UNIVERSITY OF CALIFORNIA

Los Angeles

Word-level and Feature-level Effects in Phonetic Imitation

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Linguistics

by

Kuniko Yasu Nielsen

2008

The dissertation of Kuniko Yasu Nielsen is approved.

Jody Kreiman

Bruce Hayes

Colin Wilson, Committee Co-chair

Patricia Keating, Committee Co-chair

University of California, Los Angeles

2008

To my parents:

Michiko and Yoshitaka Yasu

with love and gratitude

TABLE OF CONTENTS

Chapter 1: Introduction 1
1.1 Abstractionist versus Episodic Views 2
1.1.1 The Abstract View of Speech Perception 2
1.1.1 The Episodic View of Speech Perception 3
1.1.1.1 Exemplar-based Models 3
1.1.1.2 Strengths of Exemplar-based Models 5
1.1.1.3 Challenges for Exemplar-based Models 6
1.2 Specificity 6
1.2.1 Repetition Priming 7
1.2.2 Perceptual Learning 8
1.2.3 Spontaneous Imitation 8
1.3 Two Types of Variability: Indexical and Allophonic 10
1.3.1 Differences between Indexical and Allophonic Variability 11
1.3.2 Specificity of Indexical Variability 12
1.3.3 Specificity of Allophonic Variability 12
1.4 Units of Phonological Representations 14
1.5 Current Study 16
1.5.1 Research Questions 16
1.5.2 Experiments 18
1.6 Organization of the Dissertation 20

2 Chapter 2: Experiment 1 - Imitation of Extended VOT 22
2.1 Method 24
2.1.1 Participants 24
2.1.2 Stimuli Selection 24
2.1.2.1 Production List 25
2.1.2.2 Listening List 26
2.1.3 Stimulus Construction 26
2.1.4 Procedure 29
2.1.5 Statistical Analysis 30
2.2 Results and Discussion 31
2.2.1 Phonetic Imitation 31
2.2.2 Gender Difference in VOT 34
2.2.3 Generalizability 35
2.2.4 Word Specificity 37
2.2.5 Post-hoc Analysis of Word Specificity 38
2.3 Discussion 42
2.3.1 Generalization of VOT Imitation 42
2.3.2 Word Specificity and Episodic Traces 43
2.3.3 Gender Difference in VOT and Imitation 45
2.4 Interim Summary 46

3 Chapter 3: Experiment 2 - Imitation of Reduced VOT 48
3.1 Method 50
3.1.1 Participants 50
3.1.2 Stimuli Selection 50
3.1.3 Stimulus Construction 50
3.1.4 Procedure 53
3.1.5 Statistical Analysis 53
3.2 Results 53
3.2.1 Phonetic Imitation 54
3.2.2 Word Specificity 56
3.2.3 Generalizability 58
3.2.4 Long vs. Short VOT (Experiment 1 & 2 Comparison) 60
3.3 Discussion 61
3.3.1 Asymmetry of Imitation 61
3.3.2 Gesture Theories of Speech Perception 66
3.4 Interim Summary 67

4 Chapter 4: Experiment 3 - Imitation of High Vowel Devoicing in Japanese 70
4.1 Background: Japanese Vowel Devoicing 71
4.2 Method 74
4.2.1 Participants 74
4.2.2 Stimuli Selection 74
4.2.3 Stimulus Construction 75
4.2.4 Procedure 79
4.2.5 Acoustic Analysis 80
4.2.6 Statistical Analysis 82
4.3 Results 83
4.3.1 Phonetic Imitation 83
4.3.1.1 Gradient Analysis 83
4.3.1.2 Categorical Analysis 85
4.3.2 Generalizability 89
4.3.2.1 Phoneme-level Generalization 89
4.3.2.2 Feature-level Generalization 91
4.3.3 Word Specificity 92
4.3.3.1 Lexical Frequency 93
4.3.3.2 Presence of Exposure 94
4.3.4 Marked Environments 96
4.3.4.1 Consecutive Devoicing 97
4.3.4.2 Mid-vowel Devoicing 99
4.3.4.3 Pre /h/ Devoicing 99
4.3.4.4 Word-final Devoicing 100
4.4 Discussion 102
4.4.1 Phonetic Imitation 102
4.4.2 Lexical Frequency & Japanese Devoicing 104
4.4.3 On the Mechanism of Japanese Devoicing 107

4.4.3.1 Effect of Lexical Frequency 108
4.4.3.2 Segmental Duration 108
4.5 Interim Summary 111

Chapter 5: Discussion and Conclusion 112
5.1 Summary of Results 112
5.2 General Discussion 114
5.2.1 Phonological Representations: How episodic and abstract? 114
5.2.2 Size/Unit of Representation 116
5.2.3 Implications for Models of Speech Perception and Production 119
5.3 Directions for Future Work 123
5.4 Conclusion 125

Appendix A: English Test and Filler Words with Lexical Frequency 126
Appendix B: Japanese Devoicing Test and Filler Words with Lexical Frequency 131

References 134

LIST OF FIGURES

Figure 2-1: Example of measurement of Voice Onset Time (VOT) from a waveform .. 23
Figure 2-2: Example waveform and spectrogram of word "pass" before the VOT manipulation 28
Figure 2-3: Example waveform and spectrogram of word "pass" after the VOT manipulation, used as one of the listening stimuli in Experiment 1 29
Figure 2-4: Phonetic imitation and Gender difference on VOT 32
Figure 2-5: Individual variability of VOT and phonetic imitation 33
Figure 2-6: Generalizability of spontaneous imitation 35
Figure 2-7: Effect of lexical frequency on degree of imitation 40
Figure 2-8: Effect of presence of exposure on degree of imitation 41
Figure 3-1: [Reprinted from Keating (1984), part of Figure 3] VOT values for English utterance-initial bilabial stops before twelve vowels..... 49
Figure 3-2: Example waveform and spectrogram of word "pass" before VOT manipulation 52
Figure 3-3: Example waveform and spectrogram of word "pass" after shortening VOT by 40 ms., which was used as one of the listening stimuli in Experiment 2..... 52
Figure 3-4: Absence of Short VOT imitation 55
Figure 3-5: Individual variability of VOT and phonetic imitation 55
Figure 3-6: Gender difference of lexical frequency effect on VOT 57
Figure 3-7: Generalization of Short VOT imitation 59
Figure 3-8: Comparison of Experiments 1 (Long VOT) and 2 (Short VOT)..... 61
Figure 3-9: Histogram of VOT distribution of word-initial /p/ in the baseline recordings (two experiments combined)..... 63
Figure 4-1: A spectrogram showing the word "taishokuteate" (retirement bonus)..... 77
Figure 4-2: Spectrograms of the word "tekagami" (hand mirror) (a) before and (b) after the acoustic manipulation 78
Figure 4-3: Example of acoustic measurement..... 81
Figure 4-4: Example of acoustic measurement using intensity reading 82
Figure 4-5: Phonetic imitation of Japanese vowel devoicing in vowel duration..... 84
Figure 4-6: Phonetic imitation of Japanese vowel devoicing measured in rate of complete devoicing..... 86
Figure 4-7: Phonetic imitation of Japanese vowel devoicing in summed vowel duration 86
Figure 4-8: Individual variability of vowel duration and phonetic imitation 88
Figure 4-9: Generalization of phonetic imitation at phoneme level 90
Figure 4-10: Generalization of phonetic imitation at a feature-level..... 92
Figure 4-11: Effect of lexical frequency..... 94
Figure 4-12 [repeated from 4-9]: Effect of target speech exposure..... 95
Figure 4-13: Phonetic imitation in phonologically marked environments 98
Figure 4-14: Gender difference of word-final devoicing in Japanese 101
Figure 4-15: Hyperarticulation with reduced or increased coarticulation (From Figure 2 in Scarborough, 2004)..... 107

Figure 4-16: Schematized Japanese vowel devoicing as a phonetic process..... 109
Figure 4-17: Example of extreme partial devoicing 110

LIST OF TABLES

Table 1: Examples of test stimuli used in Experiment 1 26
Table 2: Summary of Experiment 1 Results 31
Table 3: Summary of Experiment 2 Results. The mean, median, standard error and standard deviation of VOT (ms) are shown by stimulus type 54
Table 4: Type and distribution of Experiment 3 stimuli 75
Table 5: The effects of frequency and neighborhood density on lexical difficulty (from Table 12 in Scarborough, 2004)..... 105

ACKNOWLEDGMENTS

First of all, my deepest gratitude goes to my advisor, Pat Keating. She has always been extremely supportive and generous, yet critical at the right times. She has gently guided me through this long journey with her high standard as a phonetician and her wisdom as a mother. I came to UCLA to work with Pat, and I couldn't be happier with this decision. My heartfelt thanks also go to Colin Wilson, who has been an amazing source of inspiration since my first year at UCLA. His strong confidence in my development as a researcher helped me to keep my chin up, and the quality of my work constantly benefited from his keen linguistic insights. Our weekly meetings have provided me the best brainstorming opportunities, from which this project developed.

I am also grateful to my wonderful committee members: Bruce Hayes for his insightful and sharp-witted comments, Sun-Ah Jun for our long talks on phonetics and motherhood, and Jody Kreiman for her fresh perspective on data analysis. I also owe a great debt to my undergraduate advisor, Mike Broe, who first introduced me to the wonders of Laboratory Phonology at OSU. Without his inspiration and guidance, I probably would not have chosen this path. I am also grateful for my teachers, who all inspired me in different ways: Mary Beckman, Keith Johnson, Peter Ladefoged, Kie Zuraw, Megha Sundara, Susie Curtiss, and Donca Steriade.

Many thanks are also due to my dearest friends in the p-lab and in the department: Shabnam Shademan for being the sister I never had and supporting me during the hardest times, her partner David Ross for his big heart and friendship, Rebecca Scarborough for many inspiring conversations on phonetics and life, Stephen Wilson for his brilliant advice on experiments, Tim and Heidi Arbisi-Kelm for the wonderful times and food we shared,

Jason Kandybowicz for his ever-lasting smile and big hugs, Brook Lillehaugen and Roger Standridge for being the best housemates one could possibly imagine, and Ying Lin for his harmonious smile and presence. Adam Albright, Julia Berger, Jason Bishop, Taehong Cho, Melissa Epstein, Christina Esposito, Jeffrey Heinz, Tomoko Ishizuka, Amanda Jones, Sameer Khan, Christina Kim, Sahyang Kim, Ingvar Lofstedt, Reiko Okabe, Shigeko Sekine, Molly Silman, Marcus Smith, Kristine Yu, and Chad Vicenik: you all helped me in one way or another. Thank you.

My sincere thanks also go to Cheryl Hipp and Eric Williams, who helped me with thousands of phonetic measurements. I know I couldn't have asked for more capable research assistants. Many thanks are also due to Yuko Abe and Professor Peri Bhaskararao at Tokyo University of Foreign Studies for their generous help with the data collection in Japan.

