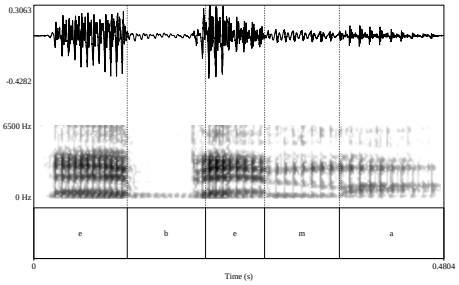
the female V-set /egema/



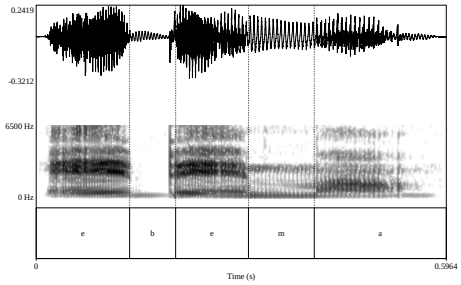
the male V-set /egima/



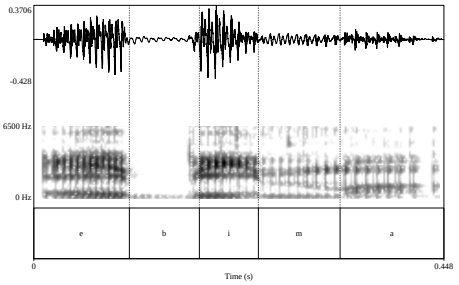
the female V-set /egima/



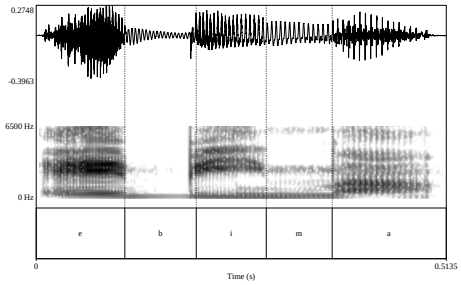
the male V-set /ebema/



the female V-set /ebema/



the male V-set /ebima/



the female V-set /ebima/

other stimuli.

Statistical Analyses

Only the results with those C-set stimuli in which the medial vowels are either /e/, /i/ or /u/ are employed in the analyses.; all the other stimuli are treated as fillers. The primary target of the analyses are the results from the voiced C-set stimuli.; the results from the voiceless C-set are employed in supplementary analyses.

First, the patterns of the responses (i) from the voiced C-set stimuli in Experiment 5, (ii) from the voiceless C-set stimuli in Experiment 2, and (iii) from the voiceless C-set stimuli in Experiment 5, are compared to each other by means of χ-squared tests and/or Fisher's exact tests, separately for each place-vowel combination, so that we can see how listeners' responses do or do not differ across voiceless and voiced stimuli on the one hand (i)–(ii), and how (in)consistent listeners' responses to voiceless stimuli were across the two experiments (iii).[2]

Next, the listener-averaged rates of the relevant vowel categories are examined. With voiceless stimuli, we are primarily interested in the rates of /i/ responses; **DB-sensitivity** should induce /i/ perception for both /i/- and /e/-coarticulation. However, with voiced stimuli, we are interested in whether the original vowels would be restored based on their coarticulation traces, because **DI-sensitivity** is something that should help restorations of /i/ from /i/-coarticulation and of /e/ from /e/-coarticulation. Thus the primary dependent measures (with arcsin square root transformations) are the listener-averaged /i/ response rates for voiceless stimuli and the listener-averaged 'original vowel restoration rates'. The examinations of the results from the voiceless stimuli serve as a double check of the results of Experiment 2. With voiced stimuli, the hypotheses to be compared are **coarticulation insensitivity** and **DI-sensitivity**.

[2]There will be many expected 0's, where no response for a given vowel category was observed. In such a case, the relevant vowel columns are removed when χ-squared tests are conducted. On the other hand, Fisher's exact tests for 2×5 tables were conducted with the implementation in the R language (version 3.0.1).

As with Experiments 1–2, the primary measures are submitted to (two-sided) one-sample *t* tests, separately for /pe/, /pi/, /ke/ and /ki/ on the one hand (the /i/ response rates), and for /be/, /bi/, /ge/ and /gi/ on the other (the 'original vowel restoration rates'), with Holm corrections of the significance values for each place-voicing combination; restoration rates for /bu/ and /gu/ are not informative with respect to the evaluation of the hypotheses and hence not submitted to *t* tests (so that the inflation of significance values in multiple comparisons would be reduced). The **coarticulation insensitivity** hypothesis predicts that the original vowel restoration rates should be at the floor, whereas the **DI-sensitivity** hypothesis predicts, for each of /be/, /bi/, /ge/ and /ga/, the *possibility* that the restoration rate is better than chance.

6.1.2 Results and Discussion

The results were divided according to the following four variables:

CV: whether the stimulus was from the C- or the V-set.

voicing: whether the medial consonant is voiced or voiceless.

consonant: whether the medial consonant was /g/ or /b/.

vowel: the identity of the medial vowel (/i/, /e/, /u/).

The numbers of the responses for each category are illustrated in Tables 6.1–6.4. To help the reader, the tables for the numbers of the responses for each category for the C-set stimuli in Experiment 2 are repeated here as Tables 6.5–6.6.

Not only with /b/ and /p/ but also with /g/ stimuli in Experiment 5 (Tables 6.1–6.3), listeners' responses were mostly /u/, the default epenthetic vowel, which suggests that coarticulation sensitivity of whatever kind failed to be exerted.

On the other hand, in the case of /k/ stimuli, listeners' responses in Experiment 5 (Table 6.4) were rather unexpected. First, the number of /i/ responses to /ki/ stimuli (28) is even larger than the number of /u/ responses to /ku/ stimuli (24), which is rather unexpected from all the

Table 6.1: The number of responses for the C-set /b/ stimuli in Experiment 5.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	3	0	1	28	32
	/i/	1	1	9	1	20	32
	/u/	0	0	0	3	29	32
Total		1	4	9	5	77	96

Table 6.2: The number of responses for the C-set /g/ stimuli in Experiment 5.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	1	2	8	21	32
	/i/	1	0	2	5	24	32
	/u/	0	1	0	0	31	32
Total		1	2	4	13	76	96

hypotheses. While only coarticulation-sensitive identifications could lead to /i/ responses to /ki/ stimuli, both coarticulation-sensitive and coarticulation-insensitive identifications should lead to /u/ response to /ku/ stimuli, and hence a natural expectation would be that, whether listeners are coarticulation sensitive or not, the number of /i/ responses to /ki/ stimuli should not be larger than the number of /u/ response to /ku/ stimuli, particularly with the stimuli in which the bursts are artificially shortened to 10 ms. This expectation was indeed met in Experiment 2 (Table 6.6) but was betrayed in Experiment 5 (Table 6.4).[3]

Next, the number of /i/ responses to /ke/ stimuli (9) lost the number of /u/ responses (14). Indeed, it is assumed that /e/-coarticulation is weaker than /i/-coarticulation (the **production presuppositions**), so such a result could be interpreted as suggesting either that **DB-sensitivity** is not real or that it is real but its effects failed to be manifested with the artificial shortening of

[3]However, we cannot declare with confidence immediately at this point that the results of Experiment 5 should be questioned. In Experiment 1 (where the bursts were not artificially shortened), there were 80 /i/ responses to /ki/ stimuli and 79 /u/ responses to /ku/ stimuli. Counting /i/ responses to /ki/ stimuli on the one hand, and /u/ responses to /ku/ stimuli on the other, as 'expected'responses, and conflating all the other responses as 'unexpected responses', the associations between 'expected' vs. 'unexpected' responses and /ki/ vs. /ku/ stimuli were examined through Fisher's exact tests. In the case of Experiment 1, the result was far from significant [$p = 1$]; in the case of Experiment 2, the conceptually natural larger number of 'expected' responses to /ku/ stimuli than to /ki/ stimuli was highly significant [$p < .001$]; in the case of Experiment 5, the conceptually unnatural larger number of 'expected' responses to /ki/ stimuli than to /ku/ stimuli failed to reach significance [$p = .337$].

Table 6.3: The number of responses for the C-set /p/ stimuli in Experiment 5.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	1	1	3	1	26	32
	/i/	0	2	7	1	22	32
	/u/	0	2	0	1	29	32
Total		1	5	10	3	77	96

Table 6.4: The number of responses for the C-set /k/ stimuli in Experiment 5.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	1	4	9	4	14	32
	/i/	0	3	28	0	1	32
	/u/	4	2	2	0	24	32
Total		5	9	39	4	39	96

Table 6.5: The numbers of responses for the C-set /p/ stimuli in Experiment 2.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	1	2	0	45	48
	/i/	0	0	6	0	42	48
	/u/	0	1	0	0	47	48
Total		0	2	8	0	134	144

Table 6.6: The numbers of responses for the C-set /k/ stimuli in Experiment 2.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	0	28	2	18	48
	/i/	0	2	32	2	12	48
	/u/	0	0	3	0	45	48
Total		0	2	63	4	75	144

the bursts to 10 ms. Whichever is the case, the result of Experiment 5 contrast with the results of Experiment 2 with respect to whether /e/ responses outnumber /u/ responses to /ke/ stimuli.

However, the comparison between the /i/ responses to /ki/ stimuli and the /u/ response to /ku/ stimuli suggests the *possibility* that the listeners in Experiment 5 were not paying close attention to the stimuli.[4] In fact, a casual look of the response patterns in the tables gives the impression that listeners' responses are relatively spread across response categories in a way expected under no hypothesis (e.g., nobody would expect /a/ responses for /ku/ stimuli). Thus the results from Experiment 5 should be seen with some caution.

With the words of caution in mind, let us proceed to statistical tests.

First, the pattern of responses for voiced stimuli in Experiment 5 (Tables 6.1–6.2) are compared to the patterns of responses for voiceless stimuli in Experiment 2 (Tables 6.5–6.6), separately for each place-vowel combination, with χ-squared tests and Fisher's exact tests. No significant difference was found for /be/ vs. /pe/ on the one hand, and for /gu/ vs. /ku/ on the other.[5]. Non-significant trends were found for /bu/ vs. /pu/ [$\chi^2(2) = 5.2741$, $p = .071$ (χ-squared); $p = .060$ (Fisher's)]. For /bi/ vs. /pi/, the result of a χ-squared test slightly failed to be significant [$\chi^2(4) = 8.5484$, $p = .0073$], but the result of Fisher's exact test was significant [$p = .0257$]. However, note that the significance values are not corrected for multiple comparisons; if corrected (for example, with a Holm correction separately for each place or for each vowel), this will not count as significant [$p = .514$]. In contrast, clearly significant differences were observed for /gi/ vs. /ki/ [$\chi^2(4) = 32.8711$, $p < .001$ (χ-squared); $p < .001$ (Fisher's)] and for /ge/ vs. /ke/ [$\chi^2(3) = 25.1709$, $p < .001$ (χ-squared); $p < .001$ (Fisher's)]; they are highly significant even when conservative Bonferroni corrections are applied to the whole comparisons. Thus it is between /gi/ vs/ /ki/ on the one hand, and between /ge/ and /ke/

[4]Given the result in footnote 3 immediately above, we could only raise it as a *possibility*.

[5]For /be/ vs. /pe/, $\chi^2(3) = 4.9572$, $p = .175$ (χ-squared); $p = .148$ (Fisher's); since there was no /a/ responses, the /a/ columns are totally removed from the χ-squared test, hence the three degree of freedom, rather than four. (The degrees of freedom in χ-squared tests are quite often something other than four. However, when Fisher's exact tests are conducted, the 'no response' vowel columns are not removed.) For /gu/ vs. /ku/, $\chi^2(2) = 3.5197$, $p = .1721$ (χ-squared); $p = .1479$ (Fisher's).

on the other, that clear differences are observed.

The result of an examination of the standardized residuals in the χ-squared tests accorded with the intuition that this is because the default coarticulation-insensitive /u/ response was the majority for the /gi/ and /ge/ stimuli whereas the coarticulation-sensitive /i/ response was the majority for the /ki/ and /ke/ stimuli; the standardized residual was the largest with the /u/ response for /gi/ (4.404) and for /ge/ (2.466), but with the /i/ response for /ki/ (5.355) and for /ke/ (4.714).

When the response patterns for voiced stimuli in Experiment 5 (Tables 6.1)–6.2 are compared to the responses patterns for voiceless stimuli in the same Experiment 5 (Tables 6.3–6.4), the results were somewhat different. For bilabials, the results of χ-squared tests and Fisher's exact tests were unanimously non-significant.[6] For /gu/ vs. /ku/, the χ-squared test result was slightly non-significant [$\chi^2(3) = 7.2242$, $p = .065$], while the result of Fisher's exact test was significant [$p = .027$]. For /ge/ vs. /ke/, both the χ-squared test result and Fisher's exact test result were significant [$\chi^2(4) = 9.9879$, $p = .041$ (χ-squared); $p = .028$ (Fisher's)], but note that those significance values are all non-corrected for multiple comparisons. In contrast, again, the χ-squared tests and Fisher's exact test result were both highly significant for /gi/ vs. /ki/ [$\chi^2(4) = 52.6933$, $p < .001$ (χ-squared); $p < .001$ (Fisher's)]; whatever corrections for multiple comparisons would end up in significance.

Indeed, an examination of the standardized residuals in the χ-squared tests reveal that, in the /gi/ vs. /ki/ comparisons, the standardized residual was the largest with the /u/ responses for the /gi/ stimuli (5.893) and with the /i/ responses for the /ki/ stimuli (6.513), suggesting that the majority responses were /u/ for /gi/ and /i/ for /ki/. Furthermore, if the difference between /ge/ and /ke/ should count as significant at all, the result of an examination of the standardized residuals was similar to the comparison between the responses for the voiced

[6]The results of χ-squared tests were: for /e/, $\chi^2(4) = 5.0741$, $p = .260$; for /i/, $\chi^2(4) = 1.6786$, $p = .797$; for /u/, $\chi^2(2) = 3.00$, $p = .359$. The results of Fisher's exact tests were: for /e/, $p = .260$; for /i/, $p = .871$; for /u/, $p = .359$.

stimuli in Experiment 5 and for the voiceless stimuli in Experiment 2; the standardized residual was the largest with /u/ for /ge/ (1.758) but with /i/ for /ke/ (2.319).[7] However, the result of the voiced vs. voiceless comparison within Experiment 5 differed from the above result of the voiced vs. voiceless comparison across Experiment 5 and Experiment 2, in that the difference between /ge/ and /ke/ is not clearly significant. This suggests that the patterns of responses to voiceless stimuli differed across Experiment 5 and Experiment 2. In order to examine such a suspicion, responses to voiceless stimuli were compared across Experiment 5 and Experiment 2.

No significant difference was found in the responses to bilabial stimuli.[8] However, the cross-experimental difference was significant for /ke/ [$\chi^2(4) = 13.2536$, $p = .010$ (χ-squared); $p = .004$ (Fisher's)], for /ku/ [$\chi^2(3) = 8.7826$, $p = .021$ (χ-squared); $p = .009$ (Fisher's)], and for /ki/ [$\chi^2(3) = 8.9316$, $p = .030$ (χ-squared); $p = .014$ (Fisher's)]. If Holm corrections are applied to those velar comparisons, they would all end up being significant.[9]

Thus listeners' responses to voiceless stimuli do seem to have differed across Experiment 5 and Experiment 2 in that the tendency for /ke/ stimuli to induce /i/ identifications differed across Experiment 5 and Experiment 2. This consideration suggests that the difference between the responses to voiceless stimuli and the responses to voiced stimuli *within* Experiment 5 might be a better indicator of how differently listeners respond to voiceless vs. voiced stimuli, than the cross-experimental comparison between the responses to voiceless stimuli in Experiment 2 and the responses to voiced stimuli in Experiment 5. The /gi/ vs. /ki/ difference was clear

[7] On the other hand, in the /gu/ vs. /ku/ comparison, the standardized residual was the largest with /u/ for /gu/ (2.517) but with /a/ for /ku/ (2.066).

[8] For /pe/, $\chi^2(4) = 4.2547$, $p = .373$ (χ-squared); $p = .309$ (Fisher's). For /pi/, $\chi^2(3) = 6.3822$, $p = .094$ (χ-squared); $p = .060$ (Fisher's). For /pu/, $\chi^2(2) = 2.4963$, $p = .287$ (χ-squared); $p = .261$ (Fisher's).

[9] In the /ki/ comparison, the standardized residual was the largest with /i/ for the Experiment 5 result (2.108) and with /u/ for the Experiment 2 result (2.598), suggesting that the coarticulation-sensitive /i/ responses were more often in Experiment 5 than in Experiment 2. In the /ku/ comparison, the standardized residual was the largest with /a/ for the Experiment 5 result (2.513) and with /u/ for the Experiment 2 result (2.386), suggesting that the mysterious /a/ responses by the listeners in Experiment 5 are responsible. In the /ke/ comparison, the standardized residual was the largest with /e/ for the Experiment 5 result (2.513) and with /i/ for the Experiment 2 result (2.655), and the /i/ standardized residual was the only positive residual for the Experiment 2 result, suggesting that the dominance of /e/ vs. /i/ responses characterize the Experiments 5 & 2 results.

across the two voiceless-voiced comparisons, and the /ge/ vs. /ke/ difference was not that clear in the within-experimental comparison. Thus the only firm difference to be concluded seems to be between /gi/ and /ki/.

Fortunately, the purpose of Experiment 5 was to examine **DI-sensitivity**, and given the **production presuppositions**, the crucial result should be listeners' responses to /i/-coarticulated velars. If **DB-sensitivity** is not real, then /i/-identification should be brought about only through **DI-sensitivity** both for /ki/ and /gi/ stimuli and hence we should expect that /i/-identifications should be the majority both for /ki/ and /gi/ stimuli. However, the within-experiment comparison (as well as the across-experimental comparison) between listeners' response patterns resulted in a significant difference between /ki/ and /gi/, and the most obvious candidate reason is that, while /i/ identifications dominated for /ki/ stimuli, the default /u/ identifications dominated for /gi/ stimuli (as indeed suggested by an examination of the standardized residuals). Such a result could naturally be interpreted in terms of the lack of the effects of **DI-sensitivity**, in which case we would have to attribute all the coarticulation-sensitive responses to **DB-sensitivity**. However, we could also assume that both **DB-sensitivity** and **DI-sensitivity** are real, in which case /i/ identifications were helped by both in the case of /ki/ stimuli but only by the latter in the case of /gi/ stimuli, and hence the different rates of /i/ identifications for /ki/ vs. /gi/ stimuli. To the extent that both interpretations are possible, the comparisons failed to support the reality of **DI-sensitivity**.

Next let us examine the /i/-identification rates for voiceless stimuli and the original vowel restoration rates for voiced stimuli. First, the /i/ identification rates for voiceless stimuli are summarized in Table 6.7. The listener-averaged /i/ response rates were submitted to two-sided one-sample t tests with the expected value being the arcsin square root transform of .2 (the chance level), with Holm corrections separately for /k/ and /p/. Within the /p/ stimuli, the /i/ identification rate for /pe/ was significantly below chance only before Holm corrections [$t(7) = -2.729$, $p = .059$ if corrected; $p = .029$ if uncorrected], while /i/ identification rates

Table 6.7: Mean /i/ response rates for voiceless stimuli in Experiment 5.

stimuli	pe	pi	pu	ke	ki	ku
mean /i/ response rate	.09	.22	.00	.28	.86	.06
S.D.	.19	.28	.00	.34	.19	.12
N	8	8	8	8	8	8

for /pi/ did not differ from chance even before Holm corrections; such results are just as expected from the **production presuppositions**, irrespective of the (un)reality of **DB-sensitivity** or **DI-sensitivity**. Within the /k/ stimuli, the /i/ identification rate was highly significantly above chance with /ki/ [$t(7) = 7.611$, $p < .001$]; however, the rate did not differ from chance with /ke/ [$t(7) = .145$, $p = .889$]. The /ki/ result was just as expected (either from **DB-sensitivity** or **DI-sensitivity**), but the /ke/ results failed to support the reality of **DB-sensitivity**. In order to examine the /ke/ results further, some supplementary analyses were conducted. First, its /u/ and /e/ identification rates were submitted to two-sided one-sample *t* tests (without corrections); if neither **DB-sensitivity** or **DI-sensitivity** was real, we would expect a significant tendency for the default /u/ epenthesis, but the /u/ identification rate was .44 (S.D.= .37) and failed to reach significance [$t(7) = 1.169$, $p = .281$, uncorrected]; a significant tendency for /e/ identification could only be attributed to **DI-sensitivity**, but the /e/ identification rate was .13 (S.D.= 14) and exhibited a trend for a lower-than-chance rate [$t(7) = -2.079$, $p = .076$, uncorrected]. Thus the expectations from **coarticulation insensitivity** or from **DI-sensitivity** failed to be supported too. Next, the /i/ and /e/ identification rates for /ke/ and /pe/ stimuli were compared through two-sided between-subject *t* tests (with no corrections). The three hypotheses (**DB-sensitivity**, **DI-sensitivity** and **coarticulation insensitivity**) predict different effects of the richer /e/-coarticulation cues within /ke/ stimuli than within /pe/ stimuli (the **production presuppositions**); **DB-sensitivity** predicts the different amounts of coarticulation cues should be reflected in /i/, rather than /e/, identification rates; **DI-sensitivity** predicts that they should be reflected in /e/, rather than /i/, identification rates; **coarticulation insensitivity** predicts that

they should have no effect. The results of the t tests were the most compatible with **DB-sensitivity**; the /i/ identification rate for /ke/ [mean = .28, SD = .34] was significantly higher than the /i/ identification rate for /pe/ [mean = .09, SD = .19; $t(7) = 2.546$, $p = .038$], whereas the /e/ identification rate for /ke/ [mean = 13, SD = .23] did not significantly differ from the /e/ identification rate for /pe/ [mean = .03, SD = .09; $t(7) = .837$, $p = .430$]. This result can best be interpreted in terms of **DB-sensitivity**.

Thus the voiceless results do not contradict the view, argued for in the previous chapter, that **DB-sensitivity** is real. However, to arrive at this conclusion, we had to appeal to a comparison between /ke/- and /pe/-results; a simple examination of the /i/ identification rate for /ke/ stimuli failed to support it. Thus, for some reason, the effect of **DB-sensitivity** has been drastically reduced in Experiment 5 than in Experiment 2.

However, if the effect of **DB-sensitivity** was rather weak in Experiment 5, how should we interpret the significant tendency for /i/ identification with /ki/ stimuli? Note that such a tendency itself could be attributed either to **DB-sensitivity** or to **DI-sensitivity**. Thus a natural interpretation would be that, while the effects of **DB-sensitivity** are somehow weakened in Experiment 5, the effects of **DI-sensitivity** are somehow strengthened and made up for the loss of **DB-sensitivity**. However, it would also be possible to attribute it solely to **DB-sensitivity**; /e/-coarticulation is assumed to be weaker than /i/-coarticulation (the **production presuppositions**), and hence an observation of successful exploitation of /i/-coarticulation cues coupled with not so much successful exploitation of /e/-coarticulation does not contradict the idea of **DB-sensitivity**.[10] The /ke/–/ki/ results could be interpreted in both ways; whether **DI-sensitivity** is real or not has to be examined through listeners' responses to voiced stimuli, to which we now turn.

[10] In the case of /k/ stimuli, a significant tendency for /e/ identification with /ke/ stimuli could only be interpreted as an effect of **DI-sensitivity**; it would not be interpretable in terms of **DB-sensitivity**. However, the non-significance of the tendency for /e/ identification with /ke/ stimuli could be interpreted in terms of the weakening of the effect of **DB-sensitivity**, where its weakening presupposes its presence.

Table 6.8: Mean vowel restoration rates for voiced stimuli in Experiment 5.

stimulus	bi	bu	be	gi	gu	ge
accuracy	.28	.91	.09	.06	.97	.03
S.D.	.25	.19	.19	.18	.09	.12
N	8	8	8	8	8	8

Recall that, in the case of voiced stimuli, the hypotheses to be compared are **coarticulation insensitivity** and **DI-sensitivity**. The good identifications with /u/ stimuli are expected from either hypothesis, so the question is whether identifications are good with non-/u/ stimuli; **DI-sensitivity** would be supported if listeners exhibited good identifications with non-/u/ stimuli. However, this expectation does not seem to be borne out, as seen in the original vowel restoration rates in the **C-set** summarized in Table 6.8.

The arcsin square root transforms of the mean accuracy scores with /be, bi, ge, gi/ were submitted to two-sided one-sample *t*-tests, with the expected value being the arcsin square root transformed value of .2, the chance level in a one-from-five choice task; the significance values are Holm corrected. Against the expectation from **DI-sensitivity**, none of /be, bi, ge, gi/ resulted in above-chance accuracy; the scores with /ge/ were significantly below chance [$t(7) = -6.084$, $p < .001$], as well as the scores with /gi/ [$t(7) = -3.729$, $p = .022$; $p = .007$ if uncorrected]; the scores with /be/ was significantly below chance before a Holm correction [$t(7) = -2.729$, $p = .059$; $p = .029$ if uncorrected], and the scores with /bi/ did not differ from chance [$t(7) = -.217$, $p = .834$ uncorrected]. Thus the results failed to support **DI-sensitivity**.

With the results from the voiceless stimuli, the effect of **DB-sensitivity** on the /ke/ result could be defended because the /i/ identification rate for the /ke/ stimuli was not significant but still above chance on the one hand, and it differed from the /i/ identification rate for the /pe/ stimuli in a way expected from **DB-sensitivity**. However, with the results from the voiced stimuli, no statistical effort to defend **DI-sensitivity** seems possible, given the significantly low vowel restoration rates, even for the /gi/ stimuli, for which restoration success was expected the

most.[11]

6.2 Experiment 6

The failure to observe effects of **DI-sensitivity** in Experiment 5 could be due to its task; identification would involve phonemic categorization, and the effects of **DI-sensitivity**, which are assumed to be auditory rather than phonological, might have been suppressed with such a task.[12] If so, discrimination could be a better means to examine such effects, because discrimination would not suppress sub-phonemic sensitivity as much as identification would.

Imitating Experiment 3, Experiment 6 examined their discriminatory abilities with 1,500 ms interstimulus intervals (ISI's), with the discrimination being AX discriminations between different talkers. As stated above, it has already been noted in the literature that listeners' responses tend to be based on phonological, rather than acoustic, representations, when (i) the experimental task is memory-demanding (e.g., ABX rather than AX), (ii) the interstimulus interval (ISI) is long (say, 1,500 ms), and (iii) the stimuli to be compared are uttered by different speakers and hence cannot be compared without resorting to linguistic representations (Dupoux et al., 1997; Dupoux et al., 2001; Matthews & Brown, 2004; Werker & Logan, 1985). Thus, although it is a discrimination task, Experiment 6 still encourages phonological processing with respect to (ii) and (iii). However, crucially, the experimental task does not require explicit phonemic categorization. If Experiment 5 failed to support **DI-sensitivity** because the task suppressed listeners' sensitivity to sub-phonemic distinctions, Experiment 6 might reveal such sensitivity.[13]

[11] Why only the /bi/ result, rather than the /gi/ result, was non-significantly above chance, is not clear.

[12] Under this interpretation, the significantly successful /i/ identification for /ki/ stimuli in Experiment 5 would be mostly attributed to **DB-sensitivity**, while the lack of a similarly successful /i/ identification for /ke/ stimuli would be attributed to the greater amount of coarticulation cues needed by the listeners in Experiment 5 than the listeners in Experiment 2.

[13] Dupoux et al. (2011) employed the value of '/i/ response rate minus /u/ response rate' as the dependent measure to examine the effect of /i/-coarticulation. Similarly, we could employ the value of 'original vowel restoration rate minus /u/ identification rate' as the dependent measure. **DI-sensitivity** could be defended if such difference scores are significantly above zero. However, for all of /be/, /bi/, /ge/, and /gi/, such difference scores were negative,

6.2.1 Method

Stimuli

The voiceless and voiced C-set stimuli in Experiment 5 were first divided into those in which the medial consonant was /g/ (the /g/ set), those in which the medial consonant was /b/ (the /b/ set), those in which it was /k/ (the /k/ set) and those in which it was /p/ (the /p/ set). Each stimulus in the /g/ set by one speaker was concatenated with each stimulus in the /g/ set by the other speaker, with an ISI of 1,500 ms. Similarly for the /b/, /k/ and /p/ sets.

The four sets were intermixed and presented in a quasi-random order through circumaural closed-back headphones (SONY MDR-ZX700).

Participants

Ten native Japanese listeners with no known hearing disability participated in the experiment for course credit at undergraduate programs at Hosei University, Tokyo (except one, who participated "just out of curiosity," and another one, who participated for an extra-curricular reason). They consisted of four males and six females (mean age = 20.2, SD = 1.14).

Procedure

The experiment was controlled by E-prime and conducted in a sound-attenuated room. They were told that the experiment was meant to examine how non-Japanese sounds are perceived by Japanese listeners, and they were instructed to judge whether the medial portion surrounded by /e/ and /ma/ are the same. The responses were made by pressing either the "1" or the "2" key on the computer ("1 = same" and "2 = different").

Again, only after scoring more than 80 % accuracy in the practice session (with the same stimuli employed in Experiment 3) were they allowed to proceed to the main experimental session.

Table 6.9: The listener-averaged accuracy scores in Experiment 6.

pair	be–bi	be–bu	bi–bu	ge–gi	ge–gu	gi–gu
mean	.50	.49	.49	.51	.63	.61
S.D.	.19	.16	.14	.15	.19	.17
N	10	10	10	10	10	10

Statistical Analyses

As in Experiments 3–4, three different measures are employed: arcsin square root transforms of listener-avearged discrimination accuracies, *d* values calculated in the Independent Observations method, and H–F values. Those measures with /be/–/bi/, /be/–/bu/, /bi/–/bu/, /ge/–/gi/, /ge/–/gu/, and /gi/–/gu/ pairs are submitted separately to two-sided one-sample *t* tests, with Holm corrections of the significance values. While **coarticulation insensitivity** leads us to expect that the accuracy should be significantly below chance for any pair, **DI-sensitivity** would be supported by the existence of a pair for which the accuracy is significantly better than chance; significant success with any pair would be a support for **DI-sensitivity**, while only significant failure with all the pairs would be a support for **coarticulation insensitivity**. We are particularly interested in /iu/ discriminations, because Dupoux et al.'s (2011) arguments for one-step models were based on comparisons between /i/ and /u/ perception by Japanese listeners.

However, pairwise comparisons among the pairs (conducted for Experiments 3–4) will not be conducted, because **DI-sensitivity** makes no specific predictions.

6.2.2 Results and Discussion

The listener-averaged accuracies are as described in Table 6.9, according to which discrimination performance is clearly not good.

The listener-averaged accuracy rates were, again, arcsin square root transformed and submitted to one-sample *t* tests with the expected value of the transform of 0.5 (two-sided). Dis-

Table 6.10: The listener-averaged d' values in Experiment 6.

pair	/be–/bi/	/be/–/bu/	/bi/–/bu/	/ge/–/gi/	/ge/–/gu/	/gi/–/gu/
mean	.78	1,04	,,48	.77	1.78	1.61
S.D.	1,09	1.06	1.80	1.26	2.25	1.31
N	10	10	10	10	10	10

Table 6.11: The listener-averaged β values in Experiment 6.

pair	/be–/bi/	/be/–/bu/	/bi/–/bu/	/ge/–/gi/	/ge/–/gu/	/gi/–/gu/
mean	1.03	.98	1.00	1.03	1.15	1.20
S.D.	.19	.16	.16	.18	.22	.20

criminations between bilabial pairs were far from being significant [with /be/–/bi/, $t(9) = -.071$, $p = .945$ uncorrected; with /be/–/bu/, $t(9) = -.241$, $p = .815$ uncorrected; with /bi/–/bu/, $t(9) = -.280$, $p = .786$ uncorrected]. The /ge/–/gi/ discriminations were also far from being significant [$t(9) = .296$, $p = .774$ uncorrected]. The /gi/–/gu/ discriminations were significantly successful only before Holm corrections [$t(9) = 2.038$, $p = .072$ if uncorrected but $p = .265$ if corrected], as well as the /ge/–/gu/ discriminations [$t(9) = 1.912$, $p = .088$ if uncorrected, but the Holm procedure dictates that this is not significant].

Next the d' values were examined. The descriptive results are summarized in Tables 6.10–6.11. The listener-averaged d' values were submitted, for each pair, to one-sample t tests with the expected value of 0 (two-sided). Again, bilabial discriminations were all non-significant (after Holm corrections) [with /be/–/bi/, $t(9) = 2.260$, $p = .151$ if corrected, $p = .050$ if uncorrected; with /be/–/bu/, $t(9) = 1.904$, $p = .179$ if corrected, $p = .089$ if uncorrected; with /bi/–/bu/, $t(9) = 1.441$, $p = .183$ uncorrected]. in contrast, the /gi/–/gu/ discriminations were significantly successful [$t(9) = 3.891$, $p = .011$ if corrected, $p = .004$ if uncorrected], while the /ge/–/gu/ discriminations only showed a trend towards significant success after Holm corrections [$t(9) = 2.506$, $p = .067$ if corrected, $p = .034$ if uncorrected], and the /ge/–/gi/ discriminations showed a trend only before Holm corrections [$t(9) = 1.932$, $p = ..085$, uncorrected].

Table 6.12: The listener-averaged H–FA scores in Experiment 6; the highest possible is 4.0, and the lowest possible score is −4.0.

pair	/be–/bi/	/be/–/bu/	/bi/–/bu/	/ge/–/gi/	/ge/–/gu/	/gi/–/gu/
mean	.00	−.10	−.10	.10	1.00	.90
S.D.	1.49	1.29	1.10	1.20	1.49	1.37
S.D.	1.46	.33	1.47	1.66	1.69	1.29

Finally, the H–FA score results, summarized in Table 6.12, were examined. The listener-averaged H–FA scores were submitted, for each pair, to one-sample t tests with the expected value of 0 (two-sided). The bilabial discriminations were again all far from being significant [with /be/–/bi/, $t(7) = .000$, $p = 1.000$, uncorrected; with /be/–/bu/, $t(7) = -.246$, $p = .811$, uncorrected; with /bi/–/bu/, $t(7) = -.287$, $p = .780$], as well as the /ge/–/gi/ discrimination [$t(7) = .264$, $p = .798$, uncorrected]. This time, the /ge/–/gu/ and /gi/–/gu/ discriminations failed to be significant even when uncorrected [with /ge/–/gu/, $t(7) = 2.121$, $p = .063$, uncorrected; with /gi/–/gu/, $t(7) = 2.077$, $p = .068$, uncorrected].

Thus the only clearly significant success was with /gi/–/gu/ discriminations, which were significantly successful only when the measure is d'. Note that the d' values are calculated by regarding the values as zero when false alarm rates were larger than hit rates, and hence the d' values rather overestimate discrimination successes.[14] Thus, if the only evidence for successful /gi/–/gu/ discriminations come from results with d', that evidence is rather weak. On the other hand, we are testing a directional hypothesis that some discriminations should be better than chance, rather than a non-directional hypothesis that some discriminations should be either better or worse than chance, which would justify one-sided tests; if the H–FA results were submitted to one-sided, rather than two-sided, one-sample t tests, /ge/–/gu/ and /gi/–/gu/ discrimination results would count as significant, but only before Holm corrections [with /ge/–/gu/, $p = .031$; with /gi/–/gu/, $p = .034$], and they would still fail to count as significant after

[14]This 'corrected' calculation was applied to eight listeners with /ge/–/gu/ and to two listeners with /gi/–/gu/, where the total number of the listeners was ten. The overestimation must be considerable with /ge/–/gu/.

Holm corrections [with /ge/–/gu/, $p = .094$, and the Holm procedure dictates that /gi/–/gu/ should be regarded as non-significant].

Thus discrimination successes did not reach clear significance with no pair, and hence the results of Experiment 6 did not fully support the reality of **DI-sensitivity**.

6.3 Experiment 7

As noted above, discriminations with a 1,500 ms ISI still encourages phonological processing, as opposed to discriminations with a shorter ISI. Thus, although Experiment 6 failed to provide firm evidence for **DI-sensitivity**, the failure could be because the ISI was too long. Thus, as with Experiment 4, Experiment 7 examines listeners' discriminations with a 250 ms ISI; given the auditory nature of **DI-sensitivity**, ISI shortening may produce evidence for it. Such expectations can be examined twice: by testing significant success for the discriminations with each pair on the one hand, and by comparing the results of Experiment 7 with those of Experiment 6.

6.3.1 Method

Stimuli

The same stimuli as Experiment 6 were employed, except that the ISI was 250 ms, rather than 1,500 ms.

Participants

Ten native listeners of Japanese (five males and five females; mean age = 20.7, SD = 1.25) with no known hearing disability participated, either for course credit at undergraduate programs at Hosei University, Tokyo, or for an extra-curricular reason. None of them had participated in the other experiments.

Table 6.13: The discrimination accuracies in Experiment|7.

pair	/be–/bi/	/be/–/bu/	/bi/–/bu/	/ge/–/gi/	/ge/–/gu/	/gi/–/gu/
mean	.54	.60	.61	.58	.78	.76
S.D.	.19	.11	.15	.58	.78	.76
N	10	10	10	10	10	10

Procedure

The same procedure as Experiment 6 was employed.

Statistical Analyses

As with Experiment 6, three different measures are employed: (i) arcsin square root transforms of the listener-averaged discrimination accuracy scores, (ii) listener-averaged Independent Observations d' values, and (iii) H–FA values. Those three measures with /be/–/bi/, /be/–/bu/, /bi/–/bu/, /ge/–/gi/, /ge/–/gu/ and /gi/–/gu/ are separately submitted to two-sided one-sample *t* tests with the expected value of the arcsin square root transform of .5 (accuracy scores) or 0 (d' values or H–FA values).

Furthermore, the arcsin square root transforms of the listener-averaged discrimination accuracy scores from Experiments 6–7 are submitted to a two-way mixed-design ANOVA, with the within-subject factor being 'consonant' and the between-subject factor being 'experiment'. According to **coarticulation insensitivity**, ISI should have no effect, whereas according to **DI-sensitivity**, ISI shortening should result in better discriminations.

6.3.2 Results and Discussion

The listener-averaged accuracy rates for each pair are described in Table 6.13. This time, the /ge/–/gu/ and the /gi/–/gu/ discriminations seem good. This impression is supported by the results of two-sided one-sample *t* tests conducted on the (arcsin square root transforms of) the accuracy rates. The /be/–/bi/ discrimination was far from being significant even if uncorrected

Table 6.14: The listener-averaged Independent Observations d' values in Experiment 7.

pair	/be–/bi/	/be/–/bu/	/bi/–/bu/	/ge/–/gi/	/ge/–/gu/	/gi/–/gu/
d'	1.03	.98	1.00	1.03	1.14	1.14
S.D.	.19	.16	.16	.18	.22	.20
N	10	10	10	10	10	10

Table 6.15: The Independent Observations β values in Experiment 7.

pair	/be–/bi/	/be/–/bu/	/bi/–/bu/	/ge/–/gi/	/ge/–/gu/	/gi/–/gu/
β	1.06	1.12	1.15	1.08	1.33	1.32
S.D.	.20	.13	.19	.12	.24	.25
N	10	10	10	10	10	10

[$t(9) = .551$, $p = .595$, uncorrected]. The /be/–/bu/ and the the /bi/–/bu/ discriminations were significantly good, but only before Holm corrections [with /bi/–/bu/, $t(9) = 2.739$, $p = .023$ if uncorrected but $p = .069$ if corrected; with /bi/–/bu/, $t(9) = 2.341$, $p = .044$ if uncorrected but the Holm procedure dictates that this is not significant]. The /ge/–/gi/ discrimination was not significant even before Holm corrections [$t(9) = 1.717$, $p = .120$]. However, both the /ge/–/gu/ and the /gi/–/gu/ discriminations were significantly successful even after Holm corrections [with /ge/–/gu/, $t(9) = 3.943$, $p = .003$ if uncorrected, $p = .010$ if corrected; with /gi/–/gu/, $t(9) = 3.539$, $p = .006$ if uncorrected, $p = .013$ if corrected].