I also wish to thank my wonderful in-laws in Norway and family in Japan. To my parents, for their boundless love and support over the years, and for the numbers of cross-Pacific visits. To my brothers and their wives, for their encouragement, and for making sure that my parents are well and happy.

Most of all, I owe my deepest gratitude to my dear husband, Jon-Fredrik. Were it not for his love and support, I would not have made it even half way through. He read every single version of this dissertation, made sure our children were happy, and reminded me how much we both love research. Lastly, words cannot express how grateful I am for Mina and Kai, for filling my life with smile and joy.

Finally, I thank the UCLA Department of Linguistics for financial support for seven years. This dissertation has been supported by a UCLA dissertation year fellowship and an NSF doctoral improvement grant (BCS-0547578).

VITA

October 12, 1968	Born, Mito, Japan
1999	B.A., Linguistics (with Distinction) and Chinese Ohio State University Columbus, Ohio
2002-2007	Graduate Research Assistant Department of Linguistics University of California, Los Angeles
2001-2008	Graduate Teaching Assistant Department of Linguistics University of California, Los Angeles
2005	Dissertation Year Fellowship University of California, Los Angeles
2006-2008	Doctoral Improvement Grant (BCS-0547578) National Science Foundation

PUBLICATIONS AND PRESENTATIONS

Keating, P., Nielsen, K., Manis, F., and Bruno, J. (December 2006). Acoustic phonetic variability and auditory word recognition by dyslexic and non-dyslexic children. Poster presented at the 152nd Meeting of the Acoustical Society of America. Honolulu, HI.

Nielsen, K. Y. (2004). Segmental Differences in the Visual Contribution to Speech Intelligibility. *UCLA Working Paper in Phonetics 103*. pp. 106-147.

___. (May 2004). Segmental differences in the visual contribution to speech intelligibility. Poster presented at the 148th Meeting of the Acoustical Society of America. New York, NY.

___. (2004). Segmental Differences in the Visual Contribution to Speech Intelligibility. *Proceedings of the 8th International Conference on Spoken Language Processing,* Jeju, Korea.

___. (May 2005). Generalization of the imitation effect within a natural class. Poster presented at the 149th Meeting of the Acoustical Society of America. Vancouver, Canada.

___. (2005). Generalization of phonetic imitation across place of articulation. *Proceedings of the ISCA Workshop on Plasticity in Speech Perception*, University College London, London, UK.

___. (2005). Investigation of SLI Children's Semantic Knowledge through Intonation Patterns. *UCLA Working Papers in Linguistics no.13, Papers in Psycholinguistics 2.* pp. 76-86.

___. (2005). Kiche Intonation. *UCLA Working Papers in Phonetics 104*. pp. 45-60.

___. (October 2005). The interaction between specificity and linguistic contrast. Poster presented at the 150th meeting of the Acoustical Society of America, Minneapolis, MN.

___. (November 2005). The specificity and abstractness of phonetic imitation. Talk presented at RIKEN BSI Forum, Wako, Japan.

___. (January 2006). VOT imitation generalizes across phonemes and natural classes. Talk presented at the 2006 LSA meeting in Albuquerque, NM.

___. (June 2006). The specificity and abstractness of phonetic imitation. Poster presented at the 10th Conference on Laboratory Phonology, Paris, France.

___. (2006). Specificity and Generalizability of Spontaneous Phonetic Imitation. *Proceedings of the 9th International Conference on Spoken Language Processing,* Pittsburgh, PA.

___. (2007). The Interaction between Spontaneous Imitation and Linguistic Knowledge. *UCLA Working Papers in Phonetics 105*. pp. 125-137.

___. (2007). Implicit Phonetic Imitation is constrained by Phonemic Contrast. *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, Germany.

___. (January 2008). Phonetic imitation of Japanese vowel devoicing. Talk presented at the 2008 LSA meeting in Chicago, IL.

Wilson, C., and ___. (May 2008). A Bayesian Network Model of Multi-level Phonetic Imitation. Talk presented at the 27th West Coast Conference on Formal Linguistics. Los Angeles, CA.

ABSTRACT OF THE DISSERTATION

Word-level and Feature-level Effects in Phonetic Imitation

By

Kuniko Yasu Nielsen

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2008

Professor Patricia Keating, Co-chair

Professor Colin Wilson, Co-chair

The imitation paradigm (Goldinger, 1998) has been used to show that speakers implicitly imitate fine phonetic details of recently heard speech. Although word-level specificity of imitation has been shown, it is unknown whether imitation also involves sub-lexical units. By using a modified imitation paradigm, this dissertation examined the psychological reality of three levels of linguistic units (i.e., word, phoneme, and feature) in imitation, as well as the interaction between phonetic imitation and phonetic/phonological factors.

The results from Experiment 1 showed that artificially lengthened English VOT was imitated by participants. Furthermore, the shift in their speech was generalized to new instances of the target phoneme /p/ and the new phoneme /k/, showing that the imitation was generalized at a feature level. At the same time, a word-specific imitation pattern was

observed in two post-hoc analyses. Experiment 2 showed that artificially shortened VOT in English, which can endanger phonemic contrast with the phoneme /b/, was *not* imitated, suggesting that the need to preserve phonemic contrast suppresses phonetic imitation. Among the participants who did demonstrate imitation in this experiment, the change was generalized to the new phoneme /k/. Experiment 3 showed that Japanese vowel deletion was imitated, despite the fact that this violated certain phonetic and phonological constraints. Further, the change was generalized to novel items containing /i/. Consistent with the finding in Experiment 1, the degree of imitation was shown to be partially word-specific.

Taken together, these results provide the following observations: 1) perceived fine phonetic details are retained in listeners' memory and subsequently affect their speech production, 2) multiple levels of phonological representations (i.e., word- and feature-level unit) simultaneously contribute to the pattern of phonetic imitation, 3) phonetic imitation is not an automatic process, but rather a process which is modulated by other factors, such as the presence of linguistic contrast. In addition, large speaker variability in the degree of imitation was observed in all three experiments.

Chapter 1: Introduction

A fundamental task in speech science is to understand how we perceive speech. Although speech signals vary greatly from token to token, the human mind perceives speech effortlessly. Traditional accounts of speech perception assume abstract and invariant phonological representations. According to this view, certain variability in the speech signal represents "noise", which has to be filtered out in the process of speech perception. Recently, this traditional view has been challenged by the episodic view of speech perception, which assumes specific and episodic representations (see Tenpenny, 1995 for a review). In this view, all the information in a speech signal is preserved in memory and utilized in speech perception. A growing body of research now suggests that fine phonetic details are retained and used in speech perception, and that both speech perception and production are more plastic than once considered, supporting the episodic view. However, the existing literature does not necessarily rule out the presence of abstract representations, because previous studies mostly concern phonetic details that are linguistically irrelevant.

This dissertation aims to investigate the nature of phonological representations by examining how perceived allophonic variations are imitated. Three experiments employing the spontaneous imitation paradigm (Goldinger, 1998) are conducted, which compares subjects' speech production before and after they are exposed to other speakers' speech. The central questions addressed in this dissertation concern 1) the size of the linguistic representations involved in speech perception and production, and 2) the abstractness and

episodicity of these representations.

1.1 Abstractionist versus Episodic Views

1.1.1 The Abstract View of Speech Perception

Traditional accounts of speech perception assume that the representations of linguistic sound structures are *abstract*, and that these representations need to be extracted from variant speech signals in the process of speech perception (e.g., Halle, 1985). That is, perceived speech signals are reduced to sequences of discrete, abstract representations, such as phonemes and features. According to this view, some types of variability in the speech signal (such as speech rate, voice quality or F0) represent 'noise' which listeners have to filter out by some sort of normalization process in order to achieve prototypical, abstract representations (such as phonemes) prior to lexical access.

In fact, most phonological theories, whether rule-based or constraint-based, assume abstract underlying representations. For example, in generative phonology (e.g., Chomsky and Halle, 1968), an utterance is a string of distinctive feature matrices. No phonological representations smaller than a distinctive feature are assumed, and patterns of fine phonetic details which cannot be captured by distinctive features are not considered to be a part of phonological grammar, but to be attributable to phonetic implementation rules.

Support for the notion of abstract representations has traditionally been provided by phonological alternations and phonotactic constraints, where categorically expressed (by distinctive features or gestures) phonological environments solely account for the systematic application of the changes or rules. However, the abstract representations of phonology have yet to be reconciled with the variability of the acoustic signal (e.g., lack of acoustic

invariance problem: Liberman and Mattingly, 1985), and many challenges remain on the way to mapping variant acoustic-phonetic information onto abstract representations, whether the representation is acoustic, gestural, or amodal. Variability in speech *production* presents another challenge for abstract representations. If representations are abstract, they require a phonetic implementation component to generate the full range of phonetic detail seen in fluent connected speech. However, some variations in speech are already shown to be related to paralinguistic factors (e.g., Labov, 1972), and thus the phonetic implementations cannot be accounted for solely by their phonological environments.

1.1.1 The Episodic View of Speech Perception

In the last two decades, this traditional abstractionist view has been challenged by the episodic view of memory, which assumes representations of any category to be specific and episodic (Semon, 1909, 1923; Tulving, 1972, 1983, 2002; Schacter et al. 1978). According to this view, every stimulus leaves a unique memory trace, and a category is represented in memory as a cloud of traces. Each time a new stimulus is presented, all traces are activated according to their similarity to this stimulus, and recognition of the stimulus takes place when the most activated traces come to consciousness.

1.1.1.1 Exemplar-based Models

The episodic view has been implemented by exemplar-based models (e.g., Hintzman, 1986; Nosofsky, 1986; Johnson, 1997; Pierrehumbert, 2001, 2002) in which detailed information in the signal is preserved as a trace or exemplar. A perceptual category is defined as the set of all the exemplars of the category, and categorization is based on sums of similarity over categories. For example, word recognition in Hintzman's MINERVA 2 model (1986, 1988)

occurs as follows: For every known word, a large number of traces exist in memory. Each time a stimulus word is presented, an analog probe communicates with all the traces in parallel, activating traces in proportion to their mutual similarity. These activated traces constitute an *echo*, whose intensity reflects the total activation in memory created by the probe. Thus greater similarity of traces to the probe increases echo intensity. The stimulus is recognized when the most intense echo comes to consciousness. In this model, any type of variability in the speech signal is a source of information for processing. Other less-extreme, and linguistically more informed, exemplar-based models include Johnson (1997), Pierrehumbert (2001, 2002), and Coleman (2002). In Johnson (1997), an exemplar is an association between a set of auditory properties (e.g., formant values) and categorical labels (e.g., the gender of the speaker and the linguistic value of the exemplar). By retaining speaker-specific information, the model makes it possible to account for the process of talker normalization in speech perception. Pierrehumbert's modular feed-forward model of speech production (2002) is a hybrid of a traditional modular view and an exemplar view: In this model, details of allophonic variability as well as speaker-information are preserved in memory and systematically associated with words, capturing attested word-specific phonetic changes. In Coleman (2002), word forms are stored as memories of phonetic (psychophysical) experiences, and statistical regularities over the phonetic space are expressed as phonological constraints. More recently, Lin (2005) showed that an unsupervised statistical learning model, which utilizes lexical exemplars, is able to group acoustically similar sounds into natural classes from the acoustic traces of isolated words.