The listener-averaged Independent Observations d' values are summarized in Table 6.14, with the listener-averaged β values summarized in Table 6.15. The listener-averaged d' values were submitted to two-sided one-sample t tests, separately for each pair, with the expected value of 0. This time, bilabial discriminations were significantly better than chance [with /be/–/bu/, $t(9) = 3.372$, $p = .019$ if corrected, $p = .008$ if uncorrected; with /bi/–/bu/, $t(9) = 3.356$, $p = .017$ if corrected, $p = .008$ if uncorrected; with /be/–/bu/, $t(9) = 2.873$, $p = .018$]. Velar discriminations were highly significantly successful [with /ge/–/gu/, $t(9) = 4.824$, $p = .003$ if corrected, $p = .001$ if uncorrected; with /gi/–/gu/, $t(9) = 4.725$, $p = .002$ if corrected, $p = .001$ if uncorrected; with /ge/–/gi/, $t(9) = 4.227$, $p = .002$].

Table 6.16: The listener-averaged H–FA scores in Experiment 7; the highest possible is 4.0, and the lowest possible score is −4.0.

pair	/be–/bi/	/be/–/bu/	/bi/–/bu/	/ge/–/gi/	/ge/–/gu/	/gi/–/gu/
mean	.30	.80	.90	.60	2.20	2.10
S.D.	1.49	.92	1.20	1.07	1.48	1.52
N	10	10	10	10	10	10

Next, the listener-averaged H–FA scores, summarized in Table 6.16, were examined. The listener-averaged H–FA scores were submitted to two-sided one-sample t tests, separately for each pair, with the expected value of 0. The /be/–/bi/ discrimination was not significant even when uncorrected [$t(9) = .635$, $p = .541$ uncorrected]. The /be/–/bu/ and /bi/–/bu/ discriminations were significantly above chance only when uncorrected [with /be/–/gu/, $t(9) = 2.753$, $p = .022$ if uncorrected but $p = .067$ if corrected; with /bi/–/bu/, $t(9) = 2.377$, $p = .041$ if uncorrected but not significant according to the Holm procedure]. The /ge/–/gi/ discrimination was not significant even when uncorrected [$t(9) = 1.765$, $p = .111$ uncorrected]. However, both the /ge/–/gu/ and /gi/–/gu/ discriminations were highly significantly successful [with /ge/–/gu/, $t(9) = 4.714$, $p = .001$ if uncorrected, $p = .003$ if corrected; with /gi/–/gu/, $t(9) = 4.358$, $p = .002$ if uncorrected, $p = .004$ if corrected].

Discriminations with pairs other than /ge/–/gu/ and /gi/–/gu/ were significantly successful when (and only when) the measure is d'; again, given the possibility of overestimating discrimination successes due to the way d' values are calculated, such results are rather weak as evidence for the reality of **DI-sensitivity**.[15] However, irrespective of the choice of the dependent measure, the /ge/–/gu/ and /gi/–/gu/ discriminations are clearly successful significantly; the /ge/–/gu/ and the /gi/–/gu/ results thus clearly support the reality of **DI-sensitivity**.

Furthermore, the results of Experiments 6–7 are compared through a two-way mixed-design ANOVA, with the within-subject factor being 'consonant' and the between-subject fac-

[15]The 'corrected' calculations of the d' values were applied to five (/be/–/bi/), three (/be/–/bu/), four (/bi/–/bu/) and one (/ge/–/gi/) listeners, where the total number of the listeners was ten.

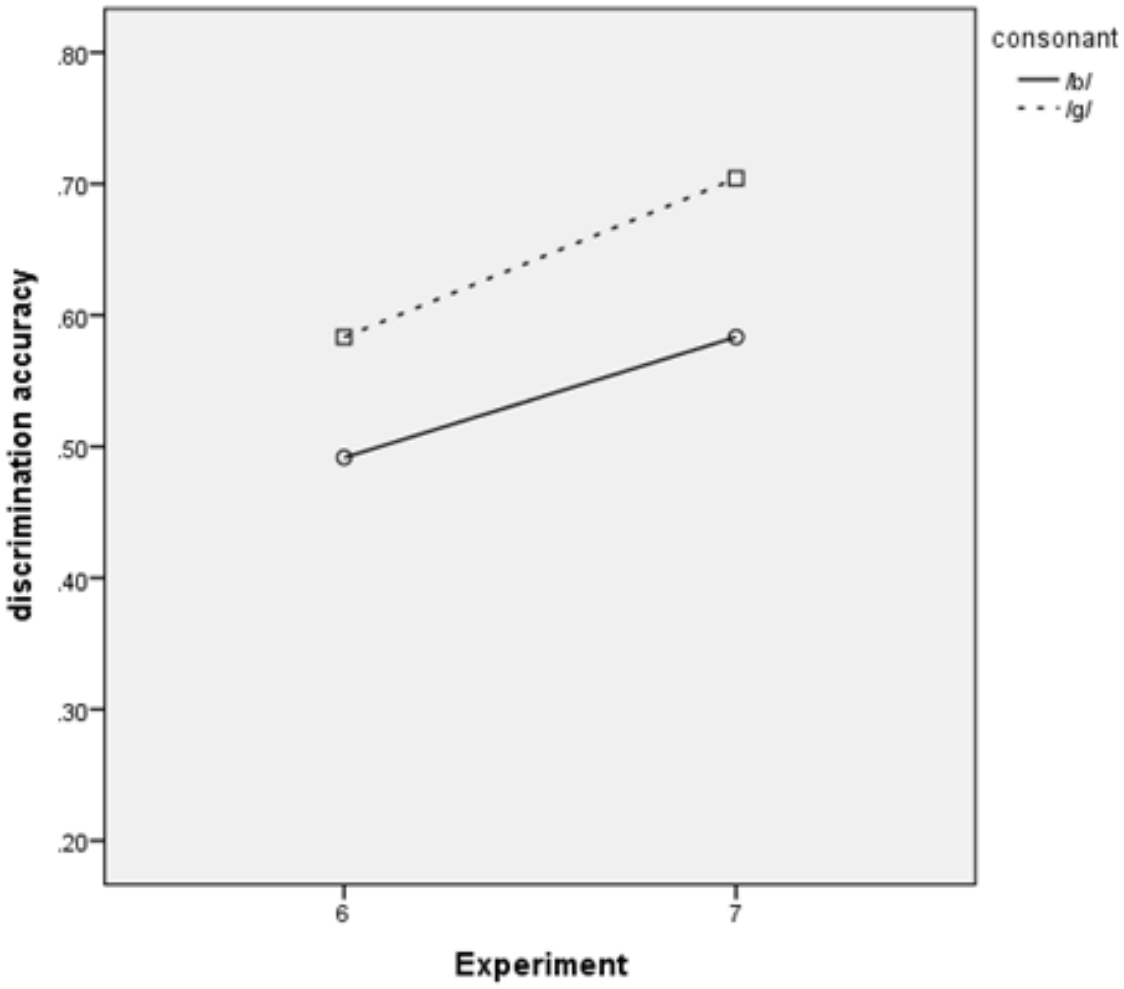


Figure 6.2: The listener-averaged accuracy scores across Experiments 6–7.

tor being ‘experiment’. Unlike **DB-sensitivity**, **DI-sensitivity** leads us to expect ISI shortening to have either no effect or a uniform facilitatory effect. On the other hand, the prediction of better discriminations with velar pairs than with bilabial pairs is common. Thus, the results of the bilabial pairs on the one hand, and of the velar pairs on the other, are conflated.

The listener-averaged accuracy scores across the two experiments are depicted in Figure 6.2, which suggests that (i) discrimination was more successful with the /g/ stimuli than with the /b/ stimuli in both experiments, and (ii) discrimination was a bit more successful in Experiment 7 than in Experiment 6 with both /b/ and /g/ stimuli. According to the results

of the ANOVA conducted with the arcsin square root transformed listener-averaged accuracy scores, velar discriminations were significantly better than bilabial discriminations [the main effect of 'consonant'; $F(1,\ 18) = 5.811,\ p = .027$], in line with the expectation from the **production presuppositions**, Furthermore, the discriminations were highly significantly better in Experiment 7 than in Experiment 6 [the main effect of 'experiment'; $F(1,\ 18) = 14.112,\ p = .001$], which supports the reality of **DI-sensitivity**. Although the interaction between 'consonant' and 'experiment' failed to be significant [$F(1,\ 18) = .133,\ p = .720$], an examination of the simple main effects suggests that better *velar* discriminations in Experiment 7 than in Experiment 6 [$F(1,\ 18) = 4.774,\ p = .042$] made a clearer contribution to the whole cross-experimental difference than better *bilabial* discriminations in Experiment 7 than in Experiment 6 [$F(1,\ 18) = 3.131,\ p = .094$], which is in line with the observation that it was the /ge/–/gu/ and /gi/–/gu/ discriminations whose significance differed cross-experimentally according to the one-sample *t* tests.

The listener-averaged Independent Observations d' values across the two experiments are depicted in Figure 6.3, Similar results were obtained from the ANOVA conducted on d' values. The velar discriminations were significantly better than the bilabial discriminations [the main effect of 'consonant'; $F(1,\ 18) = 10.052,\ p = .002$], and the whole discriminations were highly significantly better in Experiment 7 than in Experiment 6 [the main effect of 'experiment'; $F(1,\ 18) = 9.929,\ p = .006$]; although the interaction between 'consonant' and 'experiment' was not significant [$F(1,\ 18) = 1.174,\ p = .281$], an examination of the simple main effects suggests that better *velar* discriminations in Experiment 7 than in Experiment 6 [$F(1,\ 18) = 10.324,\ p = .002$] made a clearer contribution to the whole cross-experimental difference than better *bilabial* discriminations in Experiment 7 than in Experiment 6 [$F(1,\ 18) = 2.826,\ p = .096$],

The listener-averaged H–FA scores are depicted in Figure 6.4. Again, similar results were obtained from the ANOVA conducted on listener-averaged H–FA scores. Velar discrimina-

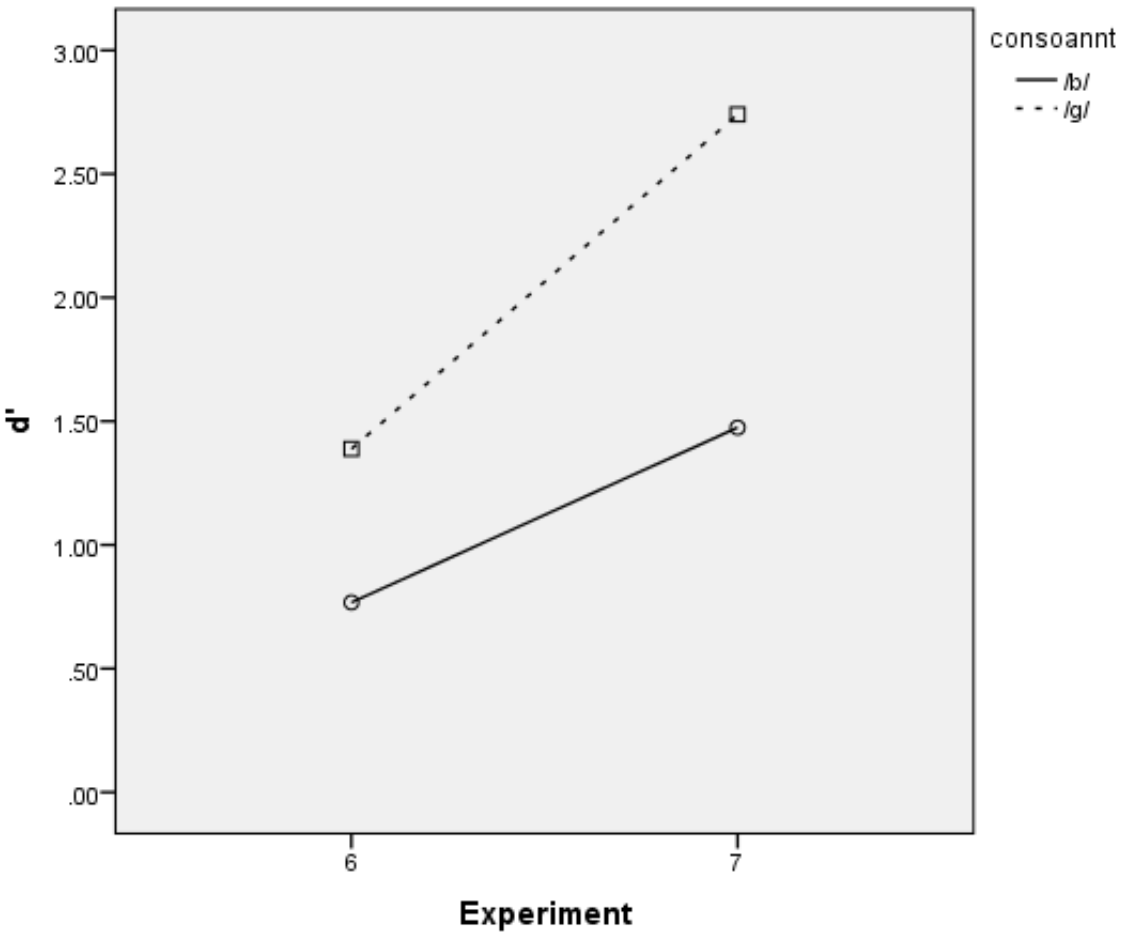


Figure 6.3: The listener-averaged Independent Observations d' values across Experiments 6–7.

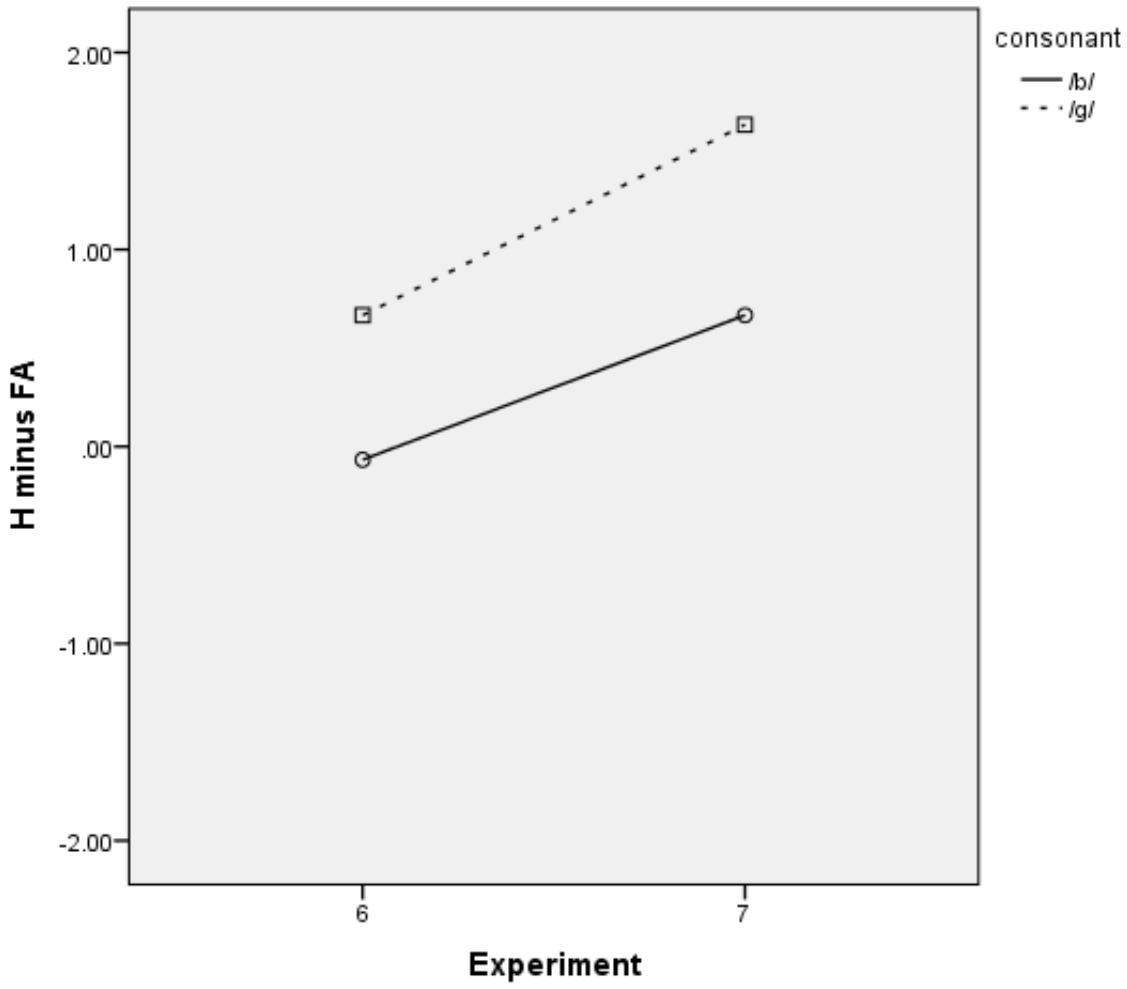


Figure 6.4: The listener-averaged H–FA scores across Experiments 6–7.

tions were significantly better than bilabial discriminations [the main effect of 'consonant'; $F(1,\ 18) = 5.874,\ p = .026$], and discriminations were significantly better in Experiment 7 than in Experiment 6 [the main effect of 'experiment'; $F(1,\ 18) = 14.059,\ p = .001$]; although the interaction between 'consonant' and 'experiment' was not significant [$F(1,\ 18) = .111,\ p = .743$], an examination of the simple main effects suggests that the better *velar* discriminations in Experiment 7 than in Experiment 6 [$F(1,\ 18) = 5.297,\ p = .034$] made a clearer contribution to the whole cross-experimental difference than the better *bilabial* discriminations in Experiment 7 than in Experiment 6 [$F(1,\ 18) = 2.689,\ p = .094$].

The ANOVA results all uniformly suggest that ISI shortening indeed facilitates discriminations between voiced pairs. Coupled with the significant discrimination success with /ge/–/gu/ and /gi/–/gu/ observed in Experiment 7, the ANOVA results also argue for the reality of **DI-sensitivity**.

6.4 Experiment 8

In Experiment 5, identification accuracy was not significantly better than chance with any voiced stimulus. Given that (i) identification encourages phonological processing more than discrimination, and (ii) **DI-sensitivity** is presumably auditory sensitivity to coarticulation cues, effects of **DI-sensitivity** are presumably more likely to be manifested in a discrimination experiment than in an identification experiment. The results of Experiments 6–7 indeed supported the reality of **DI-sensitivity**, both in terms of the significant discrimination successes with /ge/–/gu/ and /gi/–/gu/ (particularly in Experiment 7) and in terms of the facilitatory effect of ISI shortening. Thus, we have finally obtained some discrimination-based evidence for **DI-sensitivity**, but we still lack identification-based evidence.

One possible reason for the failure in obtaining a clear identification-based support for **DI-sensitivity** in Experiment 5 is that the stimuli were produced by Japanese speakers. As already

noted, closures accompanying voiced stops by Japanese speakers are presumably rather weak, so weak that they often become fricatives rather than stops (Chapter 2).In fact, the Japanese speakers for Experiments 5–7 had to be asked to iterate their relevant utterances with voiced stops until closure periods and bursts look visually distinguished on spectrograms. Because the finally employed stimuli could have been the results of somewhat artificial productions, the coarticulation traces may have been rather unnatural. This potential problem would be more serious for velars than for bilabials, because velar closures are more difficult for phonetically naive speakers to be conscious of. If so, the failure to observe effects of **DI-sensitivity** in Experiment 5 could be attributed not to the unreality of **DI-sensitivity** but rather to the combination of the nature of the task (identification) and the poor acoustic quality of the coarticulation traces in the voiced bursts produced by Japanese speakers; in short, the idea is that **DI-sensitivity** needs stronger coarticulation cues when the task is strongly phonological (identification) than when it is rather auditory (or weakly phonological at best; discrimination).

In fact, there is a reason to assume that coarticulation cues to be exploited by **DI-sensitivity** need to be stronger with a strongly phonological task (identification) than with a task whose nature is rather auditory (or weakly phonological at best; discrimination).[16] It is assumed in this thesis that **DI-sensitivity** is auditory in nature but could be exerted only in phonotactic repair. That means that its effects could be successfully exerted only when phonological processing (phonotactic repair) completes before the coarticulation traces in the auditory memory decay or get erased by the results of phonological processing. The completion of phonological processing (phonotactic repair) and the resistance to the erasure of the coarticulation traces in auditory memory by phonological processing are contradictory requirements, and to overcome such contradictory requirements, the coarticulation traces would have to be strong enough. If this consideration is coupled with the possibility that the stimuli employed in Experiments 5–7 may have lacked natural coarticulation traces, it would not be surprising that identification-

[16]This is partially stated in footnote 8 on page 150; it is repeated here in a more elaborated form.

based evidence for **DI-sensitivity** failed to be observed with stimuli produced by Japanese speakers.

If this is the right interpretation of the results of Experiment 5, we might be able to obtain identification-based evidence for **DI-sensitivity** if the stimuli are the right kinds acoustically (i.e., if the quality of the coarticulation cues is good). Thus another identification experiment was conducted; this time, /eCVta/ utterances by an American English speaker, recorded for a slightly different project, are employed as stimuli.[17] In contrast to the Japanese speakers' utterances, those /eCVta/ utterances by this American female speaker exhibited clear closure periods and bursts with no requested repetitions.[18] Although no clear difference between the Japanese speakers' and the American speaker's utterances seems noticeable from the waveforms and spectrograms in Figure 6.5, the LPC spectra in Figure 6.6 suggest some difference. Speaking more specifically, three differences are found between the Japanese speakers' /gi/ bursts and the American speaker's /gi/ burst: (i) the center of gravity was higher in the Japanese speakers' /gi/ bursts (2320 Hz and 1520 Hz) than the American speaker's /gi/ burst (987 Hz), (ii) while the F2 values of all the speakers' /i/ were in the range between 2100 Hz and 2820 Hz, a peak within this range was found in the LPC spectrum of the American speaker's /gi/ burst (2814 Hz), while no peak within this range was found neither in the LPC spectra of the Japanese speakers' /gi/ bursts,[19] and (iii) the first peak is relatively sharper followed by a longer dip (valley) in the American speaker's /gi/ burst than in the Japanese speaker's /gi/ bursts.

Although the issue of what acoustic properties count as exploitable coarticulation traces by Japanese listeners is beyond the scope of this thesis, the LPC spectra of the /ki/ bursts in the stimuli for Experiments 1–2 (shown in Figure 6.7) suggest the possibility that the American speaker's /gi/ burst with a shaper first peak followed by a longer dip enables us to observe

[17] Recall that, as long as the consonant before the 'devoicing site' is voiced, the voicing of the consonant following the 'devoicing site' does not affect the probability of vowel devoicing; thus devoicing is not more likely with /ebVta/ or /egVta/ than with /ebVma/ or /egVma/.

[18] The Japanese speakers had to repeat the productions of the stimuli many times until closures were visible in the spectrograms.

[19] The idea to focus on the LPC peak in the F2 range was borrowed from Tsuchida (1994).

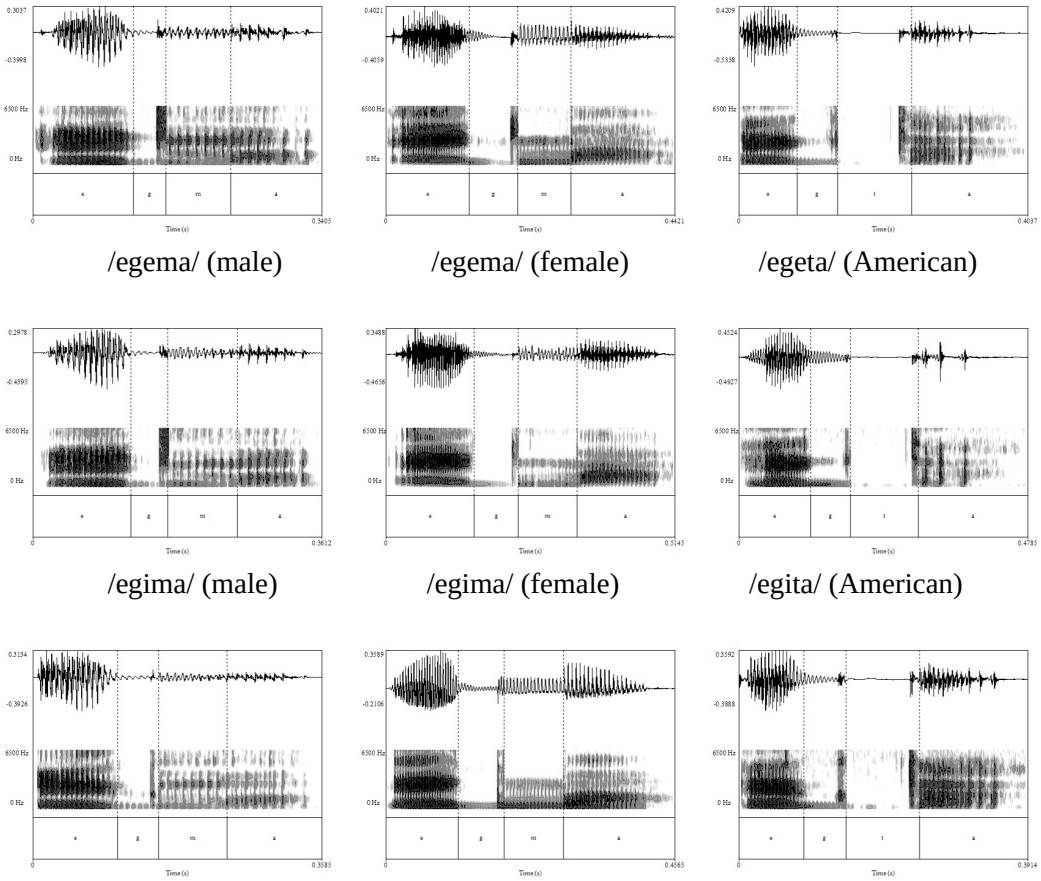


Figure 6.5: The waveforms and spectrograms of the C-set /ge, gi, gu/ stimuli for Experiments 5 and 8; the stimuli produced by the Japanese speakers are on the left and the middle column, and the stimuli produced by the American speaker are on the right column.

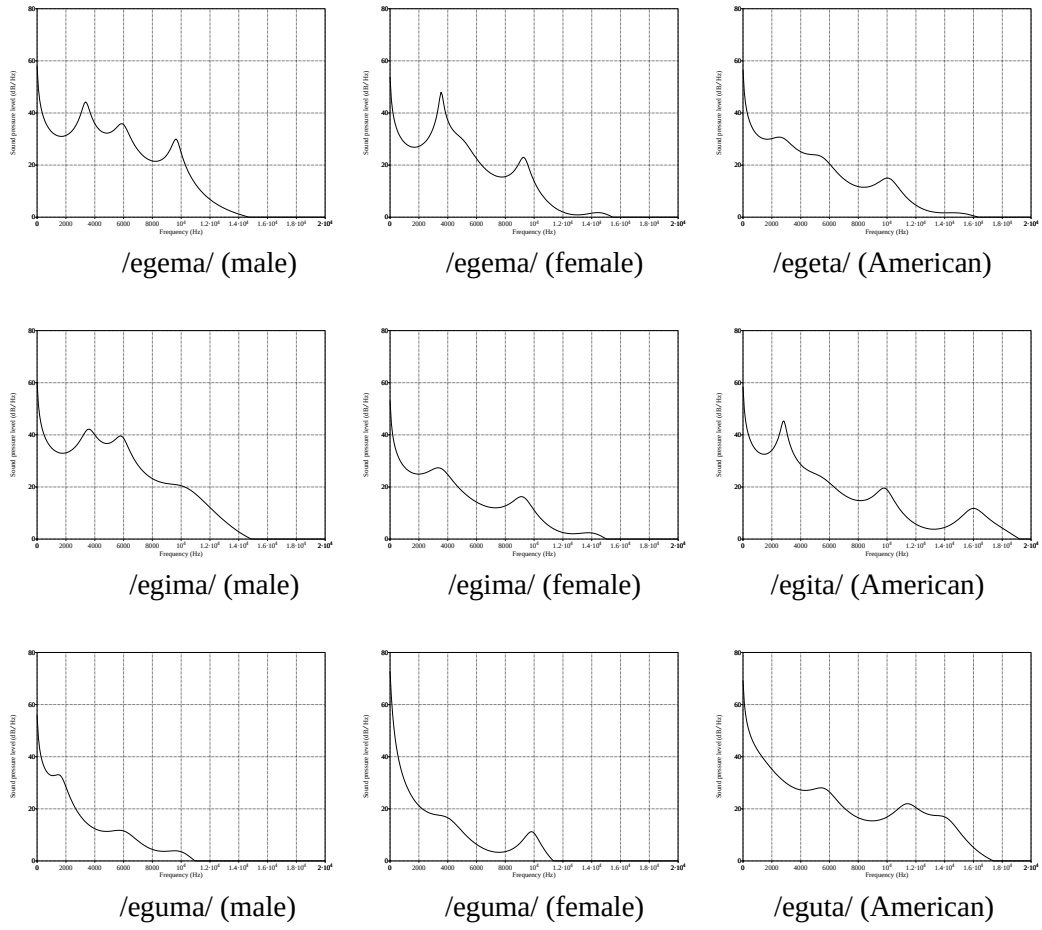
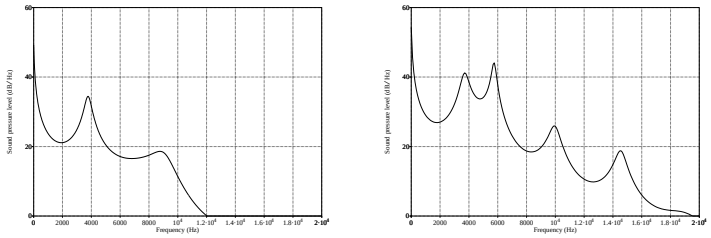
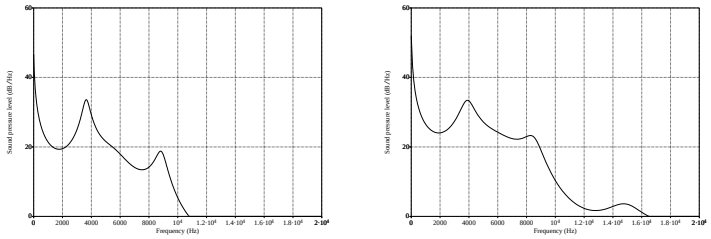


Figure 6.6: The LPC spectra of the velar bursts; the Japanese speakers' utterances are shown on the left and the middle column, and the American speaker's utterances are shown on the right column.



Experiment 1
(The first female speaker)

Experiment 2
(The first female speaker)



Experiment 1
(The second female speaker)

Experiment 2
(The second female speaker)

Figure 6.7: The LPC spectra of the /ki/ burst portions in Experiment 1 (the left panels) and in Experiment 2 (the right panels).

Japanese listeners' successful identification based on **DI-sensitivity**. Recall that /i/ identification from /ki/ bursts in Experiment 2 was not as successful as in Experiment 1. The centers of gravity were: 3806 Hz in Experiment 1 vs. 4745 Hz in Experiment 2 with the first female speaker; 3349 Hz in Experiment 1 vs. 1663 Hz in Experiment 2 with the second female speaker. The first peaks are not noticeably different across Experiments 1–2, according to Figure 6.7. Thus it is not clear what roles are played by the centers of gravity or the first peaks' frequencies on the reduced coarticulation-sensitive identifications in Experiment 2 than in Experiment 1. However, as seen in Figure 6.7, the burst stimuli employed in Experiment 1 had a sharper first peak followed by a longer dip than the burst stimuli employed in Experiment 2. Thus, if at least such a spectral shape plays some role, we might be able to observe Japanese listeners' successful /i/ identification based on **DI-sensitivity**. At the same time, however, the spectral shapes of the American speaker's /ge/ burst lack the sharp first peak followed by a long dip and look rather similar to /gu/ bursts. Thus chances are that Japanese listeners will fail to exhibit coarticulation sensitivity with the American speaker's /ge/ bursts if spectral shapes do matter.

The identification experiment was conducted a short break after Experiment 5 (within the same sessions, with the same participants).

6.4.1 Method

Stimuli

A female native American (Southern Californian) English speaker, age 21 and a beginning learner of Japanese, produced /eCVta/ utterances, where C is either /p/, /k/, /b/ or /g/, V is one of the Japanese five vowels /a, i, u, e, o/, and an accent is placed on the first mora. Her utterances were recorded in a sound-attenuated room with Marantz Solid State Recorder PMD650 with the sampling frequency of 44,100 Hz. Again, two sets of stimuli were produced from the original utterances: **the C-set** was created by deleting the bursts and the V's (the voiced portions between the bursts and the following stops' closures) except the initial 10 ms portions of the bursts of the C's, and **the V-set** was created by removing three pitch periods artitrarily chosen from the midst of the V's. The intensity of each stimuli was rescaled to the same lvel before presentation. The C- and the V-set were intermixed in a random order generated by the experiment software (Eprime) and presented to the listeners through circumaural closed-back headphones (SONY MDR-ZX700) in a sound-attenuated room. (See Figure 6.5 for example waveforms and spectrograms of the C-set stimuli.)

Participants

The participants in Experiments 5.

Procedure

This experiment was conducted immediately following Experiment 5 (as soon as each participant felt that he or she was ready). As in Experiment 5, the task was, in effect, a forced-choice vowel identification.

Statistical Analyses

Again, first the patterns of the responses to voiced and to voiceless stimuli are compared through χ-squared tests and Fisher's exact tests, separately for each place-vowel combination, to see if the response patterns differ across voiceless and voiced stimuli (with no multiple comparison corrections).

Next, the arcsin square root transforms of the listener-averaged accuracies in the original vowel restoration rates (identifications) are submitted, separately for the C-set /be/, /bi/, /bu/, /ge/, /gi/, and /gu/, to two-sided one-sample *t* tests with the expected value of the arcsin square root transform of .5, with Holm corrections of the significance values.

Given the **production presuppositions**, **DI-sensitivity** leads us to expect significantly above-chance restorations with /gi/ stimuli at least (and possibly with /ge/ stimuli). In contrast, **coarticulation insensitivity** leads us to expect /u/ identifications throughout, and hence the original vowel restorations should be significantly below chance for all of /be/, /bi/, /ge/ and /gi/ (below-chance significance in the *t* tests with all the stimuli).

Furthermore, given that the same listeners participated in Experiment 8 shortly after participating in Experiment 5, whatever differences across the two experiments could be interpreted as being due to an order effect, rather than being due to presumably improved velar burst stimuli in Experiment 8. To see which is the case, the listener-average cross-experimental difference rates with /be/, /bi/, /ge/ and /gi/ stimuli are submitted to a two-way repeated-measure ANOVA, with 'consonant' and 'vowel' being the between-subject factors. If there are not enough cross-experimental differences, the difference scores with all of the four kinds of stimuli should be close to zero. If the cross-experimental differences are due to an order effect, the order effect would distribute equally or randomly across the four kinds of stimuli, and hence the main effects of 'consonant' and 'vowel', as well as their interaction, should be non-significant. In contrast, if the cross-experimental differences are due to the improved quality of velar bursts,

Table 6.17: The numbers of responses to the C-set /b/ stimuli.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	1	2	0	1	12	16
	/i/	0	4	4	0	8	16
	/u/	0	0	0	0	16	16
Total		1	6	4	1	36	48

Table 6.18: The numbers of responses to the C-set /g/ stimuli.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	1	6	0	9	16
	/i/	0	0	13	0	3	16
	/u/	0	0	0	0	16	16
Total		0	1	19	0	28	48

the main effect of 'consonant' should be significant; furthermore, given the **production presuppositions**, coarticulation sensitivity improvement should be more manifest with /gi/ than with /ge/, and hence the main effect of 'vowel' as well as the interaction between 'consonant' and 'vowel' should be significant.

6.4.2 Results and Discussion

The total numbers the responses to the voiced stimuli are summarized in Tables 6.17–6.20.

The patterns of responses to voiced and voiceless stimuli did not differ significantly for any of the place-vowel combinations, no matter whether the tests conducted were χ-squared tests or Fisher's exact tests.[20]

Unlike Experiment 5, Japanese listeners exhibited successful exploitation of /i/ coarticulation within /gi/ bursts, while exhibiting poor exploitation of /e/ coarticulation within /ge/ bursts, as seen in the original vowel restoration rates with the **C-set** /be/, /bi/, /bu/, /ge/, /gi/, and /gu/

[20]For /e/-coarticulated bilabials, $\chi^4(3) = 2.24$, $p = .524$ (χ-squared), $p = 1$ (Fisher's); for /i/-coarticulated bilabials, $\chi^2(2) = 2.2588$, $p = .323$; for /u/-coarticulated bilabials, the response patterns were completely the same across voiceless and voiced stimuli; for /e/-coarticulated velars, $\chi^2(2) = 2.7273$, $p = .256$ (χ-squared), $p = .252$ (Fisher's); for /i/-coarticualted velars, $\chi^2(2) = 2.037$, $p = .361$ (χ-squared), $p = .600$ (Fisher's); for /u/-coarticulated velars, $\chi^2(1) = 0$, $p = 1$ (χ-squared), $p = 1$ (Fisher's).

Table 6.19: The numbers of responses for the C-set /p/ stimuli.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	3	0	0	13	16
	/i/	0	1	6	0	9	16
	/u/	0	0	0	0	16	16
Total		0	4	6	0	38	48

Table 6.20: The numbers of responses for the C-set /k/ stimuli.

		responses					Total
		/a/	/e/	/i/	/o/	/u/	
stimuli	/e/	0	0	3	0	13	16
	/i/	0	1	14	0	1	16
	/u/	1	0	0	0	15	16
Total		1	1	17	0	29	48

summarized in Table 6.21.

The arcsin square root transforms of the listener-averaged original vowel restoration rates with the /be/, /bi/, /ge/, and /gi/ stimuli were submitted separately to two-sided one-sample t tests, with the expected value of the arcsin square root transform of .2 (a chance level in a five-choice task) and with Holm corrections of significance values.[21] The vowel restoration rates were not significantly better than chance with /be/ [$t(7) = -2.079$, $p = .076$ uncorrected], or with /bi/ [$t(7) = -.338$, $p = .745$ uncorrected]. However, the original vowel restoration rate was significantly better than chance with /gi/ [$t(7) = 5.654$, $p = .004$ if corrected; $p = .001$ if uncorrected] and was significantly worse than chance with /ge/ [$t(7) = -3.723$, $p = .037$

Table 6.21: Mean original vowel restoration rates and S. D. with the C-set stimuli (to be employed in the analyses) in Experiment 8 ($N = 8$).

	be	bi	bu	ge	gi	gu
mean	.1250	.2500	1.0000	.0625	.8125	1.0000
S. D.	.2315	.3780	.0000	.1768	.2588	.0000

[21] Since the vowel restoration rates were 1.000 for /bu/ and /gu/, only four tests were conducted. However, the Holm corrections were made on the assumption that six tests were conducted; significance under this assumption means significance under the assumption that four tests were conducted.

	/be/	/bi/	/ge/	/gi/
M	.0312	-.0312	.0313	.7500
SD	.2086	.3391	.2086	.2673

Table 6.22: The 'Experiment 8 minus Experiment 5' difference scores; $N = 8$.

if corrected; $p = .007$ if uncorrected]. The observation that the original vowel was restored from /gi/ bursts but not from other bursts is a result in line with **DI-sensitivity** coupled with the **production presuppositions**.

The observation of the significant success with /gi/ clearly supports **DI-sensitivity**. Thus, in contrast to Experiment 5, we have obtained identification-based evidence with stimuli produced by an American English speaker in Experiment 8, just as expected.