1.1.1.2 Strengths of Exemplar-based Models

Johnson (1997) discusses that one of the strengths of exemplar models is that they provide a solution to the lack of invariance problem in on-line perception. According to exemplar theories, there is no need for mapping acoustic-phonetic information onto abstract phonological representations, because the two are more or less the same – the representational medium (or currency) of speech consists of the variant speech signals listeners experienced in the past.

Another appeal of this view is its ability to account for the acquisition of categories. Previous studies have shown that young infants are born with an ability to discriminate many phonetic (including non-native) distinctions without relevant experience (e.g., Kuhl et al. 1992), indicating that there are no innate phonological categories. Infants' phonological environment shapes a perceptual reorganization, which mostly takes place during the first year of life (Werker & Tees, 1984). If there are no innate phonological categories or prototypes, then infants need to discover/create them through exposure to their ambient language. As mentioned earlier, exemplar-based models assume that a perceptual category is defined as the set of experienced exemplars, providing a rather realistic account of how early experience in a given language forms its categories (e.g., Lin, 2005).

In addition, exemplar-based models predict plasticity of the categories.[1] A number of studies have shown that acquisition is never really over: For example, Harrington, Palethorpe, and Watson (2000) showed that Queen Elizabeth II has gradually shifted her pronunciation in the direction of the Southern British English which has become fashionable with younger

[1] This is not a unique property of exemplar models, however, and any theory of learning that allows a change in representation would predict plasticity.

speakers. In Sancier and Fowler (1997), a fluent Portuguese-English bilingual speaker shortened VOTs in both unaspirated (Portuguese) and aspirated (English) stops while she was in Brazil. Similarly, exemplar models account for the role of frequency of occurrence in phonetic reduction and phonological change, as discussed by Pierrehumbert (2002). These findings show either that the phonetic implementation and mapping rules that are part of abstract theories must be plastic, or that phonological representations are variant and based on cumulative experience.

1.1.1.3 Challenges for Exemplar-based Models

There are some theoretical challenges for the exemplar theories, however. The most obvious challenge concerns the capacity of memory. Even taking memory decay into consideration (e.g., Nosofsky 1986; Pierrehumbert, 2001, Wedel, 2006), preserving all stimuli does not seem to be very economical. Also, as Goldinger (1997) discusses, any exemplar model (e.g., MINERVA 2: Hintzman, 1986) has to discover "dimensions" along which to determine the similarity between a new stimulus and exemplars. In the phonological literature, "similarity" is often defined in terms of features (e.g., Frisch, 1996) and natural classes, which are both sub-phonemic. Given a common assumption that exemplar-based theories are tied to lexical words (as in MINERVA 2, for example), in order to compute fine-grained phonetic similarity, exemplar models potentially need to include sub-lexical representations.

1.2 Specificity

As mentioned above, exemplar-based theories state that every experienced stimulus leaves a unique trace in memory with detailed information. A body of research now supports this claim by showing *specificity* of episodic traces in various forms.

1.2.1 Repetition Priming

The repetition priming paradigm tests the effects of prior perceptual experience, by comparing the processing speed and accuracy of already-experienced tokens and novel counterparts with a given kind of variability (e.g., voice, speech style, lexical words, syntactic structure, etc.). When already-experienced tokens are processed faster than new ones, the result is interpreted to mean that information in the episodic trace is utilized to facilitate processing. This is referred to as a *specificity effect*.

During the past two decades, a number of repetition priming studies have shown a robust effect of specificity. Listeners perceive speech faster and more accurately when it is repeated in the same voice (Mullennix et al., 1989, among others), intonation, F0 (Church and Schacter: 1994), speech rate (McLennan and Luce, 2005) or speech style (McLennan et al., 2003). Goldinger (1996) showed that even *similar* voices may cause a priming effect: In his study, participants' performance on trials with different voices was affected by the distance between study (which participants were familiarized to) and test voices as calculated by multidimensional scaling (MDS; Kruskal and Wish, 1978; Shepard, 1980), suggesting that the episodic traces retain voice details with great precision. Goldinger's results also revealed the time course of the voice specificity effect: In an explicit recognition memory experiment (where participants were asked to classify words as either "old" or "new"), old voices were recognized more accurately than new at 5-min and 1-day delays (across 2, 6, and 10 voices conditions), but not at a 1-week delay. In a perceptual identification experiment (where participants auditorily perceived words and identified them by typing), the same-voice repetitions were recognized more accurately than different-voice repetitions at all delays. These results from repetition priming suggest that episodic traces of spoken words are

available to explicit recognition memory for at least a day, and to implicit perceptual memory for up to a week.

1.2.2 Perceptual Learning

Recent studies in *perceptual learning* have shown that perceived fine phonetic details in speech signals are utilized to modulate category assignment. Allen and Miller (2004) examined listeners' sensitivity to talker differences (indexical variability) in voice-onset-time (VOT) associated with a word initial voiceless consonant. When presented with a short vs. long VOT variant of a given talker's speech, listeners selected the variant consistent with their experience of that talker during a previous training phase. Listeners also generalized talker-specific VOT information to novel words which share the same initial consonant (e.g., *time* > *town*: phoneme-level generalization). Perceptual learning can be seen not only in memory for voices, but in phoneme categorization. Norris et al. (2003) demonstrated that listeners trained on stimuli containing ambiguous /s/-/f/ tokens subsequently showed an appropriate shift in their /s/-/f/ categorization boundaries based on the lexicality of the training stimuli, and similar to Allen and Miller's result, the change was generalized to novel words. Later, Eisner and McQueen (2005) reported that perceptual learning in the /s/-/f/ continuum is speaker specific and the learning occurs at a phonemic level, because the effect was generalized to "new" speakers only when the critical phoneme actually was produced by the original speaker (but the vowel was produced by new speakers).

1.2.3 Spontaneous Imitation

In addition to repetition priming and perceptual adaptation, recent work by Goldinger provides another way to investigate the specificity effect: through *imitation.* Goldinger

(1998) tested the extreme exemplar model MINERVA 2 (Hintzman, 1986) against speech production data from a single-word shadowing task. As predicted by MINERVA 2, subjects shifted their speech production in the direction of the model speech compared with their (pre-task) baseline production (*spontaneous imitation*, also in Goldinger, 1997). His results also confirmed the word-specific advantage predicted by the model: Low-frequency words are predicted to demonstrate stronger specificity effects, because the smaller the number of exemplars associated with a given word, the larger the weight of each new exemplar. Also, a greater number of target exposures means more exemplars associated with the modeled speech for a given word, resulting in stronger specificity. As predicted, stronger imitation effects were observed for low-frequency words than for high-frequency words, as well as among subjects who were exposed to a greater number of listening trials, which was varied from 0 to 12 exposures. Later, Goldinger (2000) replicated the effects of spontaneous imitation and word-specific advantage in a non-shadowing, printed word naming task, in which speakers' post-task productions were recorded *five days* after they listened to the model speech. Shockley et al. (2004) extended Goldinger's work by showing a significant Voice-Onset-Time (VOT) imitation effect in single-word shadowing for voiceless stops with artificially extended VOTs. After being exposed to model speech with extreme values of VOT (i.e., doubled from the natural recordings), subjects increased their VOT by 12 ms on average. More recently, Delvaux and Soquet (2007) examined influence of ambient speech by employing two regiolects of Belgium French as modeled speech, and showed similar patterns of implicit imitation in segment duration and mel-frequency cepstral coefficients.

These findings suggest that reading aloud involves more than simple print-to-sound conversions (or, phonetic implementation from discrete and abstract representation); it also

taps into memory for prior perceptual episodes which contain surface information about voice. A similar result is also found in the visual domain (i.e., lower-case specificity; Jacoby and Hayman, 1987).

1.3 Two Types of Variability: Indexical and Allophonic

As described above, previous studies have shown that some types of variability in speech signals (such as voice) are kept in memory even after lexical access. From a linguistic perspective, variability in speech signals can be classified into two categories: *indexical* and *allophonic*. Indexical variability refers to variations in speech signals that arise from differences among tokens, such as talkers, dialects, speaking rates, affective states, and so on (Abercrombie, 1967; Pisoni, 1997). For a given utterance, indexical variability typically has no consequences for its denotation (Luce et al., 2003). Allophonic variability refers to articulatory and acoustic differences among speech sounds belonging to the same linguistic category (Ladefoged, 2000), such as the difference between aspirated /t/ in 'top' and unaspirated /t/ in 'stop', as well as the difference between released and unreleased /t/ in 'cat'. Although both indexical and allophonic variability carry surface information, only allophonic variability is constrained by the phonotactics of a given language. That is, allophonic variability is mostly predictable based on phonological environments, such as adjacent segments and syllable position, while indexical variability is not predictable based on phonological environments.

In practice, however, it is difficult to make a clear distinction between the two types of variability, due to the fact that they can overlap. For example, the variability between released and unreleased /t/ could be both allophonic and indexical, for it is often associated

with speech style. Nonetheless, some variations are more strongly indexical (such as voice quality in English), while some are more strongly allophonic (such as stop aspiration in English).

1.3.1 Differences between Indexical and Allophonic Variability

A body of research indicates that indexical variability (such as voice) and allophonic variability may be stored and processed differently. For example, in a dichotic listening study, Kreiman and Van Lancker (1988) showed that indexical variability is processed in the right hemisphere. The same result was also observed in an fMRI study by Belin and Zatorre (2003). On the other hand, a number of studies have shown that more allophonic, linguistic processing predominantly occurs in the left hemisphere. For example, Frye et al. (2007) recently showed that there is a neural correlate (M100) of allophonic VOT variation: Changes in VOT (varied between 25 ms and 50 ms for the synthesized syllable /pa/) linearly modulated M100, and the effect was observed predominately in the left hemisphere. Also, Phillips et al. (2000), in an ERP phonological contrast experiment (i.e., passive listening to a /ta/ - /da/ continuum in an oddball paradigm), showed that processing of phonological contrast takes place in the left hemisphere.

In addition, the two types of variability appear to differ in terms of their processing time-course. Previous studies have suggested that the indexical specificity effects emerge relatively late in processing (Luce et al. 2003; McLennan and Luce, 2005), and last for a relatively long time (*cf.* up to a week, Goldinger 1996). In contrast, phonetic (feature) priming has been reported to emerge relatively early in processing and is relatively transient. In Goldinger et al. (1992), the effect of phonetic (feature) priming (e.g., *run* - *lamb*) was

present when the inter-stimulus interval (ISI) was 50 ms, while the effect was absent when the ISI was increased to 500 ms. McLennan et al. (2003) observed allophonic (or phonetic) specificity effects (i.e., [ærəm] primed [ærəm], but not [ædəm] nor [ætəm]), only when the lexical decision task was made easy, which resulted in shorter reaction times. According to Mullenix (1997:83), voice processing is "available for a longer amount of time in an unprocessed format than phonetic information is." Massaro (1974) noted that unlike phonetic information, which appears to be highly encoded after a time lapse of 250ms, voice information that becomes available to other processes during this period seems relatively episodic.