Note also that the same listeners participated in Experiment 8 shortly after participating in Experiment 5; thus the differences between Experiment 5 and Experiment 8 could alternatively be interpreted as an order effect, rather than in terms of the qualities of the velar bursts. To examine which of the presumably improved velar bursts or an order effect is responsible for the cross-experimental differences, for each of the /be/, /bi/, /ge/, and /gi/, the listener-averaged vowel restoration rates in Experiment 5 were subtracted from the corresponding rates in Experiment 8, the results of which are illustrated in Table 6.22. The difference scores were submitted to a two-way repeated-meaasure ANOVA, with 'consonant' and 'vowel' being the between-subject factors.[22] The results clearly supported the 'improved velar bursts' interpretation over the 'order effect' interpretation; the main effect of 'consonant' was highly significant, reflecting better vowel restorations with /g/ than with /b/ [$F(1,\ 7) = 42.476,\ p < .001$], indicating /g/'s greater contribution to the cross-experimental difference than /b/; the main effect of 'vowel' was also highly significant reflecting better /i/ restorations than /e/ restorations [$F(1,\ 7) = 32.495,\ p = .001$], suggesting /i/'s greater contribution to the cross-experimental

[22] Since some listeners exhibited less accuracy with non-/gi/ stimuli (but not with /gi/ stimuli) in Experiment 8 than in Experiment 5, resulting in negative 'difference scores', arcsin square root transformation was impossible. However, the observation of negative 'difference scores' with non-/gi/ stimuli but not with /g/ stimuli is consistent with the conclusion obtained from the ANOVA.

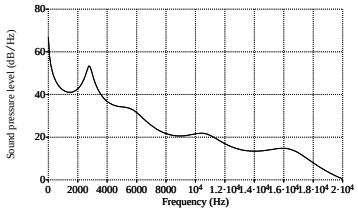


Figure 6.8: The LCP spectrum of the /ke/ burst in Experiment 8.

difference than /e/; their interaction was also highly significant [$F(1,\ 7) = 18.939$, $p = .003$], and examinations of the simple main effects revealed that the interaction was due to a greater effect with /gi/ than with /bi/ [$F(1,\ 7) = 28.974$, $p = .001$] and than with /ge/ [$F(1,\ 7) = 42.563$, $p < .001$], rather than to /be/ vs. /ge/ [$F(1,\ 7) = .000$] or /be/ vs. /bi/ [$F(1,\ 7) = .368$, $p = .563$]. Thus the ANOVA result clearly supports the idea that the cross-experimental difference is due to the improved /i/ restoration with the /gi/ stimuli in Experiment 8. Such an asymmetry would make sense under the 'improved velar burst quality' interpretation but not under the 'order effect' interpretation.

To summarize the results of Experiment 8, with the velar bursts good enough by an American speaker, effects of **DI-sensitivity** have been observed as an above-chance tendency for successful /i/ restoration from /gi/ bursts.

However, one possible problem remains. Indeed, the significant failure of /e/ identification with /ge/ stimuli is compatible with the expectation from the spectral shape. However, the rate of /i/ responses to /ke/ stimuli was also significantly *lower* than chance, if no multiple comparison correction is applied [$t(7) = -2.376$, $p = .049$, two-sided], which is rather unexpected from **DB-sensitivity** and the spectral shape of the /ke/ bursts shown in Figure 6.8. Note that the listeners in Experiment 8 are exactly those listeners in Experiment 5, in which above-chance /i/ identifications against /ke/ stimuli failed to be observed through a one-sample *t* test. Also note that, in both experiments, listeners were tested with voiceless stimuli mixed with voiced stimuli, in contrast to Experiments 1–2, in which separate groups of listeners were tested only

with voiceless stimuli and exhibited a clear tendency for /i/ identifications against /ke/ bursts. Thus the lack of a similar tendency in Experiments 5 and 8 could be attributed either to the particular group of listeners or to the experimental setting in which both voiceless and voiced stimuli are tested. Which interpretation should be adopted is not clear.

6.5 General Discussion

Dupoux et al. (2011) argued against the idea that listeners first perceive individual phonemes and subsequently repair the resulting phoneme strings according to phonotactics (two-step models). Their argument is based on the observation of Japanese listeners' /i/ perception from consonants with coarticulation traces of /i/. If the observed /i/ perception is a result of phonotactic repair (epenthesis), it would argue for the coarticulation sensitivity in phonotactic repair, which should be impossible under two-step models.

Dupoux et al.'s (2011) argument crucially hinges on the assumption that the /i/ perception in question is a result of phonotactic repair, an assumption under which their result should be seen as evidence for coarticulation-sensitivity in epenthesis. However, the results of Experiments 1–4 supported the perceptual reality of **DB-sensitivity**, according to which a voiceless [C] with enough front coarticulation is perceived as a vowel-devoiced version of /Ci/, which means that Japanese listeners' /i/ perception in such cases is a result of phonemic categorization, rather than epenthesis. Since Dupoux et al. conflated voiceless and voiced consonant stimuli, the observed /i/ perception by Japanese listeners must have partially come from **DB-sensitivity**, in which case their results do not necessarily argue against two-step models. Coarticulation sensitivity of phonemic categorization is not relevant in the evaluation of two-step models; in order to arrive at the conclusion that two-step models should be rejected, clear evidence for coarticulation sensitivity *in epenthesis* (**DI-sensitivity**) is needed. In the case of Japanese listeners, coarticulation-sensitive vowel perception (rather than the coarticulation-insensitive default /u/

epenthesis) based on *voiced* consonants would constitute such evidence, so Experiment 5–8 examined Japanese listeners' coarticulation (in)sensitivity in vowel perception based on voiced stops.

In Experiment 5, which was an identification experiment, evidence for **DI-sensitivity** (in the form of above-chance successful restoration of coarticulated non-/u/ vowels) failed to be obtained.

Given the assumed non-phonological (auditory) nature of **DI-sensitivity**, discrimination experiments were assumed to be more likely to succeed in eliciting its effects. Thus two AX discrimination experiments, with 1,500 ms ISI's (Experiment 6) and with 250 ms. ISI's (Experiment 7), were conducted. If **DI-sensitivity** is real, we would expect (i) above-chance successful discriminations, and (ii) improved discriminations with shorter ISI's. Above-chance successful discriminations were observed with /ge/–/gu/ and /gi/–/gu/ pairs (at least) in Experiment 7, and indeed improved discrimination with a shorter ISI was observed with velar pairs in a cross-experimental comparison between Experiments 6–7. Thus discrimination-based evidence for **DI-sensitivity** was obtained, but we still lacked identification-based evidence.

Assuming that voiced velar bursts by Japanese speakers, employed as stimuli in those three experiments, were rather poor in quality, another identification experiment (Experiment 8) was conducted with stimuli produced by an American English speaker; clear evidence for /i/ perception based on /gi/ bursts was observed, supporting the 'poor quality burst' interpretation of the results of Experiments 5-7 on the one hand, and the perceptual reality of **DI-sensitivity** on the other.

The support for the perceptual reality of **DI-sensitivity** constitutes evidence for Dupoux et al.'s (2011) conclusion that two-step models should be rejected; indeed, with voiceless and voiced consonant stimuli conflated, Dupoux et al.'s results should be seen as being partially due to **DB-sensitivity**, but **DB-sensitivity** is not the whole story; their results were probably also due to **DI-sensitivity**.

However, that in turn means that Japanese listeners' coarticulation sensitivity as observed by Dupoux et al. (2011) is a rather overestimation of **DI-sensitivity**, because their stimuli did contain /i/-coarticulated voiceless consonants (including /k/), for which the listeners' /i/ perception must have been helped by **DB-sensitivity**. Furthermore, the current results also suggest that, if they had employed voiced stops produced by Japanese speakers (rather than the utterances by a French speaker they employed), they could have failed to observed the sensitivity.

Chapter 7

Experiments 9–10

The results of the identification experiments reported above suggest that fronted velar bursts tend to give rise to /i/ identification (particularly when the bursts are voiceless). A natural interpretation of such results would be that the bursts were perceived as /ki/ or /gi/; if the bursts were perceived simply as /k/ or /g/, /i/ would be chosen only by chance and hence we would not have observed a significant tendency for /i/ identification. However, strictly speaking, a natural interpretation is not necessarily an inevitable interpretation. Logically it would be possible that listeners were sensitive not only to (i) front coloring within the bursts but also to (ii) the absence of "vowel" portions, in which case the bursts would have sounded like "fronted /k/ or /g/ bursts" rather than "/ki/ or /gi/"; listeners could have thought that "no vowel" was the best answer but had to pick up the second best answer (/ki/ or /gi/) because the best answer was not given as a possible choice. Since the purpose of the identification experiments was to examine Japanese listeners' /CV/ perception from bursts, it would be better if such a possibility is also examined, with further identification experiments with the "no vowel" option. Experiments 9–10 are meant to be such identification experiments.

Indeed, having "no vowel" as an additional response option has its own disadvantage. For Japanese speakers/listeners, the notions of "consonant" and "vowel" are rather foreign, which

they typically learn and talk about only in foreign language classrooms (which in turn put very little emphasis on the language's phonetics or phonology). Thus a task that explicitly mentions such notions as 'vowel' or 'consonant' (rather than the mora-based 'dan') could give a somewhat artificial impression to the listeners.[1]

Thus the previous identification experiments on the one hand, and Experiments 9–10 on the other, are complementary to each other; the task is more natural in the previous experiments than in Experiments 9–10, while Experiments 9–10 would distinguish what the the previous identification experiments did not. The same conclusion obtained in both would be trustworthy.

7.1 Experiment 9

Experiment 9 mimics Experiment 5 except that (i) only the female speaker's utterances are employed as stimuli, and (ii) the additional "no vowel" response option is added. (Due to the sixth response option and to the software availability, the experimental software was switched from E-prime to Praat.)

7.1.1 Method

Stimuli

The Japanese female speaker's utterances employed in Experiment 5; each utterance was presented to the listeners twice.

Participants

14 Japanese native listeners with no known hearing disability participated in the experiment for course credit at the "BA in English" and "MA in English" programs at Hosei University, Tokyo

[1] In fact, that was why the task in the previous identification experiments were a forced choice of the 'dan', rather than the 'vowel'.

(nine males and five females; mean age = 22, S.D. = 1.23).[2] None of them had participated in the previous experiments.

Procedure

The experiment was controlled by Praat (run on Fujitsu ESPRIMO D750/A with Microsoft Windows 7, 32 bit) and conducted in a computer lab. The participants were instructed on the computer screen (Fujitsu VL-193SEL) in Japanese that they would hear 'eXma' through a headset (ELECOM MS-HS67BK), and that their task was to identify the vowel portion of 'X', with the response options /a/, /i/, /u/, /e/, /o/, and 'no vowel'. The participants responded by clicking the spaces on the computer screen each allocated to the six response options. The experiment had no practice session.

Statistical Analyses

Only the results with those C-set stimuli in which the medial vowels are either /e/, /i/ or /u/ are employed in the analyses; all the other stimuli are treated as fillers.

First, the patterns of the responses from the C-set voiced and voiceless stimuli are compared to each other by means of χ-squared tests and/or Fisher's exact tests, separately for each place-vowel combination, so that we can see how listeners' responses do or do not differ across voiceless and voiced stimuli.

Next, the listener-averaged 'no vowel' response rates (for both voiceless and voiced stimuli), the listener-averaged /i/ response rates (for voiceless stimuli), and the listener-averaged original vowel restoration rates (for voiced stimuli), are examined.

In an identification experiment with the "no vowel" option, the question of how C-set stimuli are perceived could by interpreted in two different ways: (i) which of the six response options they tend to choose (the one-task interpretation), or (ii) whether they perceive a vowel,

[2]The participants in Experiments 1–8 were students at various undergraduate and/or postgraduate programs. In contrast, the participants in Experiments 9–10 were all students at the "English" programs.

and when they do, which vowel they tend to choose (the two-tasks interpretation). Under the one-task interpretation, the question would be understood as a single question, but under the two-tasks interpretation, the question would be understood as consisting of two sub-questions.

The two-tasks interpretation seems to have two advantages over the one-task interpretation. For one thing, the /a/, /e/, /i/, /o/, and /u/ options on the one hand, and the 'no vowel' option on the other, are qualitatively different options; the two-tasks interpretation respects this qualitative difference, which the one-task interpretation neglects. For another, for those listeners who perceived some vowel but yet had some difficulty categorizing the percept into one of the five Japanese vowels, the task would indeed consist of the decision of the presence or the absence of a vowel followed by the decision of categorizing the vowel percept; the two-tasks interpretation respects this two-step nature of the task, which the one-task interpretation does not.

However, the two-tasks interpretation assumes that listeners first classified the six options into 'some vowel' responses and 'no vowel' responses, an assumption which will probably be correct for some listeners, but not for every listener. For those listeners who do not classify the six options in this way (including those who did not pay attention to the stimuli or the task in the first place), the two-tasks interpretation is simply wrong; rather, the one-task interpretation should be taken. Thus the choice from the one-task and the two-tasks interpretations seems to depend on the particular listeners. Clearly, we do not know which interpretation is more appropriate for which listeners.

In the following statistical analyses, the one-task interpretation is primarily adopted. This decision is based on two considerations.

First, when the 'no vowel' response rates are examined, the chance level would be 1/6 under the one-task interpretation and 1/2 under the two-tasks interpretation. Note that 'no vowel' response rates significantly larger than chance would argue against the idea that CV's are indeed perceived from bursts. That means that setting the chance level for 'no vowel' to

be 1/6 would result in a rather tougher test for the idea of CV perception from bursts, than setting the chance level to be 1/2 (because the 'no vowel' rates should be more easily dictated to be significant if the expected value is set to 1/6 than if it is set to 1/2). If the idea of CV perception from bursts survives the tougher test under the one-task interpretation, it would certainly survive the looser test under the two-tasks interpretation.

The second consideration comes from the practical computational consideration. Note that the primary question in this thesis, as stated in Chapter 4, is what vowel does or does not function as a default-override in CV percepts, not whether CV percepts are obtained from bursts. That is, our primary interests are in whether the /i/ responses rates for voiceless stimuli on the one hand, and the original vowel restoration rates for voiced stimuli on the other, are above chance. Under the one-task interpretation, the listener-averaged rates as a whole can simply be employed and compared to 1/6. However, under the two-tasks interpretation, the computation becomes rather complicated. First, the 'no vowel' responses have to be removed from the data, and only the remaining portions of the data should be employed in the one-sample *t* tests, with the chance level being (the arcsin square root transform of) 1/5 (a chance level for a forced choice from five options). However, each stimulus is played twice, which means that (A) listeners who chose 'no vowel' twice, (B) those who chose it only once, and (C) those who did not choose it at all, should all be expected. For example, assume that listener B chose 'no vowel' once and chose /i/ once. Also assume that listener C chose /i/ twice. If we simply take listener-averaged /i/ response rates, we would end up giving an equal status to B's and C's /i/ response rate; in order to avoid this, we would have to weight their responses in some way, a complication not required under the one-task interpretation.

Thus the one-step interpretation is primarily adopted in the statistical analyses, and the two-step interpretation is appealed to only when necessary. That means that the listener-averaged 'no vowel' response rates (for both voiceless and voiced stimuli), the listener-averaged /i/ responses rates (for voiceless stimuli), and the listener-averaged original vowel restoration rates

Table 7.1: The numbers of responses for the C-set /p/ stimuli in Experiment 9.

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	0	1	0	14	13	28
	/i/	0	0	9	0	8	11	28
	/u/	0	0	0	0	16	12	28
Total		0	1	10	0	38	36	84

Table 7.2: The numbers of responses for the C-set /k/ stimuli in Experiment 9.

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	0	14	1	5	8	28
	/i/	0	1	19	0	3	5	28
	/u/	0	0	0	0	20	8	28
Total		0	1	33	1	28	21	84

(for voiced stimuli), are all submitted to two-sided one-sample *t* tests with the expected value being 1/6 (all under arcsin square root transforms), with Holm corrections separately for each medial consonant.

As for the latter two measures, both **DB-sensitivity** and **DI-sensitivity** predicts an above-chance tendency for /i/ identification with the /i/-coarticulated /k/ burst; **DB-sensitivity** predicts the possibility of an above-chance tendency for /i/ identification with the /e/-coarticulated /k/ burst, which is rather unexpected from **DI-sensitivity**. On the other hand, indeed a significant tendency for non-default /i/ identifications against /i/-coarticulated /g/ bursts would support **DI-sensitivity**, in Experiment 9 we expect that listeners should not exhibit such a tendency with the /i/-coarticualted /g/ burst, because the stimuli are produced by a Japanese speaker; if, as suggested by the result of Experiment 5, Japanese speakers' voiced bursts are too poor in quality in inducing coarticulation-sensitive identifications, original vowel restoration failure with voiced stimuli is expected in Experiment 9 not because **DI-sensitivity** is not real but rather because the stimuli are not the right kinds.

Table 7.3: The numbers of responses for the C-set /b/ stimuli in Experiment 9.

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	1	4	1	9	13	28
	/i/	0	1	9	0	7	11	28
	/u/	0	0	0	0	19	9	28
Total		0	2	13	1	35	33	84

Table 7.4: The numbers of responses for the C-set /g/ stimuli in Experiment .

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	0	8	0	10	10	28
	/i/	1	0	8	0	11	8	28
	/u/	0	0	0	0	24	4	28
Total		1	0	16	0	45	22	84

7.1.2 Results and Discussion

First the response times were compared to the durations of the sound files in order to eliminate those responses that were made before the end of the stimuli. No response was eliminated through this procedure.[3]

The total numbers of responses are summarized in Tables 7.1–7.4.

For each place-vowel combination, the response patterns were compared across the voicing dimension. Intuitively, the most remarkable difference between responses to voiced and voiceless stimuli is that /i/ responses were the majority responses to the /ki/ stimulus but not to the /gi/ stimulus. This intuition was confirmed by the results of χ-squared tests and Fisher's exact tests, which were significant only with /i/-coarticulated velars [$\chi(4) = 11.7452$, $p = ..019$; Fisher's, $p = .007$]. According to the result of the χ-squared test conducted on responses to the /ki/ and /gi/ stimuli, the largest residual among the responses to /gi/ came from /u/ responses (1.512), while the largest residual among the responses to /ki/ came from /i/ responses (1.407), suggesting that the significance is largely due to the contrasting tendencies of voiced

[3]In the previous identification or discrimination experiments, E-prime could be set up so that responses before the end of the stimuli were impossible in the first place and hence this procedure was not necessary.

Table 7.5: Mean 'no vowel' response rates in Experiment 9 ($N = 14$).

	pe	pi	pu	ke	ki	ku
mean	.4643	.3929	.4286	.2857	.1786	.2857
S.D.	.4144	.4010	.4746	.3780	.2486	.4258

	be	bi	bu	ge	gi	gu
mean	.4643	.3929	.3214	.3571	.2857	.1429
S.D.	.4986	.4463	.4210	.4127	.4258	.3631

Table 7.6: Mean /i/-identification rates with voiceless stimuli in Experiment 9 ($N = 14$).

	pe	pi	pu	ke	ki	ku
mean	.0357	.3214	.0000	.5000	.6786	.0000
S. D.	.1336	.3725	.0000	.4385	.4210	.0000

and voiceless stimuli to elicit /u/ and /i/ responses respectively; thus we have again observed more coarticulation sensitivity in the voiceless than in the voiced case.[4]

Next, the listener-averaged 'no vowel' response rates, summarized in Table 7.5, were examined. For each consonant-vowel combination, the listener-averaged 'no vowel' response rates were arcsin square root transformed and submitted to two-sided one-sample t tests with the expected value of the arcsin square root transform of 1/6. With no consonant-vowel combination, the 'no vowel' response rate significantly differ from chance ($= 1/6$); thus the results do no constitute evidence against the idea of CV percepts from bursts.[5]

Next, the /i/-identification rates with voiceless stimuli, summarized in Table 7.6 were examined. For each place-vowel combination, the listener-averaged /i/ identification rates were arcsin square root transformed and submitted to two-sided one-sample t tests with the expected

[4]The results with the other stimuli were as follows: with /e/-coarticulated bilabials, $\chi(4) = 4.887$, $p = .2991$ (χ-squared), $p = .236$ (Fisher's); with /i/-coarticulated bilabials, $\chi(3) = 1.0667$, $p = .785$ (χ-squared), $p = 1$ (Fisher's); with /u/-coarticlated bilabials, $\chi(1) = .3048$, $p = .5809$ (χ-squared, with Yate's continuity correction), $p = .5815$ (Fisher's); with /e/-coarticalted velars, $\chi(3) = 4.5253$, $p = .21$ (χ-squared), $p = .186$ (Fisher's); with /u/-coarticulated velars, $\chi(1) = .9545$, $p = .329$ (χ-squared, with Yate's continuity correction), $p = .329$ (Fisher's).

[5]For /pe/, $t(13) = 1.775$, $p = .099$; for /pi/, $t(13) = 1.168$, $p = .264$; for /pu/, $t(13) = 1.268$, $p = .227$; for /ke/, $t(13) = .178$, $p = .861$; for /ki/, $t(13) = -1.342$, $p = .203$; for /ku/, $t(13) = .158$, $p = .877$; for /be/, $t(13) = 1.475$, $p = .164$; for /bi/, $t(13) = 1.049$, $p = .313$; for /bu/, $t(13) = .477$, $p = .641$; for /ge/, $t(13) = .811$, $p = .432$; for /gi/, $t(13) = .158$, $p = .877$; for /gu/, $t(13) = -1.287$, $p = .221$. Note that no multiple comparison correction is applied.

value being the arcsin square root transform of $= 1/6$. (The results with /u/-coarticulated stimuli were excluded from those analyses, because they all reached the floor.)

With the /i/-coarticulated /p/ stimulus, the /i/ identification rate did not differ from chance [$t(13) = .540$, $p = .599$, uncorrected]. With the /e/-coarticulated /p/ stimulus, the /i/ identification rates were highly significantly *below* chance [$t(13) = -6.496$, $p < .001$]. In contrast, with the /i/-coarticulated /k/ stimulus, the /i/ identification rate was significantly *above* chance [$t(13) = 3.652$, $p = .006$ if corrected, $p = .003$ if uncorrected], while with the /e/-coarticulated /k/ stimulus, the /i/ identification rate was marginally above chance [$t(13) = 1.982$, $p = .069$].[6] Again, given the **production presuppositions**, **DB-sensitivity** predicts that the /i/-identification rates should be higher with /ke/-stimuli than with /pe/-stimuli, a prediction not made by **DI-sensitivity**, the listener-averaged /i/ response rates were compared across the /ke/ and /pe/ stimuli with a two-sided paired *t* test, and the result confirmed the prediction of **DB-sensitivity** [$t(13) = 4.192$, $p = .001$]. Further note that, although the *t* test was two-sided, we are testing a directional **DB-sensitivity** hypothesis that /i/ identification rates should be better than chance, rather than a non-directional hypothesis that they should simply differ from chance; if a directional hypothesis justifies a one-sided test, then the significance value of the /i/-identification rates with the /ke/ stimulus would be halved [$p = .035$], which is clearly significant. Thus the identification results with voiceless stimuli are best interpreted by assuming the reality of **DB-sensitivity**.

Finally the listener-averaged rates of original vowel restorations, summarized in Table 7.7, are examined. The listener-averaged original vowel restoration rates were arcsin square root transformed and submitted to two-sided one-sample *t* tests with the expected value being the transform of the chance level ($= 1/6$), with Holm corrections separately for each place. (The

[6] Because the /i/-identification rates against /ku/ reach the floor, two *t* tests were conducted for /k/ stimuli. The lower uncorrected *p* value came from the /ki/ result, and hence the *p* value should be multiplied by two, according to the Holm procedure. The corrected value was significant, so the Holm procedure dictates that the next lowest *p* value (i.e., the one obtained with /ke/) should be corrected by multiplying the number of the remaining comparisons. However, the number of the remaining comparisons is one, so the correction by multiplying one returns the original *p* value; hence, with /ke/, the uncorrected *p* value is the corrected *p* value.

Table 7.7: Mean original vowel restoration rates with voiced stimuli in Experiment 9 ($N = 14$).

	be	bi	bu	ge	gi	gu
mean	.0357	.3214	.6786	.0000	.2857	.8571
S. D.	.1336	.4210	.4210	.0000	.3780	.3631

result with the /ge/ stimulus were excluded from the analyses, because it reached the floor.) Naturally, the original vowel restoration rates were highly significantly above chance with /bu/ [$t(13) = 3.652$, $p = .006$ if corrected, $p = .003$ if uncorrected] and with /gu/ [$t(13) = 19.504$, $p < .001$]; such results are expected whether or not listeners are coarticulation sensitive. The restoration rates were highly significantly *below* chance with /be/ [$t(13) = -6.496$, $p < .001$]; in conjunction with the floor result with /ge/, this suggests that almost no coarticulation sensitivity was exhibited for /e/-coarticulation within voiced bursts. Finally, the restoration rates did not differ from chance with /bi/ [$t(13) = .477$, $p = .641$] and with /gi/ [$t(13) = .178$, $p = .861$].

Thus the overall results obtained in the previous identification experiments without a 'no vowel' option are also obtained in Experiment 9. More specifically, (i) an observation of a clear tendency for /ki/ identifications from /i/-coarticulated /k/, (ii) an observation of some tendency for /ki/ identifications from an /e/-coarticulated /k/, and (iii) the failure to obtain evidence for coarticulation sensitivity with voiced bursts, were all replicated. In conjunction with the previous identification experiments, the results of Experiment 9 thus strengthen the conclusion that, with stimuli produced by Japanese speakers, identification-based experiments could be obtained only for **DB-sensitivity**.

7.2 Experiment 10

Both in Experiment 5 and in Experiment 9, Japanese speakers' productions were employed as stimuli and identification-based evidence for **DI-sensitivity** failed to be obtained. The identification-based evidence for **DI-sensitivity** we have obtained came from Experiment 8,

in which an American English speaker's productions were employed as stimuli. If an identification experiment with a 'no vowel' response option should complement the results of identification experiments without such a response option, then an identification experiment with that response option and with American English speaker's productions should also be conducted. Experiment 10 is meant to be such an experiment.

7.2.1 Method

Stimuli

The stimuli employed in Experiment 8 were re-used; each stimulus was played only once.

Participants

The participants in Experiment 9 also participated in Experiment 10 (shortly after Experiment 9).

Procedure

The same as Experiment 9, except that the participants were instructed to hear 'eXta' (rather than 'eXma') and that their task was to identify the vowel portion of 'X'.

Statistical Analyses

Basically, the same as Experiment 9.

7.2.2 Results and Discussion

Again, the response times were compared to the durations of the sound files to make sure that no response was made before the end of the stimuli. No response was eliminated through this procedure.

Table 7.8: The numbers of responses for the C-set /p/ stimuli in Experiment 10.

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	0	1	0	10	3	14
	/i/	0	0	1	0	11	2	14
	/u/	0	0	0	0	11	3	14
Total		0	0	2	0	32	8	42

Table 7.9: The numbers of responses for the C-set /k/ stimuli in Experiment 10.

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	0	1	0	9	4	14
	/i/	0	0	10	0	2	2	14
	/u/	0	0	0	0	10	4	14
Total		0	0	11	0	21	10	42

The total numbers of responses are summarized in Tables 7.8–7.11. Again, /u/-coarticulated stimuli only elicited /u/ responses or 'no vowel' responses.[7]

Again, the response patterns for voiced and voiceless stimuli were compared with χ-squared tests and Fisher's exact tests. Our primary interest is in whether the comparison reaches significance for /i/-coarticulated velars, because /i/-coarticulated voiceless velars clearly tend to elicit the non-default /i/ identifications; however, the comparison was far from being significant with /i/-coarticualted velars [$chi(2) = 1.3961$, $p = .4976$ (χ-squared); $p = .5758$ (Fisher's)], suggesting that coarticulation sensitivity did not differ a lot between /ki/ and /gi/. All the other comparisons failed to reach significance.[8]

Next, the listener-averaged 'no vowel' response rates, summarized in Table 7.12, are examined. The listener-averaged 'no vowel' response rates were arcsin square root transformed and submitted to two-sided one-sample *t* tests with the expected value of 1/6. As with Exper-

[7] Since each stimulus was played twice in Experiment 9 but only once in Experiment 10, the total number of the responses to each stimulus in Experiment 10 is half of that in Experiment 9.

[8] For /e/-coarticulated bilabials, $\chi(4) = 4.0294$, $p = .402$ (χ-squared), $p = .4061$ (Fisher's); for /i/-coarticulated bilabials, $\chi(2) = 3.9373$, $p = .1396$ (χ-squared), $p = .1843$ (Fisher's); for /u/-coarticulated bilabials, $\chi(1) = .2121$, $p = .6451$ (χ-squared, with Yate's continuity correction), $p = 1$ (Fisher's); for /e/-coarticulated velars, $\chi(3) = 5.5897$, $p = .1334$ (χ-squared), $p = .1318$ (Fisher's); for /u/-coarticlated velars, $\chi(1) = 0$, $p = 1$ (χ-squared, with Yate's continuity correction), $p = 1$ (Fisher's).

Table 7.10: The numbers of responses for the C-set /b/ stimuli in Experiment 10.

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	1	0	1	7	5	14
	/i/	0	0	4	0	6	4	14
	/u/	0	0	0	0	11	3	14
Total		0	1	4	1	24	12	42

Table 7.11: The numbers of responses for the C-set /g/ stimuli in Experiment 10.

		responses						Total
		/a/	/e/	/i/	/o/	/u/	'no vowel'	
stimuli	/e/	0	1	5	0	4	4	14
	/i/	0	0	7	0	3	4	14
	/u/	0	0	0	0	11	3	14
Total		0	1	12	0	18	11	42

iment 9, the 'no vowel' response rate was significantly higher than chance with no consonant-vowel combination; thus, again, the results do not constitute evidence against the idea of CV percepts from bursts.[9]

Next, /i/-identificaiton rates for voiceless stimuli, summarized in Table 7.13, were examined. The rates were arcsin square transformed and submitted to two-sided one-sample *t* tests with the expected value of the chance level ($= 1/6$). (The results from the /u/-coarticulated stimuli were excluded from the analyses because they reached the floor.) The /i/-identification

Table 7.12: Mean 'no vowel' response rates in Experiment 10 ($N = 14$).

	pe	pi	pu	ke	ki	ku
mean	.2143	.1429	.2143	.2857	.1429	.2857
S.D.	.4258	.3631	.4258	.4688	.3631	.4688

	be	bi	bu	ge	gi	gu
mean	.3571	.2857	.2143	.2857	.2857	.2143
S.D.	.4973	.4688	.4258	.4688	.4688	.4258

[9]For /pe/, $t(13) = -.470$, $p = .646$; for /pi/, $t(13) = -1.287$, $p = .221$; for /pu/, $t(13) = -.470$, $p = .646$; for /ke/, $t(13) = .144$ $p = .888$; for /ki/, $t(13) = -1.287$, $p = .221$; for /ki/, $t(13) = .144$, $p = .888$; for /be/, $t(13) = .673$, $p = .513$; for /bi/, $t(13) = .144$, $p = .888$; for /bu/, $t(13) = -.470$, $p = .646$; for /ge/, $t(13) = .144$, $p = .888$; for /gi/, $t(13) = .144$, $p = .888$; for /gu/, $t(13) = -.470$, $p = .646$. Note that no multiple comparison correction is applied.

Table 7.13: Mean /i/-identification rates for voiceless stimuli in Experiment 10 ($N = 14$).

	pe	pi	pu	ke	ki	ku
mean	.0714	..0714	.0000	.0714	.7143	.0000
S. D.	.2673	.2673	.0000	.2673	.4688	.0000

Table 7.14: Mean original vowel restoration rates with voiced stimuli in Experiment 10 ($N = 14$).

	be	bi	bu	ge	gi	gu
mean	.0714	.2857	.7857	.0714	.5000	.7857
S. D.	.2673	.4688	.4258	.2673	.5189	.4258

rates were significantly below chance with /pe/ [$t(13) = -2.748$, $p = .017$ uncorrected], /pi/ [$t(13) = 2.748$, $p = .033$ if corrected, $p = .017$ if uncorrected],[10] and with /ke/ [$t(13) = -2.748$, $p = .017$]. In contrast, the /i/-identification rate was significantly *above* chance with /ki/ [$t(13) = 3.564$, $p = .007$ if corrected, $p = .003$ if uncorrected]. Thus, as with Experiment 8 without the 'no vowel' response option, the American English speaker's /ki/ bursts, but not her /ke/ bursts, elicited /i/ identifications.

Finally, the original vowel restoration rates for voiced stimuli, summarized in Table 7.14, were examined. The original vowel restoration rates were arcsin square root transformed and submitted to two-sided one-sample *t* tests with the expected value of the transform of the chance level ($= 1/6$) and with Holm corrections separately for each place. Naturally, the restoration rates were significantly *above* chance with /bu/ [$t(13) = 4.552$, $p = .002$ if corrected, $p = .001$ if uncorrected] and with /gu/ [$t(13) = 4.552$, $p = .002$ if corrected, $p = .001$ if uncorrected]; the restoration rates were significantly *below* chance with /be/ [$t(13) = -2.748$, $p = .033$ if corrected, $p = .017$ if uncorrected] and with /ge/ [$t(13) = -2.748$, $p = .033$ if corrected, $p = .017$ if uncorrected]. With /bi/, the restoration rate was far from being significantly different from chance [$t(13) = .144$, $p = .888$], while with /gi/, the restoration rate did not differ from

[10] The precise significance value was very slightly smaller with /pi/ than with /pe/ and hence the Holm correction was applied only to the /pi/ results. However, if Bonferroni corrections were applied instead, that is, both to the /pe/ and the /pi/ results), both results are still significant.

chance either [$t(13) = 1.675$, $p = .118$].

Taken literally, the /gi/ result would suggest that, even when the stimuli were produced by an American English speaker, identification-based evidence for **DI-sensitivity** failed to be obtained if the 'no vowel' response option is added. Indeed, if the 'no vowel' responses were excluded from the analyses and a two-sided one-sample *t* test is conducted only on the epenthetic responses (with the expected value being the transform of $1/5$), as dictated by the two-tasks interpretation of the experimental task,[11] the restoration rate with the /gi/ stimulus becomes significantly higher than chance [$t(9) = 2.650$, $p = .026$, with no correction]. Thus it could be argued that the results of Experiment 8 seemed to support the reality of **DI-sensitivity** because listeners often did not perceptually epenthesize but were sensitive to coarticulation and hence picked up /gi/ as a second best choice, with the best choice 'no vowel' not being available.

However, note that the one-sample *t* test was two-sided, although in fact we are testing the directional hypothesis that original vowel restorations with /gi/-bursts should be significantly better than chance, rather than the non-directional hypothesis that they should simply differ from chance. If a directional hypothesis justifies a one-sided test, then the significance value with /gi/ should be halved, in which case the value becomes $p = .0589$, not very far from being significant. Thus the best interpretation of the /gi/ result would be that, although the addition of 'no vowel' response option indeed may make a difference, Japanese listeners' epenthetic perception exhibiting **DI-sensitivity** is indeed real; the effect of **DI-sensitivity** was, in contrast to that of **DB-sensitivity**, too weak to be detected easily with a conservative two-sided test.

7.3 General Discussion

The hypotheses tested were (i) /ki/ and /ke/ bursts should induce /i/ identification (**DB-sensitivity**), and (ii) /gi/ bursts should induce /i/ identification (**DI-sensitivity**). Both hypotheses are direc-

[11] Recall that, according to the two-tasks interpretation, the listeners first classify the six response options into the 'some vowel' group vs. 'no vowel' option, and only when the 'some vowel' group is chosen would they attempt to identify the stimulus with one of the five Japanese vowels.

tional, for which two-sided one-sample *t* tests are rather conservative; if one-sided tests are employed, all the predictions are confirmed (except the /ke/ result in Experiment 10) even when the sixth option of 'no vowel' is added.

Indeed, while conservative two-sided tests were enough to support the predictions concerning /ke/ (Experiments 1, 2) and /gi/ (Experiments 8) when the sixth option of 'no vowel' was not given, less conservative one-sided tests were needed to support them (/ke/ in Experiment 9 and /gi/ in Experiment 10). Thus it will have to be conceded that identification experiments without such an option (Experiments 1, 2, 5, and 8) may have overestimated the effects of **DB-sensitivity** and **DI-sensitivity**. However, overestimating the magnitude of an effect does not mean claiming the existence of an effect which is not real; with one-sided tests, we will have to concede too that, even with the sixth option of 'no vowel', people do tend to perceive /ki/ from /ke/ bursts and /gi/ from /gi/ bursts. Given that identification experiments with no such option (Experiments 1, 2, 5, 8) and those with such an option (Experiments 9–10) are complementary, and given that directional hypotheses justify one-sided tests, the fact that similar results were obtained from both kinds of identification experiments strengthens the conclusion that both **DB-sensitivity** and **DI-sensitivity** are real.

On the other hand, with one-sided tests, the rates of /i/-identification were significantly *above* chance in Experiment 9 (where the stimuli were uttered by a Japanese speaker) but were significantly *below* chance in Experiment 10 (where the stimuli were uttered by an English speaker). This contrast resembles the contrast between Experiment 5 (with stimuli by Japanese speakers) and Experiment 8 (with stimuli by an English speaker); the /i/ identifications against /ke/ did not significantly differ from chance in Experiment 5 but was significantly below chance (if uncorrected) in Experiment 8. Thus /ke/ bursts by an English speaker seems to lack coarticulation cues that count as 'front' for Japanese listeners; what acoustic characteristics make bursts front enough for Japanese listeners is not clear (which is beyond the scope of this thesis).

Chapter 8

Concluding Summary

In this final chapter, the main findings of this thesis are first summarized (section 8.1), after which their implications and some remaining problems are illustrated (section 8.2).

8.1 The Main Findings

This thesis asked which of (two versions of) one-step models, two-step models and lexicalist models should be adopted. This section summarize the arguments for (a specific version of) one-step models.

8.1.1 Against Lexicalist Models

Dupoux, Kakehi, Hirose, Pallier & Mehler (1999) demonstrated robust perceptual /u/ epenthesis by Japanese listeners. Dupoux et al. (2001) and Mazuka et al. (2011) argued for the sub-lexical nature of such phonotactic sensitivity. However, Fais et al. (2005) argued that such /u/ perception should be seen as assimilation of the stimuli to the attested sound patterns, a conclusion that suggests a lexicalist reduction of phonotactic sensitivity. Furthermore, although most of Dupoux et al.'s (2001) results were against a lexicalist account of phonotactic sensitivity, some seeming counterexamples were also observed.

Experiments 1–4 (as well as Experiments 5 and 9) confirmed the sublexical reality of **DB-sensitivity**, which leads Japanese listeners to phonemically categorize a voiceless consonant [C] with enough front coloring as /Ci/. Observations of /i/ perception from /i/-coarticulation traces within a voiceless consonant (Beckman & Shoji, 1984; Ogasawara, 2013; Ogasawara & Warner, 2009; Tsuchida, 1994) could be interpreted not only in terms of **DB-sensitivity** but also in terms of **DI-sensitivity**. However, /i/ perception from /e/-coarticulation traces within a voiceless consonant could only be interpreted in the former. Such /i/ perception from /e/-coarticulation traces was observed when the /e/-coarticulated voiceless consonant was immediately followed by a voiced consonant [m]; unless the sublexical reality of **DB-sensitivity** is admitted, such observations are hard to interpret.

If **devoicing-based coarticulated sensitivity** is real, Fais et al.'s (2005) observations on the one hand, and Dupoux et al.'s (2001) observations that seemed to run counter to the sublexical nature of phonotactic repair on the other, should be seen as instances of devoicing-based phonemic categorization, rather than as instances of phonotactic repair, and hence they should be regarded as simply irrelevant to the question of whether phonotactic repair is sublexical or lexical. Thus the supported reality of **DB-sensitivity** (Experiments 1–4) defends Dupoux et al.'s (2001) and Mazuka et al.'s (2011) claim of the sublexical nature of phonotactic sensitivity, by enabling us to regard alleged evidence against their claim (observed by Fais et al., 2005; Dupoux et al., 2001) as irrelevant. In other words, the reality of **DB-sensitivity** supported by the results of Experiments 1–4 (and Experiments 5 and 9) argue against lexicalist models with respect to phonotactic sensitivity, by defending Dupoux et al.'s (2001) and Mazuka et al.'s (2011) claim from the alleged counter-evidence.

Furthermore, because the confirmation of **DB-sensitivity** offered by the results of Experiments 1–4 (and Experiments 5 and 9) is the confirmation of its *sublexical* reality; thus the results of Experiments 1–4 (and Experiments 5 and 9) also argue against a lexicalist *reduction* of coarticulation sensitivity (although this conclusion validates the interpretation of Ogasawara

& Warner's 2009 results, according to which **DB-sensitivity** is real lexically too and hence a 'direct' route from the speech signal to the lexicon should be added to the Merge models of Norris et al., 2000 and Norris & McQueen, 2008).