1.3.2 Specificity of Indexical Variability

As mentioned earlier (1.2), a robust effect of specificity has been shown in a number of studies including repetition priming (e.g., voice, Mullennix et al., 1989; intonation and F0, Church and Schacter, 1994; speech style, McLennan et al., 2003; similar voices, Goldinger, 1996) and spontaneous imitation (e.g., overall voice, Goldinger, 1998, 2000). All of these variables that induced specificity were indexical.

1.3.3 Specificity of Allophonic Variability

Contrary to indexical variability, few repetition priming studies have investigated the specificity of allophonic variability. McLennan et al. (2003) provides rare evidence for the presence of arguably allophonic specificity effects: In a series of long-term repetition-priming experiments employing flaps (the neutralized allophone of /d/ and /t/ in American English, as in *ri<u>d</u>er* and *wri<u>t</u>er*), they found both allophonic and phonemic

specificity in different tasks. As described in 1.3.1, when the task was an easy lexical decision resulting in shorter reaction times, only allophonic level priming effects were observed, while when the task was made more difficult resulting in longer reaction times, allophonic and phonemic level priming effects were found (i.e., [æɾəm] primed [æɾəm], [ædəm], and [ætəm]). They interpreted this result as evidence for both surface (allophonic) and underlying (phonemic) form-based representations. Note, however, that the tested allophonic variability (i.e., flap) can also be indexical, and in fact was construed as speech style (non-flaps were labeled as 'careful' and flaps were labeled as 'casual').

Thus there appear to be significant differences in the way indexical information and allophonic information are processed and stored. However, pure exemplar models (e.g., MINERVA 2) do not readily make such a distinction. Given that ‘every’ exemplar is assumed to be preserved in memory, we would expect no difference in terms of specificity effects across the two types of variability. On the other hand, the abstractionist view would predict a clear distinction: Allophonic details, which carry crucial information to decode linguistic message, may be lost after recovering phonemes or recognizing words, while indexical details such as voice or speech rate may be retained for other purposes. That is, specificity of allophonic variability provides a testable ground for abstract vs. episodic viewpoints: If it is observed, it would provide a clear challenge to the abstractionist view of phonological representation, while it would provide support for any theory that, in contrast, claims that phonetic detail is retained in memory and used in speech production.

1.4 Units of Phonological Representations

As discussed by Pierrehumbert (2001), exemplar-based models provide an account for the attested plasticity in speech production and perception, as well as for the specificity effects observed in various tasks. Thus it is rather uncontroversial that phonological representations cannot be *purely* abstract. At the same time, the evidence that listeners can show sensitivity to episodic details does not rule out abstract representations. One reason is the difference between the two types of variability, allophonic and indexical, as just described above. Another reason why observing specificity cannot rule out the presence of abstract representations is that we do not fully understand the way phonological representations are structured in memory. That is, if phonological representations are multi-leveled, it is conceivable that each level of representation can differ in terms of its abstractness (e.g., episodic lexicon and abstract features). Therefore, in order to investigate the abstractness and episodicity of phonological representations, it is crucial to examine what phonological representations are involved in speech perception and production, and *how* abstract and episodic each level of representation is.

The psychological reality of lexical word and phoneme has been frequently demonstrated and is relatively uncontroversial (e.g., Levelt, 1989; Norris et al., 2003, Vitevich and Luce, 1998, 1999). On the other hand, support for the notion of featural representations has traditionally been provided by phonological alternations and phonotactic constraints, and their existence has been less investigated in experimental settings (except for a *gesture,* the fundamental unit of phonological representations in articulatory phonology, e.g., Goldstein et al. 2007; Browman and Goldstein, 1989). Nonetheless, Kraljic and Samuel (2006) recently showed that perceptual leaning occurs at a *feature level*: After hearing

ambiguous /d/-/t/ phonemes during a lexical decision task, listeners' categorical boundary shift was generalized to /b/-/p/ continua. Further, Maye et al. (2008) showed that subjects exposed to a novel accent (with lowered front vowels) increased their endorsement of modified forms (e.g., *wetch*) as real words in a lexical decision task following the exposure. The effect was generalized to new words containing front vowels, and, with respect to featural learning, to a lesser extent to back vowels. These are consistent with the finding by Goldrick (2004), in which learned phonotactics were generalized from /f/ to /v/. Note that in this dissertation, the term 'feature' will be used to refer to sub-phonemic representations, without distinguishing theoretically contrastive views such as distinctive features vs. gestures.

Given the previous findings on dissociations between perception and production (e.g., near-mergers by Labov, 1994), finding the production counterpart of sub-lexical generalization will provide a stronger argument for the presence of sub-lexical representations, and will yield insight into how closely perception and production interact with each other. As mentioned above, Goldinger's (1998) results of word-specific patterns of imitation provide compelling evidence for an episodic view of the mental lexicon. However, his results do not reveal whether sub-lexical units were also influenced, because the listening (= training) and production lists were identical. In other words, by making the two lists different (i.e., making the listening list a *subset* of the production list), it is possible to test if/how imitation can be generalized to unheard items. And by manipulating the degree of similarities between words that are in the listening list and words that are not (e.g., sharing the same phoneme, feature), we can test what level of representations are influenced by the imitation effects.

1.5 Current Study

1.5.1 Research Questions

It is evident that neither an abstract view nor an extreme exemplar view alone can account for all the findings in the literature. Therefore, the crucial question to be asked is not *whether* phonological representations are abstract or episodic, but rather *how* abstract and/or episodic they are, and what levels of representations are involved in the process of imitation. In order to address these questions, the current dissertation pursued the following research objectives.

1 Imitability of Allophonic Variability

Although the imitation paradigm is an interesting way to show how changes in surface information affect our speech production, the previous results of word-specificity (e.g., Goldinger, 1998, 2000) do not confirm the exemplar theories' prediction that *every* detail in the perceived speech signal is preserved in memory, because the previous results were obtained through perceptual judgments of overall voice, which mainly carries indexical information. As described earlier, if some realistic range of allophonic variability can induce phonetic imitation, it will suggest that fine phonetic details in speech signal (both indexical and allophonic) are not reduced during speech perception/word recognition and preserved in memory, providing additional support for the episodic view of phonological representations.

2 Generalizability of Phonetic Imitation

Unlike the effect of perceptual learning, which can be generalized across phoneme and natural class (see 1.4), no study has investigated whether the imitation effect can be generalized to novel stimuli. If imitation of allophonic variability can be generalized to a

different word or phoneme, it indicates that listeners code the information in incoming speech at the phoneme and/or feature level. This will further suggest that the unit of phonological representation (whether it is abstract or episodic) responsible for the imitation effect has to be smaller than a lexical word, despite the common assumption that exemplar theories are tied to lexical words.

3 Word-level Specificity of Phonetic Imitation

Another research objective is to replicate the word-level specificity observed in Goldinger (1998, 2000) through acoustic measurement of a phonetic feature, namely [spread glottis], realized as VOT (Experiments 1 and 2) and as voiced vowel duration (Experiment 3). The exemplar view predicts a stronger specificity for more recently experienced words, and thus we would expect a larger imitation effect for words to which subjects are exposed in the experiment (target) than for words to which they are not exposed (novel). The exemplar view also predicts a stronger specificity for low-frequency words than for high-frequency words, because the smaller the number of exemplars associated with a given word, the larger the weight of each new exemplar. In Pardo (2006), interacting talkers (who were engaged in a map task) increased their similarity in phonetic repertoire (*phonetic convergence*) during conversational interaction, and the convergence persisted into a post-task session. Pardo argued that episodic traces could not have caused the convergence, because the delay would increase the influence of long-term memory traces on repeated words, leading to reduced phonetic convergence. However, phonetic imitation (in her case, phonetic convergence) persisting into a post-task session does not necessarily refute the effect of episodic traces, for a detailed decaying mechanism of echoes has not been specified. In fact, Goldinger (2000)

showed that in a word naming task, the effect of imitation as well as the effect of lexical frequency and number of exposures were still present five days after the training sessions. Although the experiments in this dissertation involve a non-social setting, it is similar to Pardo's post-task session in terms of the timing (the test recording is made directly following the listening session). If we observe word-level specificity through degree of imitation, it will provide further support for the role of episodic traces in phonetic imitation and convergence.

4 Interaction between Imitation and Phonological Structure

The relationship between episodic memory (or its specificity) and phonological structure is still unknown. For example, if a given target speech is phonetically very close to a category boundary, would speakers still imitate it? Previous studies in speech perception have shown that perceptual space is warped around the category boundary (e.g., the magnet effect, Kuhl et al. 1992). Further, Ganong (1980) revealed that lexical knowledge biases lower-level phonetic perceptual processing (e.g., more /b/ responses in a /bif/-/pif/ continuum than in a /bis/-/pis/ continuum). If exemplars consist of raw percepts, target speech which approaches a phonemic boundary should still induce imitation. On the other hand, if exemplars consist of information that is already filtered by our linguistic knowledge, we would expect to see an asymmetry between target speech that does/does not approach a phonemic boundary.

1.5.2 Experiments

In order to address the research objectives described above, three experiments were conducted using a new experimental design based on the imitation paradigm (Goldinger, 1998). The first innovation of the current dissertation is that it extends the earlier studies of spontaneous imitation by using different word lists for baseline/test recordings (= production

list) and target-exposure (= listening list). A non-shadowing word-naming task was employed, which allowed the production and listening lists to differ, and further enabled the generalizability of imitation effects to be examined. Testing the generalizability of imitation is crucial in investigating the specificity and abstractness of imitation, and in examining the unit(s) of phonological representations influenced by the target exposure. The second innovation of the current dissertation is that it assesses the degree of imitation by acoustic measurements. In earlier studies (Goldinger, 1998, 2000; Pardo, 2006), the degree of spontaneous imitation or phonetic convergence was measured through AXB perceptual assessments: A different set of listeners heard three versions of the same lexical items, and judged which item produced by the shadowing talker, pre-task or shadowed token (A or B), sounded like a better imitation of the model X. While overall perceptual assessments integrate multiple acoustic phonetic dimensions and thus provide a more holistic (and thus possibly more sensitive) measure of imitation itself, they do not provide information about what is being imitated. In addition, this perceptual assessment involves subjective similarity judgments by listeners, which would introduce other factors such as listeners' attention or sensitivity to the stimuli. Not only does acoustic measurement of one phonetic dimension provide a more objective and precise measure of the imitation itself, it also provides detailed, quantitative information about the variability among speakers and lexical items. For this reason, the term *phonetic imitation* is used hereafter in this dissertation, as opposed to vocal imitation or spontaneous imitation, the terms used by Goldinger (1997, 1998).

Experiment 1 was designed to examine the imitability of target speech with lengthened VOT on the voiceless stop /p/, and its specificity and generalizability at three levels of phonological representations (i.e., word, phoneme, and feature). The manipulated

acoustic feature was VOT (voice onset time), corresponding to the phonetic feature [+spread glottis]. Word initial VOT in the target speech was lengthened within a realistic range to simulate gradient allophonic variability.

Experiment 2 aimed to determine the automaticity of phonetic imitation and its sensitivity to phonological contrast, by comparing how different types of target stimuli were imitated. Unlike lengthened VOT (as in Experiment 1), shortened VOT in voiceless stops could introduce linguistic ambiguity, namely confusion with the voiced category. If spontaneous phonetic imitation is an automatic process which reflects the collection of episodic traces, the presence of phonemic contrast should not constrain the effect and thus we would expect similar patterns of imitation for the two types of modeled stimuli. On the other hand, if different degrees of imitation are observed, it will suggest that the phonetic imitation is not automatic, but rather a more complex process that is constrained by other factors, such as knowledge of phonemic contrast.

Experiment 3 was designed to investigate the extent to which phonological factors unrelated to contrast preservation also affect imitation of fine phonetic details. The phonetic imitation of Japanese vowel devoicing, an allophonic variability with both categorical and gradient realizations of the vowels, was examined. The target speech contained extreme degrees of devoicing, which was created by manipulating (= eliminating) the duration of devoiceable vowels in devoicing environments.

1.6 Organization of the Dissertation

The remainder of this dissertation is organized as follows: Chapter 2 describes Experiment 1, in which the imitability of lengthened VOT as well as its generalizability to novel stimuli are

examined. Chapter 3 describes Experiment 2, which examines the phonetic imitation of shortened VOT and its interaction with phonemic contrast. Chapter 4 describes Experiment 3, which examines the phonetic imitation of Japanese vowel devoicing, and also examines the extent to which phonological factors unrelated to contrast preservation also affect imitation of fine phonetic details. Chapter 5 summarizes the results presented in this dissertation, and provides a general discussion as well as future research directions.

Chapter 2: Experiment 1 - Imitation of Extended VOT

Experiment 1 examined the specificity and abstractness of the spontaneous imitation effect at three levels of phonological representation, namely, word, phoneme, and feature.

As described in Chapter 1, Goldinger's (1998, 2000) results from spontaneous imitation provided compelling evidence for the effect of episodic traces in speech production, and further for an episodic view of the mental lexicon. However, his results did not reveal whether sub-lexical units were also influenced, because the listening (= training) and production lists were identical and no within-word analysis was performed.[2] The present study extended the earlier studies of spontaneous imitation by using a non-shadowing word-naming task, which allows the listening and production lists to differ and novel (or, unheard) words to be introduced into the production list. Using this modified experimental paradigm, the generalizability of the imitation effect at two levels of sub-lexical units, namely, phonemic and featural representations, was examined.

The property manipulated in Experiment 1 was VOT (voice onset time), corresponding to the phonetic feature [+spread glottis], on the voiceless stop /p/ in word-initial position. VOT refers to the interval between release of stop closure and onset of voicing, and is a standard way to measure aspiration. Figure 1 illustrates VOT measurement from a waveform ("*apa*"): As seen, VOT is measured as the temporal difference between

[2] Goldinger (1998) employed a single-word shadowing paradigm, where the listening and production lists have to be identical; Goldinger (2000) did not employ shadowing, but again the production and listening lists were identical.

stop closure release (a) and voicing onset (b). VOT was chosen because of its high inter- and intra-speaker variability (Allen et al., 2003), imitability (*cf.* Shockley et al. 2004, in shadowing), as well as its ease of acoustic manipulation for stimulus construction.

The modeled phoneme was /p/. Phoneme-level generalization was tested with novel items with the modeled phoneme /p/ in initial position. Feature-level generalization was tested with novel items with a new phoneme /k/, also in initial position. The phoneme /k/ was chosen because it falls into the same natural class (a set of segments which share acoustic or articulatory features) as /p/, sharing the modeled feature [+spread glottis].

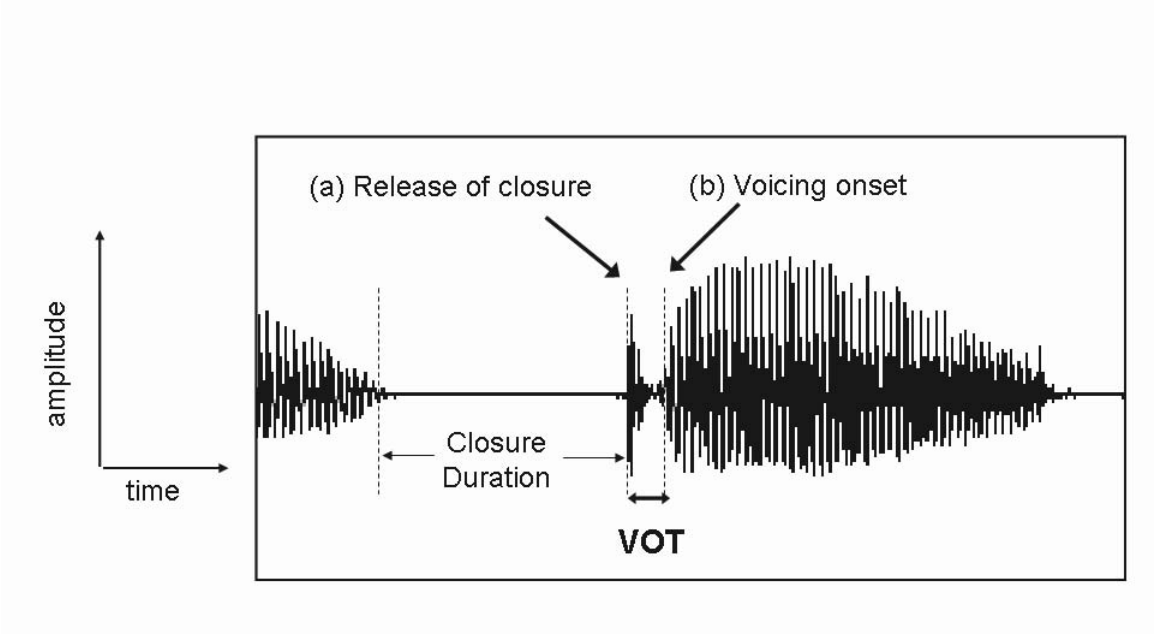

Figure 2-1: Example of measurement of Voice Onset Time (VOT) from a waveform
VOT is defined as the interval between (a) release of closure and (b) voicing onset.

Another innovation of the current study is the method used for imitation assessment, namely, acoustic measurements of recorded speech. In Goldinger (1998, 2000) as well as Pardo (2006), phonetic imitation or convergence was measured through AXB perceptual assessments. As discussed in Chapter 1, although overall perceptual assessments provide a

more configural and holistic measure of imitation, it does not provide information about what is being imitated. Acoustic measurements of one phonetic dimension provide a precise measure of the imitation, and also provide quantitative information about the variability.

The current study also aimed to replicate the word-level specificity observed in Goldinger (1998, 2000) through acoustic measurements of VOT, by comparing high and low-frequency words, and also by comparing heard (=target) and unheard words in the experiment. Lastly, the effect of participants' gender on the degree of imitation was also examined, since previous studies in phonetic imitation/convergence reported conflicting effects of gender (e.g., Namy et al., 2002; Pardo, 2006).

2.1 Method

2.1.1 Participants

Twenty-seven monolingual native speakers of American English (14 female) with no reported history of speech or hearing deficits or extensive contact with foreign languages served as participants for Experiment 1. They were recruited from the UCLA student population, and received course credit for their participation.

2.1.2 Stimuli Selection

English words with initial /p/ and /k/ were first selected from CELEX2 (Baayen et al., 1995) as test words, and were then classified into two categories: high frequency and low frequency. The lexical frequency was determined from both Kučera and Francis (1967) and CELEX2 (Baayen et al., 1995): The threshold for the low- and high-frequency category was below 5 and above 50 (per million), respectively, in Kučera and Francis, and below 300 and above 1000 for CELEX2. In addition, words with initial sonorants, which varied in lexical

frequency, were selected as fillers.

For the test words (i.e., words with initial /p/ and /k/), other lexical or phonological factors known to influence word recognition and/or production, such as phonological neighborhood density (Luce, 1986; Luce and Pisoni, 1998; Scarborough, 2004; Wright, 1997, 2004), word familiarity (Wright, 1997; Amano et al., 1999), word length, and stress pattern, were counterbalanced between the two frequency groups. The phonological neighborhood density was obtained from the Washington University in St. Louis Speech and Hearing Lab Neighborhood Database[3], and was controlled between the two frequency groups for a given syllable length (i.e., monosyllabic, disyllabic, and trisyllabic words). All words had high familiarity (> 6.0 on the 7-point Hoosier Mental Lexicon scale, Nusbaum et al., 1984). All test words (i.e., words with initial /p/ and /k/) had initial stress, while the stress pattern of fillers (i.e., words with initial sonorants) was varied. There were no words with initial onset cluster or initial voiced stops. For a complete list of words used in Experiment 1, see Appendix A.

2.1.2.1 Production List

The production list consisted of 150 words: Among them, 100 were words beginning with /p/ (80 *target* words played in the listening phase, further divided into 40 high-frequency words and 40 low-frequency words, plus additional 20 *novel* words which were not played during the listening phase), and 20 were *novel* words beginning with /k/. Each frequency group (low/high) among the target words consisted of 10 monosyllabic words, 20 disyllabic words and 10 trisyllabic words. All novel words had low frequency, and were disyllabic. The

[3] http://neighborhoodsearch.wustl.edu/Neighborhood/NeighborHome.asp

remaining 30 words were fillers with initial sonorants.

2.1.2.2 Listening List

The listening list was a subset of the production list, and consisted of 120 words, including the 80 *target* words from the production list (i.e., the 40 high-frequency words and 40 low-frequency words beginning with /p/), and 40 filler words that were different from the ones in the production list. Table 1 shows examples of stimuli used in Experiment 1. For a complete list of words used in this experiment, see Appendix A.