Since neither of phonotactic sensitivity or coarticulation sensitivity is reducible to lexical sensitivity, then, lexicalist models should be rejected.

Furthermore, whether /CV/ perception is more difficult from vowel-devoiced stimuli than from non-devoiced stimuli was examined. Ogasawara & Warner's (2009) results already suggest that the nature of the task (sublexical vs. lexical) affects the effect of vowel devoicing, but Experiments 1–2 have shown that the amount or quality of coarticulation traces also affect the effect of vowel devoicing; no significant effect of vowel devoicing was observed with full velar burst stimuli in Experiment 1, which replicates Ogasawara & Warner's /i/ monitoring results, but a significant inhibitory effect was observed with bilabial bursts in Experiments 1–2 on the one hand, and with shortened velar bursts in Experiment 2 on the other, replicating Beckman & Shoji's (1984) (and Cutler et al.'s, 2009) results.

8.1.2 The Choice between One- and Two-step Models

With lexicalist models invalidated, the remaining candidates are one- and two-step models. In Chapter 3, it was pointed out that one-step models could be implemented in two different versions: the **suprasegmental matching** version, according to which the speech signal is matched against the acquired repertoire of suprasegmental units (such as syllables), rather than the repertoire of phonemes, and the **slot filling** version, according to which phonemes are perceived but only as fillers for slots in the structured frames of suprasegmental units. The two versions differ with respect to the perceptual status of sub-syllabic elements (assuming the suprasegmental units in questions are syllables); according to the former, such elements as phonemes are only recognized as a result of post-perceptual meta-analysis, whereas according to the latter, such elements are indeed perceived.

Dehaene-Lambertz et al. (2000), Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999), Dupoux et al. (2011) all seem to have considered only the **suprasegmental matching** version, and Dupoux, Kakehi, Hirose, Pallier, & Mehler (1999) and Dupoux et al. (2011) attempted to argue against the claimed evidence for listeners' perceptual access to sub-syllabic elements by Berenet et al. (2007), Kabak & Idsardi (2007), Moreton (2002), and Pallier et al. (1993). It was argued in Chapter 3 that (i) Berent et al.'s, Kabak & Idsardi's, Pallier et al.'s, and Moreton's results do not necessarily show listeners' perceptual access to sub-syllabic elements and hence are compatible with both versions of one-step models, but (ii) the results by Matthews & Brown (2004) indeed constitute evidence for listeners' perceptual access to sub-syllabic elements and hence argue against the **suprasegmental matching** version of one-step models. However, the (ii) conclusion only implies the invalidity of the **suprasegmental matching** version and is compatible with the **slot filling** version. Thus the (ii) conclusion leaves the choice between one- and two-step models open; it simply implies that, if one-step models should be adopted at all, it should be the **slot filling** version.

Thus the remaining candidates are two-step models on the one hand, and the **slot filling** version of one-step models on the other. Both admit phonotactic parsing, but they differ with respect to whether phonemic parsing (categorization) and phonotactic parsing are serial (two-step models) or parallel (the **slot filling** version of one-step models). Thus the choice from the two concerns the nature of the speech parser.

Since coarticulation-sensitive /i/ perception from voiceless consonants could well be attributed to **DB-sensitivity**, which could well be regarded either as coarticulation sensitivity in the 'first-step' phonemic categorization stage in two-step models, or as the phonemic categorization running in parallel with phonotactic parsing in the **slot filling** version of one-step models, the choice could not be made without examinations of coarticulation sensitivity in vowel perception that could not be attributed to **DB-sensitivity**, to be exerted in phonemic categorization, rather than in phonotactic repair. Thus Japanese listeners' coarticulation sensitivity

in vowel perception from voiced stops (**DI-sensitivity**) was examined in Experiments 5–10. Although evidence for such coarticulation sensitivity failed to be obtained in Experiments 5–6 and 9, it was indeed obtained in Experiments 7–8 and 10 (and a comparison between Experiment 6 and Experiment 7). This suggests that indeed Japanese listeners exhibit coarticulation sensitivity (not only in phonemic categorization but also) in phonotactic repair with the right conditions, which in turn suggests that (the **slot filling** version of) one-step models should be adopted, rather than two-step models.

Thus the **slot filling** version of one-step models should be adopted, rather than two-step models or lexicalist models (or the **suprasegmental matching** version of one-step models).

8.2 Implications and Remaining Problems

8.2.1 The Rich Perceptual Ontology and the Parallel Architecture of the Perceptual System

The thesis has argued for the **slot filling** version of one-step models. This version claims that phonemes and suprasegmental units (such as syllables) are perceived in parallel. This claim entails the perceptual reality of both segmental units (such as phonemes) and suprasegmental units (such as syllables), where the perceptual significance of the well-formedness of the latter (phonotactics) could not be reduced to the familiarity with attested patterns.

In addition, the experimental results in this thesis validates the suggested interpretation of Ogasawara & Warner's (2009) /i/ monitoring and lexical decision experiments, according to which a 'direct' route from the speech signal to the lexicon should be assumed in addition to the 'indirect' route from the speech signal to the lexicon mediated by sublexical processing.

According to such conclusion, the perceptual ontology is rich; phonemic segments, sublexical suprasegmental units, and words are all perceptually real. However, the perceptual system with such rich ontology does not necessarily implies slower perceptual processes than a per-

ceptual system with poor ontology (e.g., the **suprasegmental matching** version of one-step models, which deny the perceptual reality of phonemic segments), because phonemic parsing (phonemic categorization) and suprasegmental parsing (phonotactic repair) are assumed to run in parallel. Furthermore, a 'direct' route from the speech signal to the lexicon is also assumed, which presumably runs in parallel with the sublexical parsing. For a processing system in general, two kinds of computational complexity can be distinguished: *memory* complexity (what have to be memorized, i.e., ontology) and *time* complexity (the efficiency of computation). However, a system that is complex with respect to memory is not necessarily complex with respect to time.

8.2.2 Loanword Phonology and the Relation between the Two Kinds of Coarticulation Sensitivity

Another implication concerns loanword phonology. As is already clear, the reality of **DB-sensitivity** offers a straightforward perceptual account of the origin of the /ki, ku/ alternation in loanwords in Japanese. For example, *Texas* is adopted as /tekisasu/, rather than /tekusasu/, because the voiceless velar burst in the original word is fronted and hence ended up being phonemically categorized as /ki/ by Japanese listeners. Furthermore, its *sublexical* reality offers a straightforward account of the fact that a voiceless velar stop followed by an ash [æ] induce the insertion of a glide /j/ in loanwords (cf. Lovins, 1973:72, as cited by Irwin, 2011:97): /kjaQto/ 'cat', /kjaputeN/ 'captain', /sukjaN/ 'scan'). Presumably, the voiceless velar stops are fronted due to [æ]. Then *sublexical* **DB-sensitivity** should lead Japanese listeners to perceive the burst as /ki/, in spite of the immediately following *voiced* vowel [æ], That means that [kæ] should be rendered to /ki/ followed by /a/ (given that [æ] itself is assimilated to /a/). However, a glide /j/ is nothing but a transition from an /i/-like tongue position, and hence /kia/ and /kja/ are not very distinct. Glide insertion in examples such as /me:kjaQpu/ 'make up' (Takehiko Makino, p. c., 1996) can be explained in a similar manner; the velar stop induced /i/ perception,

and the perceived /i/ is combined with the first vowel of the next word ([ʌ], assimilated to /a/), resulting in glide insertion (/kja/).[1]

However, one caveat is necessary here. Labrune (2012) notes that not only /k/ but also /g/ tends to induce glide insertion before [æ] (e.g., /g̱jarari:/ 'gallery', /g̱jaNburu/ 'gamble', /g̱jaQpu/ 'gap'); according to him, /k, g/ in English are more fronted than /k, g/ in Japanese,[2] which leads Japanese listeners to the perception of /i/, in which case [gæ] should be perceived as /gia/, not very distinct from /gja/. We could attribute the assumed /i/ perception from a fronted *voiced* velar burst to **DI-sensitivity**.[3] However, that would mean that glide insertion after a voiceless velar stop could also be explained in terms of **DI-sensitivity**. Note that glide insertion could be caused only by **DI-sensitivity** after *voiced* stops but by both **DB-sensitivity** and **DI-sensitivity** after *voiceless* stops. Further note that effects of **DB-sensitivity** are presumably exerted more easily than effects of **DI-sensitivity**. Then we would expect that glide insertion in loanwords should be more often after *voiceless* velar stops than after *voiced* velar stops. Indeed, non-glide insertion between a *voiced* velar stop and [æ] do exist (e.g., /gasu/ 'gas', /gamu/ 'gam'), while I could come up with no instance of non-glide insertion between a *voiceless* velar stop and [æ]. However, to confirm such an expectation, a systematic search of loanwords, as well as the philological research on their origins, is needed, which has to be left for future research.

---

[1]It is widely known that voiceless stop phonemes in English get aspirated in syllable-initial positions. Thus it could be argued that the voiceless velar burst in *cat* is followed by a voiceless aspiration 'segment'. However, whether the velar stops are aspirated or not does not seem to be responsible for the successful exploitation of **DB-sensitivity**; for example, in *scan* or *make up*, the voiceless velar stop is presumably not aspirated, but yet we observe glide insertion (/sukj̱an/).

One related note concerning alleged phonotactic effects. It has long been known that, in the case of English listeners, the voiceless stop phoneme /C/ in /CV.../ uttered by English speakers tend to be perceived as voiceless if immediately preceded by /s/ (as in the original utterances) but as voiced if /s/ is deleted. The classical account is that this is due to English phonotactics (e.g., Mann & Repp, 1981b:1155). However, an informal observation suggests that both Japanese and Chinese listeners exhibit exactly the same perceptual pattern to such natural and excised English utterances. While this informal observation suggests that the classical account is not enough, it suggests that /k/ in *scan* does sound voiceless to Japanese listeners.

[2]That may or may not be correct; according to Vance (1987; 2008), /k, g/ in Japanese are more fronted than /k, g/ in English.

[3]Careful readers will notice that the role of **DI-sensitivity** (coarticulation-sensitive phonotactic repair) is already suggested by /u/ epenthesis after palatalized nasals in loanwords from French (e.g., /burugo:nj̱u/ 'Bourgogne'), as opposed to the moraic nasal rendition with no epenthesis after non-palatalized nasals (e.g., /sukjaN/ 'scan').

On the other hand, the crucial insight offered here is that seeming instances of vowel epenthesis are often the result of phonemic categorization, rather than of phonotactic repairs. The following examples more or less seem to be covered by that insight:

(4) a. Ich [ɪç] (German) > /iQhi̲/

(as found in Japanese orthography transcriptions in German textbooks)

b. Ludwig [ludvɪç] (German) > /ruuːdobiQhi̲/

(5) a. Bach [bax] (German) > /baQha̲/

b. loch [lox] (Scottish) > /roQho̲

(the Japanese orthography notation found in tourist guidebooks)

/hi/ is often realized as [ç] (with /i/-coarticulation traces being realized as palatalization) in Japanese; [x] is not employed in regular productions in Japanese, but the mimetic expressions for laughter, /aQhaQha/ and /oQhoQho/, are, at least sometimes, realized as [aʔx^{a}aʔx^{a}] and [oʔx^{o}oʔx^{o}o] respectively (Kamiyama, 2008). Thus the role of epenthesis (a repair of phonotactic violations) in loanword phonology could well be smaller than previously assumed; the need for a reexamination of loanword phonology in terms of phonemic categorization is also suggested.

8.2.3 Categorical and Non-Categorical Coarticulation Sensitivity

Yet another implication, or a remaining problem, is the relation between **DB-sensitivity** and **DI-sensitivity**. It was suggested that effects of **DI-sensitivity** are rather hard to observe because they could only be observed in phonotactic repair, which is phonological, but its successful exploitation requires that coarticulation traces in auditory memory should not have been completely erased by completed phonological processing (which would presumably suppress sub-phonemic distinctions); thus they would be subject to rather contradictory require-

ments, and the more difficulty in obtaining evidence for it, as compared to evidence for **DB-sensitivity**, would make sense given such contradictory requirements. However, recall the cross-experimental comparisons between Experiments 3–4. The observed better /ke/–/ku/ discriminations with a longer ISI (1,500 ms) than with a shorter ISI (250 ms) were interpreted in terms of larger effects of **DB-sensitivity** with a longer ISI.

In general, worse discriminations with a longer ISI are more usual observations in the literature, which could be interpreted in terms of auditory memory decay (Pisoni, 1973) or in terms of interference by phonological processing (Kawasaki, et al., 2012). However, worse discriminations with a longer ISI would make sense only when phonological processing suppresses auditory distinctions. In contrast, the above /ke/–/ku/ discrimination result is an observation of better discriminations with a longer ISI, which suggests that phonological processing magnified the perceptual effect of a rather weak auditory distinction between front (/e/) and back (/u/) coarticulation.

Why does phonological processing magnify a weak auditory distinction, at least sometimes? An obvious candidate answer, which is in fact already assumed more or less so far, is that phonological processing leads to categorical percepts. If so, both 'rich /e/-coarticulation traces' and 'poor /e/-coarticulation traces' are categorized as /i/, and both 'poor /u/-coarticlation traces' and 'rich /u/-coarticulation traces' as /u/, in phonological processing (**DB-sensitivity**); hence the discrimination performance for 'poor /e/-coarticulation traces' vs. 'poor /u/-coarticulation traces' tends to approach the discrimination performance for 'rich /e/-coarticulation traces' vs. 'rich /u/-coarticulation traces', as phonological processing proceeds with a longer ISI. In contrast, auditory processing is presumably non-categorical and does not have such a 'non-veridical' effect; hence the discrimination performance for 'poor /e/-coarticulation traces' vs. 'poor /u/-coarticulation traces' does not approach the discrimination performance for 'rich /e/-coarticulation traces' vs. 'rich /u/-coarticulation traces', no matter auditory processing is easier with a shorter ISI. Thus, in effect, the phonological **DB-sensitivity** should 'magnify' the per-

ceptual effect of auditory differences.

However, if **DB-sensitivity** leads to categorical percepts (although the percepts are, in effect, vowel percepts), the discrimination functions should be more discontinuous and categorical when they are based on **DB-sensitivity** than when they are based on **DI-sensitivity**. Empirical confirmations of this expectation with discrimination experiments employing synthetic stimulus continuum have to be left for future research.

8.2.4 Additional Tests to Distinguish Two Versions of One-step Models

Based on Matthews & Brown's (2004), this thesis has assumed the superiority of the **slot filling** version over the **suprasegmental matching** version of one-step models. Two lines of further research are suggested by this assumption.

First, the **slot filling** version as presented in this thesis is rather too simplistic, in the sense that the slots in the phonological frames are simply assumed to be either 'consonant slots' or 'vowel slots', with no finer specifications. Such an assumption would work for Matthews & Brown's (2004) results, but would contradict observations of perceptual conversions (Massaro & Cohen, 1983; Hallé et al., 1998). In order to incorporate Massaro & Cohen's or Hallé et al.'s observations of perceptual conversion, we will have to assume that the phonological frames in Thai provide consonant cluster slots with no further specifications, while the phonological frames in English or French provide an onset cluster slot with some specific specifications on which particular consonant clusters could or could not fill in the slots.

According to Matthews & Brown, the Thai phonotactics do allow /-CC-/ clusters in general, although /-bd-/ clusters are somehow unattested; if Thai listeners' perception is guided by structural phonotactic constraints (according to which /-CC-/ clusters should be fine, with /-bd-/ clusters merely constituting accidental gaps, rather than systematic phonotactic violations), Thai listeners should not epenthesize when presented with /-bd-/ clusters, whereas if their perception is guided by familiarity with attested sequences, they would epenthesize when

presented with /-bd-/ clusters. Their results supported the former prediction; Thai listeners were successful in discriminating /-bd-/ and /-bad-/ (where /a/ is the default epenthetic vowel in Thai). Such a finding receives a straightforward account in the **slot filling** version; consonant cluster slots were available and hence there was no need to epenthesize.

However, if phonological frames in general only consist of 'consonant slots' and 'vowel slots', with no finer specifications on the slots, the **slot filling** version would fail to cover observations of perceptual conversions from one consonant cluster type to another. For example, as reviewed in Chapter 3, Massaro & Cohen (1983) observed English listeners' perceptual conversion of /dl/ onsets to /dr/ onsets; Hallé et al. (1998) observed French listeners' perceptual conversion of /dl, tl/ onsets to /gl, kl/ onsets. If the absence of such onsets means only that such onsets are accidental gaps, then such conversions should not have been observed. Rather, the relevant phonotactic constraints in the English and French cases are specific bans on /dl, tl/ onsets. Thus, in order to make sense both of Matthews & Brown's (2004) results on the one hand, and of Massaro & Cohen's (1983) and Hallé et al.'s (1998) results on the other, we will have to assume that, in contrast to Thai, English and French phonological frames are equipped with more specific restrictions than 'consonant'.

This conclusion leads us to the question of which unattested clusters in which languages are mere accidental gaps or results of systematic bans, and why; the **slot filling** version has to be refined to accommodate various further, probably cross-linguistically different, constraints on the structural frames.

Second, while the **suprasegmental matching** version is an old idea, the **slot filling** version is rather novel.[4] However, Matthews & Brown's (2004) results are the only evidence I am aware of that supports the **slot filling** version over the **suprasegmental matching** version (while I am aware of no evidence that supports the **suprasegmental matching** version over the **slot filling**

[4] A similar idea was not new in production research (e.g., Shattuck-Hufnagel, 1979), but I know of no such proposal in perceptual research.

version). Thus it would be preferable if additional evidence could be obtained. The crucial difference between the two versions is whether 'content' (specific phonemic segments) and 'structure' (phonological frames with slots to be filled in by specific phonemic segments) are perceptually separable or not; the **slot filling** version claims that they are separable, while the **suprasegmental matching** version claims that they are inseparable. Thus an observation of an effect of 'structure' independent of 'content' would support the **slot filling** version.

For example, we could modify Pallier et al.'s (1993) experiments with French listeners (reviewed in Chapter 3) to distinguish the two versions. Recall that they divided the trials to 'inductors' and 'tests', and one or two inductor trials were presented before each test trial; the inductor trials' role was to attract listeners' attention to a specific position within the stimuli. For example, if the inductor trial is 'b' detection within *dou-blure* (with the syllable boundary indicated with a hyphen), listeners' attention would be drawn to the onset of the second syllable, and a subsequent test trial of 'p' detection within *ca-price* would count as phoneme detection in the attended position, whereas if the inductor trial is 'b' detection within *sub-merge*, the same test trial would count as phoneme detection within a non-attended position. Their results suggested that:

- the detection of /C_1/ (e.g., 'p') within /...C_1-C_2.../ (e.g., *cap-ture/*) is faster if the listeners attended to the coda of the first syllable within the stimuli than if they attended to the onset of the second syllable within the stimuli,

- the detection of /C_1/ (e.g., 'b') within /...C_1C_2-.../ (e.g., *ta-bleau*) is faster if they attended to the onset of the second syllable than if they attended to the coda of the first syllable.

However, in their experiments, attentions were allocated on (a) the coda position of the first syllable vs. (b) the onset position of the second syllable; thus the allocation could be interpreted either on the coda vs. the onset position, or on the first vs. the second syllable. Under the (a) interpretation, their results would support listeners' perceptual access to sub-syllabic positions

and hence support the **slot filling** version over the **suprasegmental matching** version, but under the (b) interpretation, their results do not contradict the **suprasegmental matching** version. Because both interpretations are possible, their results do not distinguish the two versions.

This in turn means that their experiments could be modified so that only the (a) interpretation would be possible. For example, we could compare two types of 'inductor trials':

onset inductor trials: those in which the onset phoneme is to be detected (e.g., /k/ in *cap-ture*).

coda inductor trials: those in which the coda phoneme in the same syllable is to be detected (e.g., /p/ in *cap-ture*).

followed by two kinds of 'test trials':

onset test trials: those in which the onset phoneme is to be detected (e.g., /s/ in *sub-merge*).

coda test trials: those in which the coda phoneme (in the same syllable) is to be detected (e.g., /b/ in *sub-merge*).

Phoneme detections in onset test trials preceded by onset inductor trials or in coda test trials preceded by coda inductor trials would count as 'attended position condition', while phoneme detection in onset test trials preceded by coda inductor trials or in coda test trials preceded by onset inductor trials would count as 'non-attended position condition'; if listeners' RT's differ across the two conditions, it could only be attributed to listeners' attentions on (a) sub-syllabic positions, rather than (b) their attention on the first vs. the second syllable. In other words, the **slot filling** version leads us to expect that listeners' RT's should differ across the 'attended position' and the 'non-attended position' condition, while the **suprasegmental matching** version leads us to expect that the two conditions should not differ. Such an experiment would be possible with French listeners (for example).

The above hypothetical experimental design presumably involves lexical access. Alternatively, a structural priming experiment with no lexical access (in the non-prime trials) might be possible. For example, consider a hypothetical language in which /b/ is realized as [b] in onsets and [p] in coda, whereas /p/ is realized as [p] in onsets and as a glottal stop in codas; thus [p] should be phonemically perceived as /b/ if it is assumed to be in the coda position, but as /p/ if it is assumed to be in the onset position.[5] Assume two kinds of prime stimuli:

- CVC
- CVCC

which are both phonotactically fine. Assume the **slot filling** version for the moment. After a series of CVC kinds of primes, then, the CVC frame is presumably activated, while the CVCC frame is not; if listeners are then presented with stimuli of the form:

[CVCpV]

the listeners should tend to syllabify the stimuli as:

CVC-pV (with the [p] portion perceived as the onset of the second syllable)

rather than

CVCp-V (with the [p] portion belonging to the coda of the first syllable)

so that listeners should tend to perceive /p/ rather than /b/. However, after a series of CVCC primes, which would activate the CVCC frame, listeners should tend to syllabify stimuli of the form:

[CVCpV]

as

[5] More generally, the hypothetical language is one in which the same physical signal counts as allophones of two different phonemes depending on its position within the syllable; the two phonemes do not necessarily have to be /b/ and /p/.

CVCp-V (with the [p] portion belonging to the coda of the first syllable)

rather than

CVC-pV (with the [p] portion constituting the onset of the second syllable)

and hence listeners should tend to perceive /b/ rather than /p/.

Now assume the **suprasegmental matching** version, according to which 'form' is inseparable from 'content' and hence structural priming should not be real. Thus the above effects of the primes on subsequent tendency with respect to phonemic identifications should not be observed.

Thus, if there is a language with such a property,[6] such a structural priming experiment would also distinguish the **slot filling** version and the **suprasegmental matching** version.[7]

Such lexical or non-lexical experiments have to be left for future research.

8.2.5 What Acoustic Properties Count as Exploitable Coarticulation Traces

Finally, coarticulation-based /i/ identification was less successful with shortened voiceless bursts in Experiment 2 than with full voiceless bursts in Experiment 1 on the one hand, and with voiced bursts produced by Japanese speakers in Experiment 5 than with voiced bursts produced by an American speaker in Experiment 8. Furthermore, /i/ identification from /ke/ bursts was less successful in Experiments 5 and 9 than in Experiment 2 (all with burst duration shortening), and was not successful at all in Experiments 8 and 10.

The spectral shapes (sharper first peaks followed by longer dips or valleys) were initially suggested as responsible for successful /i/ identification, but the results of Experiments 8 and 10 (with an American English speaker's productions as stimuli) betrayed that suggestion.

[6] That is, a language in which (i) the same physical signal counts as allophones of two different phonemes depending on its position within a syllable, (ii) syllabifications are not signaled by non-allophonic cues (such as stress, as in English), and (iii) phonotactic constraints on coda clusters are relatively liberal.

[7] The hypothetical structural priming experiment is partially inspired by Costa & Sebastian-Gallés(1998) priming studies in production. I thank Alice Turk for making me aware of their work.

Thus it is not clear what acoustic properties result in successful /i/ identifications from stop bursts.

Experiment 2 was simply meant to be a demonstration that more difficulty in /i/ perception from /i/-coarticulation than from voiced [i] comes and goes depending on the stimuli; Experiment 8 was simply meant to be a demonstration that Japanese listeners exhibit **DI-sensitivity** at least with some stimuli; Experiment 10 was simply meant to be a demonstration of the reality of **DI-sensitivity** effects with the sixth option of 'no vowel'. An examination of what acoustic properties count as exploitable coarticulation traces is (as noted at the end of Chapter 6) simply beyond the scope of this thesis and has to be left for future research.

Bibliography

Azevedo, M. M. 2005. *Portuguese: A Linguistic Introduction*. Cambridge, UK: Cambridge University Press.

Bailey, T. M., and Hahn, U. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language, 44*, 568–591.

Beckman, M. 1996. When is a syllable not a syllable? In T. Otake and A. Cutler, eds. *Phonological Structure and Language Processing: Cross-Linguistic Studies*, Berlin and New York: Mouton de Gruyter, pp. 95–123.

Beckman, M., and Shoji, A. 1984. Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese. *Phonetica, 41*, 61–71.

Benneau, A. 1997. Relevant spectral information for the identification of vowel features from bursts. *5th European Conference on Speech Communication and Technology*, Rodos Palace Hotel International Convention Centre, Greece.

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition, 104*, 591–630.

Best, C. T., McRoberts, G. W., and Sithole, N. M. 1988. Examination of Perceptual Reorganization for Nonnative Speech Contrasts: Zulu Click Discrimination by English-Speaking

Adults and Infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 345–360.

Blevins, J. 1995. The syllable in phonological theory. In J. A. Goldsmith, ed., *The Handbook of Phonological Theory*. Cambridge, MA: Blackwell, pp. 206–244.

Brown, R. W., & Hildum, D. C. 1956. Expectancy and the perception of syllables. *Language, 32*, 411–419.

Church, K. W. 1987. Phonological parsing and lexical retrieval. *Cognition, 25*, 53–70.

Coleman, J. S., & Pierrehumbert, J. 1997. Stochastic Phonological Grammars and Acceptability. In *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonology*. Somerset, NJ: Association for Computational Linguistics. pp. 49–56.

Costa, A., and Sebastian-Gallés, N. 1998. Abstract phonological structure in language production: Evidence from Spanish. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(4)*, 886–900.

Cutler, A., & Norris, D. 1979. Monitoring sentence comprehension. In W. E. Cooper & E. C. T. Walker (eds.) *Sentence processing: Psycholinguistic studies presented to Merril Garret*. Erlbaum.

Cutler, A., Otake, T., and McQueen, J. M. 2009. Vowel devoicing and the perception of spoken Japanese words. *Journal of the Acoustical Society of America, 125*, 1693–1703.

Dehaene-Lambertz, G., Dupoux, D., and Gout, A. 2000. Electrophysiological correlates of phonological processing: A cross-linguistic study. *Journal of Cognitive Neuroscience, 12*, 635–647.

Dupoux, E., Fushimi, T., Kakehi, K., and Mehler, J. 1999. Prelexical locus of an illusory vowel effects in Japanese. *Eurospeech '99 Proceedings: ESCA 7th European Conference on Speech Communication and Technology*.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. 1999. Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perpection and Performance, 25*, 1568–1578.

Dupoux, E., Pallier, C., Kakehi, K., and Mehler, J. 2001. New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes, 5*, 491–505.

Dupoux, E., Pallier, C., Sebastian, N., and Mehler, J. 1997. Destressing "Deafness" in Frensh? *Journal of Memory and Language, 36*, 406–421.

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., and Peperkamp, S. 2011. Where do illusory vowels come from? *Journal of Memory and Language, 64*, 199–210.

Faber, A., & Vance, T. J. 2000. More acoustic traces of "deleted" vowels in Japanese. In Nakayama, M., & Quinn Jr., C. J., eds.,*Japanese/Korean Linguistics, Vol. /*, Stanford: CSLI Publications, pp. 100–113.

Fais, L., Kajikawa, S., Werker, J., and Amano, S. 2005. Japanese listeners' perceptions of phonotactic violations. *Language and Speech, 48*, 185–201.

Fant, G. 1969. Stops in CV-syllables. Quarterly Progress and Status Report, Department for Speech, Music and Hearing, Kungliga Tekniska hʹogskolan.

Fowler, C. A. 2006. Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics, 68(2)*, 161–177.

Francis, A. L., and Ciocca, V. 2003. Stimulus presentation order and the perception of lexical tones in Cantonese. *Journal of the Acoustical Society of America, 114(3)*, 1611–1621.

Frisch, S. A., Large, N. R., & Pisoni, D. B. 2000. Perception of Wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*, 481–496.

Hallé, P., Segui, J., Frauenfelder, U., & Meunier, C. 1998. Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance, 24*, 592–608.

Haraguchi, S. 1999. Accent. In Tsujimura, N., ed. *The Handbook of Japanese Linguistics*, Malden and Oxford: Blackwell, pp. 1–30.

Hay, J., Pierrehumbert, J., & Beckman, M. 2004. Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple, eds., *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press. pp. 58–74.

Irwin, Mark. 2011. *Loanwords in Japanese*. Amsterdam and Philadelphia: John Benjamins.

Jusczyk, P. W., Friederici, A., Wessels, J., Svenkerud, V., and Jusczyk, A. 1993. Infants' sensitivity to the sound pattern of native language words. *Journal of Memory and Language, 32*, 402–420.

Jusczyk, P. W., and Luce, P. A. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630–645.

Kabak, B. 2003. *The Perceptual Processing of Second Language Consonant Clusters*. PhD diss., the University of Delaware.

Kabak, B., & Isdard, W. 2003. Syllabicity conditioned perceptual epenthesis. *Proceedings of the Berkeley Linguistics Society 29*, pp. 233–245.

Kabak, B., and Idsardi, W. J. 2007. Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? *Language and Speech, 50(1)*, 23–52.

Kakehi, K., Kato, K., & Kashino, M. 1996. Phoneme/syllable perception and the temporal structure of speech. In T. Otake & A. Cutler, eds., *Phonological Structure and Language Processing: Cross-Linguistic Studies*. Berlin: Mouton de Gruyter, pp. 125–143.

Kamiyama, T. 2008. *Datsu Nihongo Namari: Eigo (+α) Zissen Onseigaku (Get Rid of your Japanese Accent: An Introduction to Practical Phonetics of Japanese*. Suita: Osaka University Press.

Kaneko, E. 2006. Vowel selection in Japanese loanwords from English. *LSO Working Papers in Lingusitics (Proceedings of WiGL 2006)*. Madison: Linguistics Student Organization, University of Wisconsin-Madison, pp. 49–62.

Kawasaki, T., Matthews, J., Tanaka, K., & Odate, H. 2012. Persistent sensitivity to acoustic detail in non-native segments: The perception of English interdentals by Japanese listeners. ms., Hosei University and Chuo University.

Kingston, J., Kawahara, S., Mash, D., and Chambless, D. 2011. Auditory contrast versus compensation for coarticulation: Data from Japanese and English listeners. *Language and Speech, 54(4)*, 499–525.

Kondo, M. 1997. Mechanisms of vowel devoicing in Japanese. Unpublished Ph.D. dissertation, University of Edinburgh.

Kubozono, H. 1999. Mora and syllable. In Tsujimura, N., ed. *The Handbook of Japanese Linguistics*, Malden and Oxford: Blackwell, pp. 31–61.

Labrune, L. 2012. *The Phonology of Japanese*. New York: Oxford University Press.

Liberman, A. M. 1955. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27(4),* 769–773.

Lotto, A. J., and Holt, L. L. 2006. Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics, 68(2),* 178–183.

Lotto, A. J., and Kluender, K. R. 1998. Gestural contrast effects in speech peception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60,* 602–619.

Lovins, J. B. 1973. *Loanwords and the Phonological Structure of Japanese*. Unpublished Ph.D. dissertation, University of Chicago.

Luce, P., and Large, N. R. 2001. Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes, 16,* 565–581.

Maddox, W. T., and Estes, W. K. 1997. Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(3),* 539–559.

Maekawa, K., and Kikuchi, H. 2005. Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In van Weijer et al. eds. *Voicing in Japanese*. Berlin: Mouton de Gruyter, pp. 205–228.

Mann, V. A. 1980. Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28(5),* 407–412.

Mann, V. A. 1986. Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r." *Cognition, 24,* 169–196.

Mann, V. A., and Repp, B. H. 1981a. Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69*, 548–558.

Mann, V. A., and Repp., B. H. 1981b. Perceptual assessment of fricative–stop coarticulation. *Journal of the Acoustical Society of America, 69(4)*, 1154–1163.

Massaro, D. W., & Cohen, M. M. 1983. Phonological context in speech perception. *Perception & Psychophysics, 34*, 338–348.

Massaro, D. W., & Cohen, M. M. 191. Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology, 23*, 558–614.

Mateus, M. H., and d'Andrade, E. 2000. *The Phonology of Portuguese*. Oxford: Oxford University Press.

Matthews, J., and Brown, C. 2004. When intake exceeds input: Language specific perceptual illusions induced by L1 prosodic constraints. *International Journal of Bilingualism, 8*, 5–27.

Mazuka, R., Cao, Y. Dupoux, E., Christophe, A. 2011. The development of a phonological illusion: A cross-linguistic study with Japanese and French infants. *Developmental Science, 14*, 693-699.

McClelland, J. L., and Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

McMillan, N. A., and Creelman, C. D. 2005. *Detection Theory: A User's Guide*. (2nd edn.) Mahwah and London: Lawrence Erlbaum.

McNicol, D. 1972. *A Primer to Signal Detection Theory*. Sydney: Allen and Unwin.

McQueen, J.. M., Otake, T., and Cutler, A. 2001. Rhythmic cues and possible-word constraints in Japanese speech segmentation. *Journal of Memory and Language, 45*, 103–132.

Mehler, J., Dupoux, E., and Segui, J. 1990. Constraining models of lexical access: The onset of word recognition. In G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, MA: MIT Press, pp. 236–262.

Monahan, P. L., Takahashi, E., Nakao, C., and Idsardi, W. 2009. Not all epenthetic contexts are equal: Differential effects in Japanese illusory vowel perception. In S. Iwasaki, H. Hoji, P. M. Clancy, S. D. Sohn, eds., *Japanese/Korean Linguistics, 17*. Stanford: CSLI Publications.

Morelli, F. 2003. The relative harmony of /s+Stop/ onsets: Obstruent clusters and the sonority sequencing principle. In C. Féry & R. van de Vijer, (ed.), *The Syllable in Optimality Theory*. Cambridge, UK: Cambridge University Press, pp. 356–371.

Moreton, E. 2002. Structural constraints in the perception of English stop-sonorant clusters. *Cognition, 84*, 55–71.

Nakamura, M. 2003. The articulation of vowel devoicing: A preliminary analysis. *On-in Kenkyuu, 6*, 49–58.

Nihon Hoosoo Kyooka. 1998. *Nihongo Hatuon Akusento Jiten [Dictionary of Japanese Pronunciation and Accent Patterns]*. rev. ed. Tokyo: Nihon Hoosoo Shuppan Kyookai.

Norris, D., and McQueen, J. M. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115(2)*, 357–395.

Norris, D., McQueen, J. M., and Cutler, A. 2000. Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Seicnes, 23(3)*, 299–370.

Ogasawara, N. 2013. Lexical representation of Japanese vowel devoicing. *Language and Speech, 56(1)*, 5–22.

Ogasawara, N., and Warner, N. 2009. Processing missing vowels: Allophonic processing in Japanese. *Language and Cognitive Processes 24*, 376–411.

Otake, T., Yoneyama, K., Cutler, A., and van der Lugt, A. 1996. The representation of Japanese moraic nasals. *Journal of the Acoustical. Society of AMerica, 100(6)*, 3831–3842.

Pallier, C., Sebastian-Gallés, N., Felguera, T., Christophe, A., and Mehler, J. 1993. Attentional allocation within the syllabic structure of spoken words. *Journal of Memory and Language, 32*, 373–389.

Pierrehumbert, J. 1994. Syllable structure and word structure. In P. Keating, ed., *Papers in Laboratory Phonology III*, Cambridge, UK: Cambridge University Press, pp. 168–188.

Pierrehumbert, J. 2001. Stochastic phonology. *GLOT, 5*, 1–13.

Pierrehumbert, J. 2003. Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy, eds., *Probabilistic Linguistics*, Cambridge, MA: MIT Press, pp. 177–228.

Pisoni, D. B. 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13(2)*, 253–260.

Pitt, M. A. 1998. Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & Psychophysics, 60*, 941–951.

Pitt, M. A., and McQueen, J. M. 1998. Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39*, 347–370.

Pollard, C., and Sag, I. A. 1987. *Information-based Syntax and Semantics, Vol. 1*. Stanford: CSLI Publications.

Pollard, C., and Sag, I. A. 1994. *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Raphael, L. J. 2005. Acoustic cues to the perception of segmental phonemes. In D. B. Pisoni & R. E. Remez. eds. *The Handbook of Speech Perception*. Oxford: Blackwell, pp. 182–206.

Sato, Y. 1993. The durations of syllable-final nasals and the mora hypothesis in Japanese. *Phonetica, 50,* 44–67.

Segui, J., Frauenfelder, U., & Hallé, P. 2001. Phonotactic constraints shape speech perception: Implications for sublexical and lexical processing. In E. Dupoux, ed., *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*, Cambridge, MA: MIT Press, pp. 195–208.

Shibatani, M. 1990. *The Languages of Japan*. Cambridge, UK: Cambridge University Press.

Shieber, S. M. 1986. *An Introduction to Unification-Based Approaches to grammar*. Stanford: CSLI Publications.

Shattuck-Hufnagel, S. 1979. Speech errors as evidence for a serial ordering mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (eds.) *Sentence processing: Psycholinguistic studies presented to Merril Garret*. Hillsdale, J. J.: Lawrence Erlbaum Associates, pp. 295–342.

Smith, J. L. 2006. Loan phonology is not all perception: Evidence from Japanese loan doublets. In T. J. Vance and K. Jones. eds. *Japanese/Korean Linguistics, Vol. 14*, pp. 63–74. Stanford: CSLI Publications.

Treiman, R., Gross, J., & Cwikiet-Glavin, A. 1992. The syllabification of /s/ clusters in English. *Journal of Phonetics, 20,* 383–402.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. 2000. English speakers' sensitivity to phonotacitc patterns. In M. B. Broe & J. B. Pierrehumbert, eds., *Papers*

in Laboratory Phonology V: Acquisition and the Lexicon, Cambridge, UK: Cambridge University Press, pp. 269–282.

Tsujimura, N. 1996. *An Introduction to Japanese Linguistics*. Cambridge, MA, and Oxford: Blackwell.

Tsuchida, A. 1994. Fricative-vowel coarticulation in Japanese devoiced syllables: Acoustic and perceptual evidence. *Working Papers of the Cornell Phonetics Laboratory, 9*, 183–222.

Vance, T. 1987. *Introduction to Japanese phonology*. Albany, N. Y.: State University of New York Press.

Vance, T. 2008. *The Sounds of Japanese*. Cambridge: Cambridge University Press.

Varden, J. K. 1998. *On high vowel devoicing in standard modern Japanese: implications for current phonological theory*. Unpublished Ph.D. dissertation, University of Washington.

Vitevitch, M. S., & Luce, P. A. 1998. When words compete: Levels of Processing in perception of spoken words. *Psychological Science, 9*, 325–329.

Vitevitch, M. S., & Luce, P. A. 2005. Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language, 52*, 193–204.

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*, 306–311.

Warren, R. M., & Warren, R. P. 1970. Auditory illusions and confusions. *Scientific American, 223*, 30–36.

Werker, J. F. 1991. The ontogeny of speech perception. In I. G. Mattingly & M. Studdert-Kennedy, eds., *Modularity and the Motor Theory of Speech Perception: Proceedings of a Conference to Honor Alvin M. Liberman*, Hillsdale: Lawrence Erlbaum, pp. 91–109.

Werker, J. F. 1995. Exploring developmental changes in cross-language speech perception. In Gleitman, L. R., and Liberman, M. eds. *An Invitation to Cognitive Science, Vol. 1: Language. (2nd edn.)*. Cambridge, MA: MIT Press, pp. 87–106.

Werker, J. F., and Logan, J. S. 1985. Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37*, 35–44.

Winitz, H., Scheib, M. E., and Reeds, J. A. 1972. Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *The Journal of the Acoustical Society of America, 51*, 1309–1317.

Yuen, C. L. 2000. The perception of Japanese devoiced vowels. *Proceedings of the Chicago Linguistic Society, 36*, 531–547.