Table 1: Examples of test stimuli used in Experiment 1*

Frequency / Stimulus Type	Target /p/ (80)	Novel /p/ (20)	Novel /k/ (20)
High	parent, power		
Low	pebble, pirate	pillar, portal	canine, kosher

***Participants only listened to target words (shown in the darker grid) during the listening phase, while they produced all the test words in both baseline recording and post-exposure recording blocks.**

2.1.3 Stimulus Construction

A phonetically trained male American English speaker recorded the 80 target (/p/ initial) words in the listening list. The speaker first produced the words in the list normally, and then one more time with *extra aspiration*. His speech was digitally recorded, and the VOTs for the normally produced initial /p/ were measured (mean=72.46 ms, SD=12.14 ms). In order to lengthen VOT, the following manipulation was conducted: The initial parts of hyper-aspirated tokens (i.e., the stop burst and aspiration) and the normally produced tokens

(excluding the stop burst and aspiration) were spliced using PCquirer (Scicon R&D, CA) such that the resulting VOT was *lengthened* by 40 ms (mean=113.26 ms, SD=10.82 ms). Figures 2 and 3 illustrate examples of the word "pass" *before* and *after* the VOT manipulation, respectively: As seen in Figure 2, the naturally produced VOT of the word was 68ms. The burst and aspiration portion (① in Figure 2**)** was replaced by that of hyperaspirated counterpart (② in Figure 3) in order to lengthen the VOT by 40 ms (which thus also lengthened the overall word duration by 40 ms). The point of splicing was not exactly at the voice onset, but typically around 10 to 20 ms *before* the voice onset, depending on the transition pattern of each token. This splicing method was chosen, as opposed to extending the middle part of VOT as in Shockley et al. (2004), in order to maximally preserve natural formant transitions from burst to aspiration as well as from aspiration to voice onset. To ensure that the VOT range of the manipulated stimuli was substantially longer than that of the normally produced tokens, the minimal VOT threshold was set at 100 ms for the after-manipulation tokens, and consequently further lengthening was conducted for the seven words whose original VOTs were below 60 ms.

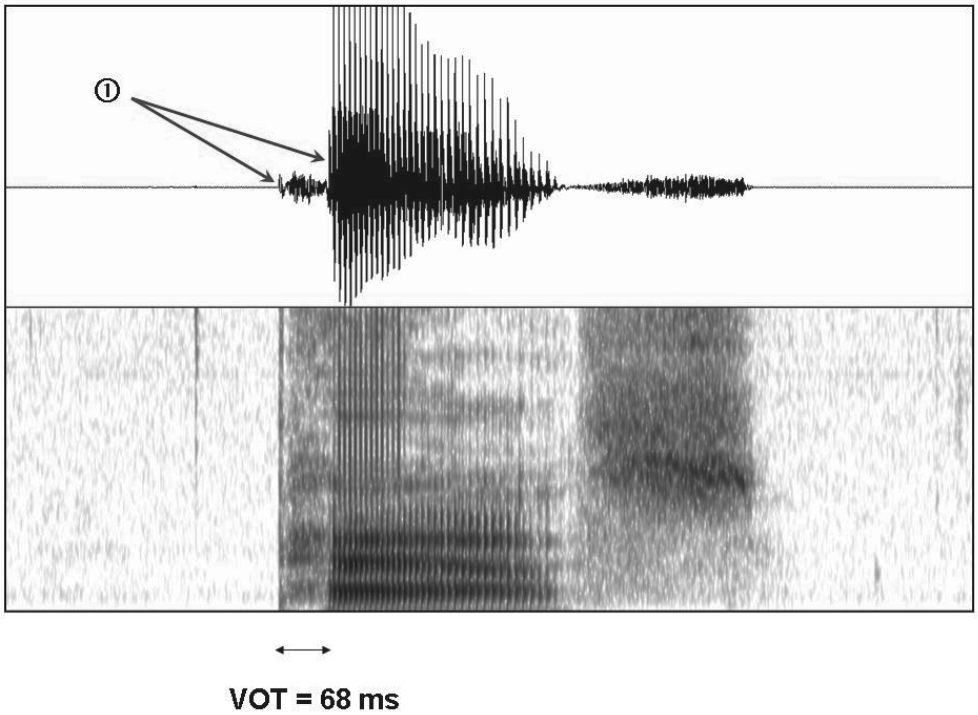

Figure 2-2: Example waveform and spectrogram of word "pass" before the VOT manipulation
The original VOT (①) was 68 ms.

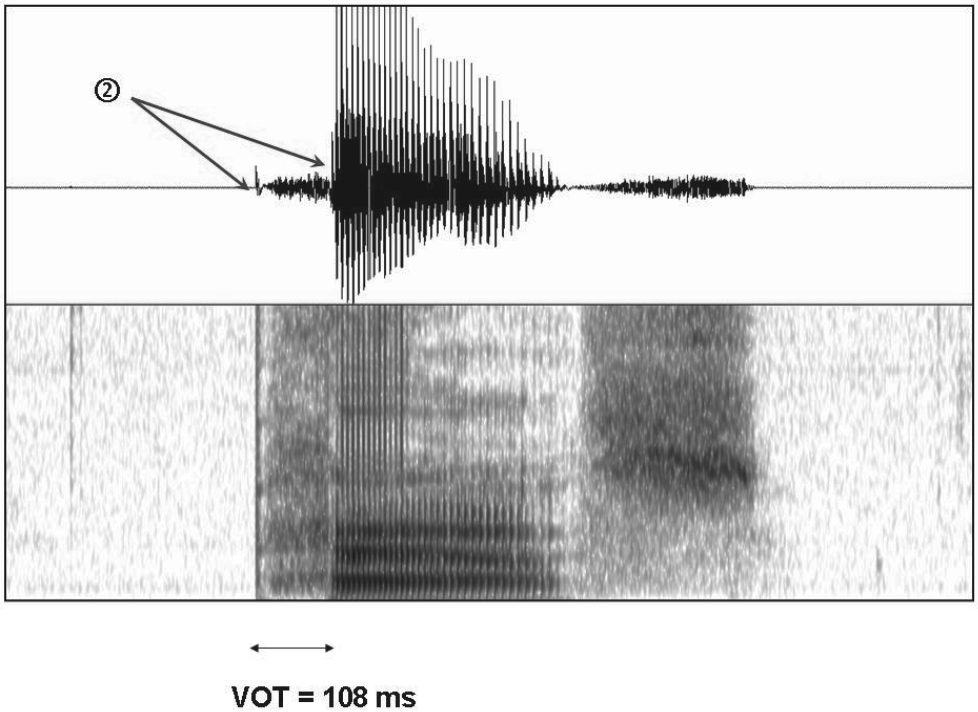

Figure 2-3: Example waveform and spectrogram of word "pass" after the VOT manipulation, used as one of the listening stimuli in Experiment 1
After lengthening VOT by 40 ms, the new token's VOT was 108 ms.

2.1.4 Procedure

The experiment used a slightly modified version of the word-naming imitation paradigm (Goldinger, 2000), in that a warm-up silent reading block was added prior to the recording. This modification was made in order to reduce possible hyper-articulation for first readings of low frequency words in the baseline recording, which was observed in our pilot data. The experimental stimuli were presented using Psyscope 1.2.5 (Cohen et al., 1993). Each participant was seated in front of a computer in a sound booth located in the UCLA Phonetics Laboratory. Each session was divided into 4 blocks: 1) warm-up, 2) baseline recording, 3) listening, and 4) post-exposure recording. In the warm-up block, the words in the production

list were visually presented on a computer screen, one at a time, every 2 seconds, and the participants were asked to read the words *silently* without pronouncing them. In the baseline recording block, the same list of words were presented again, but this time the participants were asked to read the words aloud into a TELEX M-540 microphone, providing a baseline recording. The actual wording of the instruction displayed on the computer screen was the following: "Please identify the word you see on the screen by speaking it into the microphone." In the listening block, the participants were instructed to listen carefully to two repetitions of the words in the listening list using SONY MDR-V250 headphones. There was no additional task during this block. Finally, in the post-exposure recording block, the participants read the words (in the production list) aloud for the second time, providing another recording. Across the four blocks, the words were presented in random order for each subject. Participants' tokens were digitally recorded into a computer at a sampling rate of 22100 Hz, and later the word-initial VOTs were measured by the author and a phonetically-trained research assistant. Both waveforms and spectrograms were examined for the acoustic measurements, using PCQuirer (Scicon R&D, Inc.).

2.1.5 Statistical Analysis

The within-subjects factors in Experiment 1 were 1) Production Type (Baseline vs. Post-Exposure), 2) Lexical Frequency (High vs. Low), 3) Presence of Exposure (Target vs. Novel), and 4) Segment (/p/ vs. /k/). Gender was included as the between-subjects factor. VOT was the dependent variable. Statistical by-subjects analysis of the data was based on analysis of variance tests (ANOVA) performed with the software package SPSS.

2.2 Results and Discussion

Table 2: Summary of Experiment 1 Results

Stimulus Type	Production Type	Mean (ms)	Median (ms)	Std. Error of the Mean (ms)	Std. Deviation (ms)
Target /p/ low	**Baseline**	65.73	60.13	2.98	15.50
	Post-exposure	73.41	75.89	3.18	16.53
Target /p/ High	**Baseline**	65.76	59.91	3.05	15.85
	Post-exposure	72.09	69.05	3.18	16.51
Novel /p/ Low	**Baseline**	63.73	61.31	2.88	14.95
	Post-exposure	70.27	69.60	2.64	13.73
Novel /k/ Low	**Baseline**	75.87	74.50	2.69	13.99
	Post-exposure	80.66	83.30	2.52	13.12

The mean, median, standard error and standard deviation of VOT (ms) are shown by stimulus types. As can be seen, the standard deviations are large in general, due to the individual variability of VOT. On the other hand, the standard errors of means are quite small, showing that the participants' shifts in their production (= imitation) were consistent.

2.2.1 Phonetic Imitation

Table 2 shows a summary of the results from Experiment 1. A repeated measures ANOVA analysis with Production Type (Baseline vs. Post-Exposure) as the within-subjects factor and Gender as the between-subjects factor revealed a significant effect of Production Type (= phonetic imitation) [$F(1, 25)=16.53$, $p<0.01$*] as well as Gender [$F(1, 1)=12.60$, $p<0.01$*], while the interaction between the two factors was not significant [$F<1$, $p>0.1$].

Figure *2-4* shows the effects of phonetic imitation and gender on VOT, and plots VOT in milliseconds under Production Type (Baseline vs. Post-Exposure) and gender (female and male). The light bars show the mean VOT of the baseline productions, and the dark bars show the mean VOT of the post-exposure productions. The error bars represent the standard

error of the mean. As seen in the figure, post-exposure productions show longer VOTs than baseline productions for both genders, revealing that the VOT imitation effect is present even when the task involves non-shadowing elicitation-style production. This finding is consistent with previous studies (e.g., Goldinger, 2000); about eight minutes after the participants heard the target speech, they sustained and imitated the modeled speech's detailed surface phonetic information (i.e., extended aspiration), without being instructed to do so.

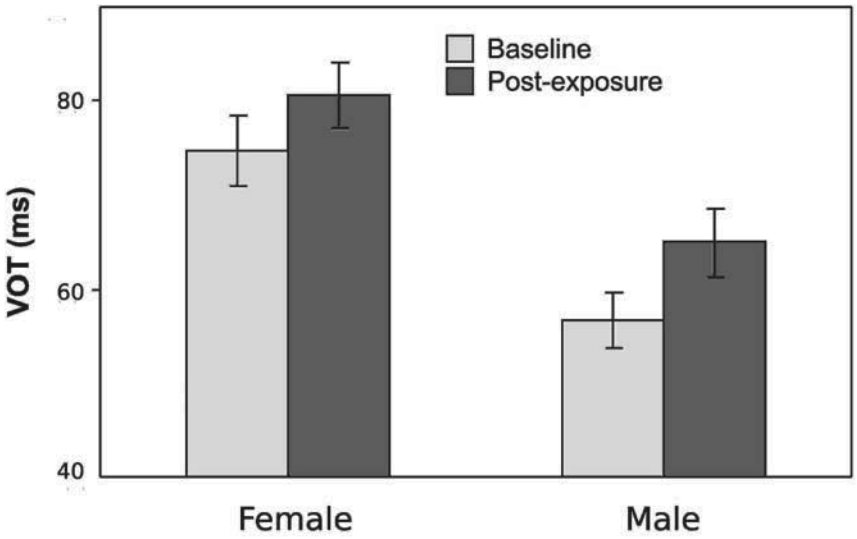

Figure 2-4: Phonetic imitation and Gender difference on VOT
Mean VOT and standard error of the mean in Experiment 1 (target = lengthened VOT), presented separately for the baseline and post-exposure productions, and for participants' gender, showing that 1) lengthened VOT was imitated, and 2) overall VOT was longer for female participants.

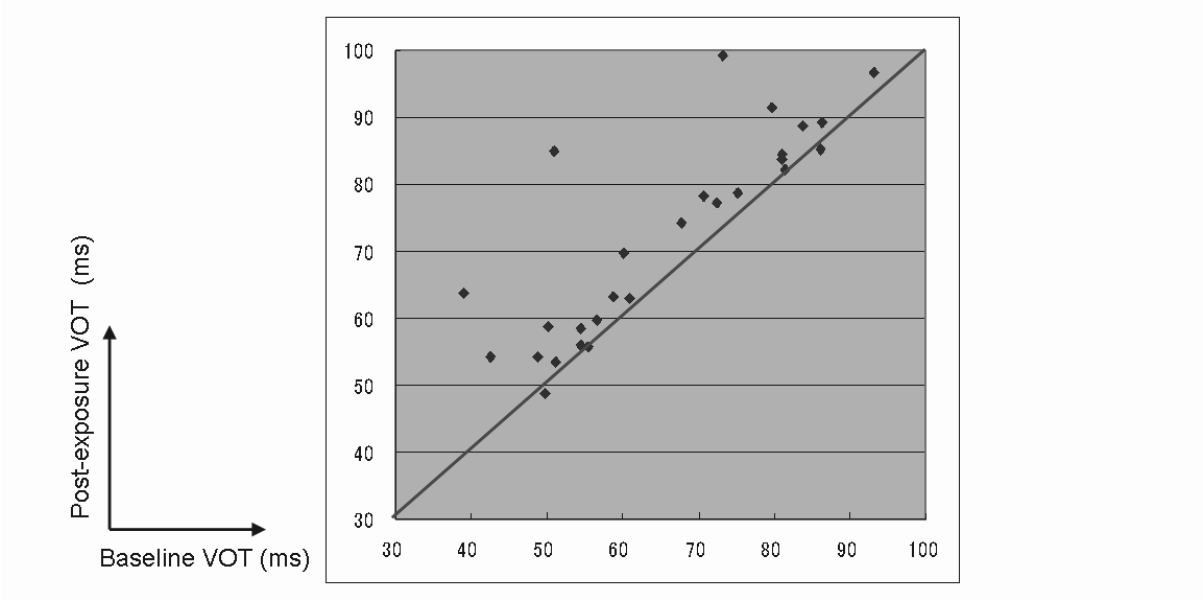

Figure 2-5: Individual variability in VOT and phonetic imitation
Mean VOT of individual participants in Experiment 1 (Lengthened VOT). The baseline VOT is plotted on the x-axis, and post-exposure VOT is plotted on the y-axis (R^2=0.71). The diagonal line's slope is one and indicates no imitation (baseline = post-exposure).

In addition to the gender difference in VOT, the data revealed individual variability in VOT as well as degree of imitation. Figure 2-5 illustrates the speaker variability in VOT and degree of imitation in Experiment 1. As can be seen, the mean VOT ranged from 39 ms to 99 ms. Although most participants increased their VOT after they listened to the target speech with lengthened VOT, the magnitude of the changes varied from participant to participant.

It could be hypothesized that the imitation is due to global changes in the subjects' speech style. That is, after being exposed to the target stimuli with extended VOT, subjects may have changed their register to more "careful speech". An argument against this interpretation is provided by a post-hoc analysis of whole-word duration. If the observed

change in VOT is due to a shift in distribution of exemplars, only the manipulated variable should be affected. On the other hand, if the change is due to more careful speech, we would expect to see overall longer word duration. For this reason, the whole-word duration of the low-frequency target words was measured from 8 randomly chosen participants' data. While their VOT showed a significant difference between baseline and post-exposure productions in a paired two-tailed t-test [$t(1,7)=4.67$, $p<0.01$*], there was no significant difference in whole-word duration between baseline and test productions ($t<1$, $p>0.1$). Given these results, it is unlikely that global aspects of speech are solely responsible for the spontaneous phonetic imitation observed here.

2.2.2 Gender Difference in VOT

The results from Experiment 1 also revealed a significant gender difference in VOT: As seen in Figure 2-4, overall VOT was longer for female participants than male participants [$F(1, 1)=12.60$, $p<0.01$*]. This result is consistent with previous literature which claimed that females tend to have longer VOT values for voiceless stop consonants than do males (e.g., Ryalls *et al.*, 1997; Swartz, 1992; Whiteside *et al.*, 2004). Note, however, that a strong correlation between VOT and speaking rate has also been reported (Allen et al., 2003), and thus the observed gender difference in VOT has to be taken with caution. In Allen et al. (2003), the same pattern of gender difference was observed in their absolute VOT data, however, this gender difference was not evident when differences among the talkers in speaking rate were controlled. The current experiment did not control speaking rate, and thus further investigation is required to determine gender difference in VOT.[4]

[4] The post-hoc analysis of whole-word duration (1.2.1) also showed a significant gender difference of word

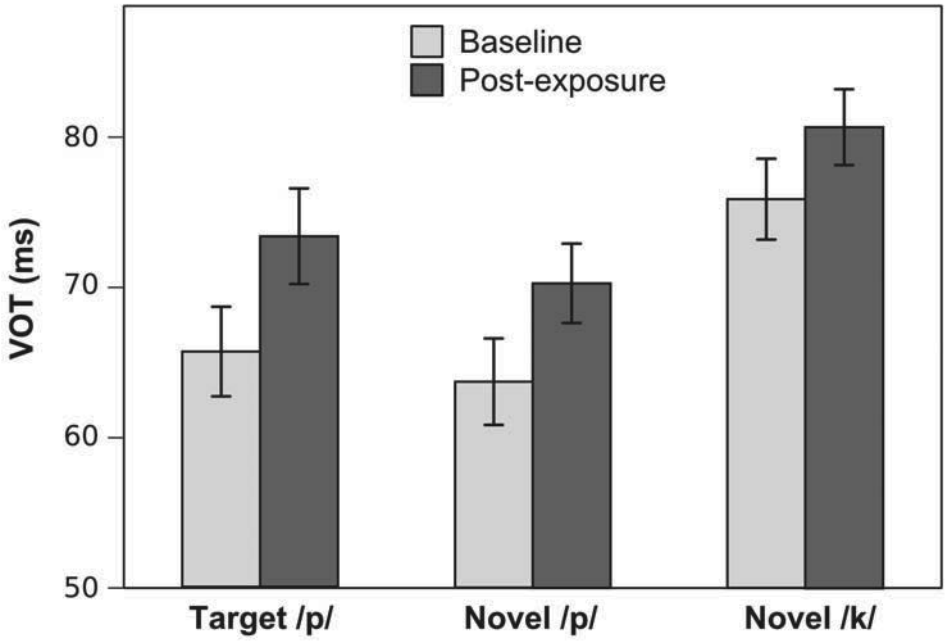

Figure 2-6: Generalizability of spontaneous imitation
Mean VOT and standard error of the mean in Experiment 1 (Lengthened VOT), presented separately for the baseline and post-exposure productions, and for the three types of stimuli. The participants heard only Target /p/. For each stimulus type, the difference between baseline and post-exposure VOTs was significant.

2.2.3 Generalizability

In order to test how the imitation effect is generalized to new stimuli with the same initial phoneme /p/, a repeated-measures ANOVA analysis was performed with Production Type and Presence of Exposure as the within-subjects factors, and Gender as the between-subjects factor. Note that this comparison was made between only the twenty novel words (all disyllabic with initial /p/ and low lexical frequency) and the twenty disyllabic target words with low frequency (see Table 2, and also *Target /p/* and *Novel /p/* in Figure 2-6). Significant

duration (single-factor ANOVA:[$F(1,1)=11.36$, $p<0.05$*]), supporting Allen et al.(2003).

main effects for both Production Type [F(1,25)=19.71, $p<0.001$*] and Presence of Exposure [F(1,25)=9.62, $p<0.01$*] were observed, showing that participants' VOT was longer for post-exposure production than baseline (=imitation), and that their VOT was also longer for Target /p/ words than Novel /p/ words. On the other hand, the interaction between the two factors was not significant [F(1,25)=1.42, $p>0.1$], indicating that the degree of imitation was the same across the two stimulus types. That is, lengthened VOT on the modeled phoneme /p/ was imitated and generalized to other /p/-initial words which participants did not listen to. The effect of gender was significant [F(1,1)=11.79, $p<0.01$*], while there was no interaction with either within-subject factor [F<1, $p>0.1$].

Another repeated measures ANOVA was performed with Production Type (baseline vs. post-exposure; imitation) and Segment (/p/ vs. /k/) as the within-subjects factors to determine whether the imitation effect is generalized to the new segment /k/. Note that both groups of stimuli were nested under *novel* items, and thus neither group of words tested here was played in the listening block. Similar to the results for items that were played in the listening block, there was a significant effect of imitation [F(1,25)=23.58, $p<0.01$*]. The effect of Segment (/p/ vs. /k/) was also significant [F(1,25)=271.15, $p<0.01$*], as expected due to their normally-differing VOT among places of articulation (e.g., Lisker and Abramson, 1964; Cho and Ladefoged, 1999; Zue, 1980), while there was no significant interaction between the two factors [F(1,25)=1.03, $p>0.1$]. The effect of gender was significant [F(1,1)=8.11, $p<0.01$*]. The interaction between gender and segment was significant [F(1,25)=10.11, $p<0.01$*] while there was no interaction between gender and Production Type [F(1,25)=1.43, $p>0.1$].

That is, the degrees of imitation observed among Novel /p/ words and Novel /k/ words were statistically same. Thus, the imitation of lengthened VOT on the modeled phoneme /p/ was generalized to /k/ initial words.

In sum, the imitation effect was generalized to novel stimuli that participants did not hear during the listening block: Compared with their baseline, the participants produced significantly longer VOT in the post-exposure production for the novel words with initial /p/ (see *Novel /p/ Low* type in Figure 2-6). This result indicates that the locus of spontaneous phonetic "imitation" is sub-lexical. More importantly, our data also showed that the imitation effect was further generalized to a new phoneme /k/ (see *Novel /k/ Low* type in Figure 2-6), which shares the manipulated feature [+spread glottis]. Consistent with Kraljic and Samuel (2006), there was no interaction between the degree of shift (in our case, imitation) and tested segment [$F(1,25)=1.03$, $p>0.1$], indicating that the degree of imitation for the two segments was not significantly different. This suggests that the locus of the imitation observed in Experiment 1 is at a *feature level*, because if the imitation occurred at *both* phoneme *and* feature level, we would expect to see a stronger imitation effect for Novel /p/ words (which share the modeled phoneme) than Novel /k/. Therefore, with respect to a phoneme-level representation, our result is inconclusive.

2.2.4 Word Specificity

In order to replicate the word-level specificity effects found in Goldinger (1998, 2000) through acoustic measurements of a phonetic feature, the current study controlled both *Lexical Frequency* and *Presence of Exposure* as independent variables. Word specificity would mean that low frequency words show a stronger imitation effect, and likewise words

with more exposure (in the current study, target as opposed to novel). Statistically speaking, word specificity would mean significant interactions between imitation (Production Type) and Lexical Frequency as well as imitation and Presence of Exposure. In order to test the effect of Lexical Frequency, a repeated-measures ANOVA analysis with two within-subjects factors (Production Type and Lexical Frequency) and a between-subjects factor (Gender) was performed. Although the result revealed a significant main effect of Production Type [$F(1,25)=16.53$, $p<0.001$*], neither the effect of Lexical Frequency [$F(1,25)=2.19$, $p>0.1$] nor the interaction between the two factors [$F(1,25)=2.88$, $p>0.1$] was significant. The effect of gender was significant [$F(1,1)=12.6$, $p<0.01$*], while there was no interaction with either within-subject factor [$F<1$, $p>0.1$].

Similarly, although a significant main effect of Presence of Exposure was found, the expected interaction between the factor and Production Type was not significant [$F(1,25)=1.42$, $p>0.1$] (see *2.2.3, Generalizability* above). That is, the effect of word specificity found in Goldinger (1998 and 2000) was not obtained in our data, either through lexical frequency or number of exposures. As described in 2.2.3, the effect of gender was significant [$F(1,1)=11.79$, $p<0.01$*], while there was no interaction with either within-subject factor [$F<1$, $p>0.1$].

2.2.5 Post-hoc Analysis of Word Specificity

Given that the effect of word-specificity is tested through the magnitude of imitation in the current study, the presence of clear imitation is prerequisite. However, as described in 2.2.1 and seen in Figure 2-5, large individual variability in degree of imitation was observed. In fact, there were some participants who did not imitate at all, confounding possible effects.

For this reason, a post-hoc analysis was performed only for the participants whose average VOT in the post-exposure phase increased by more than 5% from their baseline production. Among twenty-seven participants in Experiment 1, fifteen participants (5F and 10M) met this criterion. A repeated-measures ANOVA analysis with two within-subjects factors (Production Type and Lexical Frequency) and a between-subjects factor (Gender) revealed a significant interaction [$F(1,13)=5.78$, $p<0.05$*], as well as significant main effects of both Production Type [$F(1,13)=16.20$, $p<0.01$*] and Lexical Frequency [$F(1,13)=7.58$, $p<0.05$*]. The effect of gender was not significant [$F(1,1)=3.44$, $p>0.05$], and there was no interaction with either within-subject factor [$F<1$, $p>0.1$]. Figure 2-7 summarizes the result of the post-hoc analysis, and plots mean VOT and standard error of the mean for the Low and High frequency stimulus types. As seen, the degree of imitation (the difference between baseline and post-exposure productions) is greater for low-frequency words. This result indicates that lexical frequency significantly affected the magnitude of phonetic imitation, as predicted by the exemplar view. Recall, however, that this result is found only in the post-hoc analysis, and not in the original planned analysis of all subjects' data.

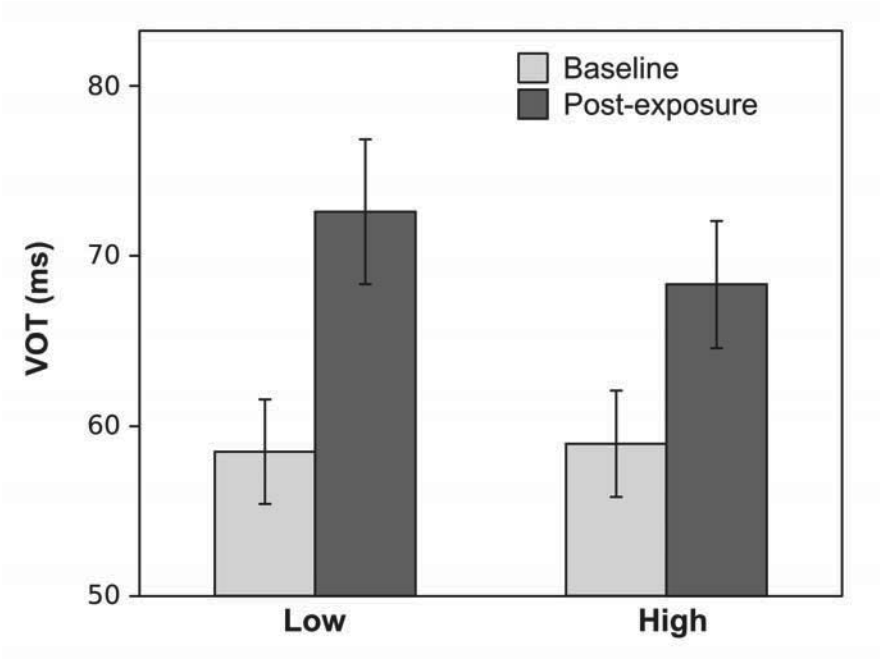

Figure 2-7: Effect of lexical frequency on degree of imitation
Mean VOT and standard error of the mean (for participants whose average VOT increased by more than 5%), presented separately for the baseline and post-exposure productions, and for the Low and High frequency stimulus types (both types were heard by the participants). There was a significant effect of lexical-frequency: Greater imitation was observed among low-frequency words than high-frequency words.

In order to test the effect of word-level specificity through Presence of Exposure (Target vs. Novel Items), another post-hoc repeated-measures ANOVA analysis was performed for the same subset of subjects, with Production Type and Presence of Exposure as within-subject factors, and Gender as a between-subject factor. As expected, the result revealed significant main effects of the two factors (Production Type: [$F(1,13)=19.57$, $p<0.001$*], Presence of Exposure: [$F(1,13)=7.41$, $p<0.05$*]), as well as a significant interaction between the two factors [$F(1,13)=4.93$, $p<0.05$*]. The effect of gender was not significant [$F(1,1)=3.3$, $p>0.05$], and there was no interaction with either within-subject factor [$F<1$, $p>0.1$].

Figure 2-8 2-8 summarizes the result of the post-hoc analysis, and plots mean VOT and standard error of the mean for Target and Novel stimulus types. As seen, the degree of imitation (the difference between baseline and post-exposure productions) is greater for Target words, which participants heard during the experiment. This result indicates that Presence of Exposure significantly affected the magnitude of phonetic imitation, once again confirming the prediction by the exemplar view. Again, recall that this result is found only in the post-hoc analysis. On the other hand, the interaction between Production Type and Segment (/p/ vs. /k/) remained not significant in this post hoc analysis [$F<1$, $p>0.1$] (see 1.2.3 above), leaving the question of phoneme-level representation inconclusive.

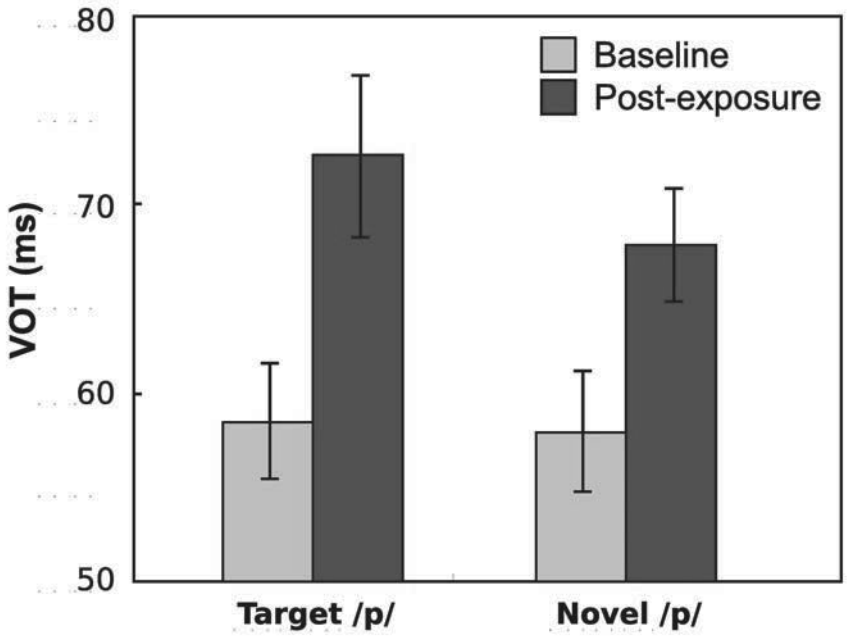

Figure 2-8: Effect of presence of exposure on degree of imitation
Mean VOT and standard error of the mean (for participants whose average VOT increased by more than 5%), presented separately for the baseline and post-exposure productions, and for the target and novel stimulus types (only target type was heard by the participants). The degree of imitation was larger for the target words which participants listened to, showing a word-specific pattern of imitation.

2.3 Discussion

2.3.1 Generalization of VOT Imitation

Our results showed that participants retained and imitated the modeled speech's detailed surface phonetic information without any overt instructions in a non-social setting. This result is in agreement with previous studies on spontaneous phonetic imitation (Goldinger, 1998, 2000, Shockley et al., 2004) and phonetic convergence (Pardo, 2006). More importantly, by introducing a new experimental design for eliciting spontaneous imitation based on a non-shadowing task, we have demonstrated that the "imitation" was *generalized* to new instances of the modeled phoneme /p/, as well as to a new phoneme /k/, neither of which were listened to by the participants during the experiment. This suggests that fine phonetic details in perceived speech signals (e.g., greater degree of [+spread glottis]) are coded at a feature level, and that this coding is systematically reflected in speech production.

This finding is consistent with the assumption in generative linguistics (and in many other linguistic theories) that words are made up of discrete speech sounds (phonemes) that are themselves complexes of features (e.g., Chomsky and Halle, 1968). Although the psychological reality of lexical word and phoneme has been frequently demonstrated and is relatively uncontroversial (e.g., Norris et al., 2003, Vitevich and Luce, 1998, 1999), the existence of featural representations has been less investigated. Along with recent experimental studies that suggest that featural representations are involved in phonotactic learning (Goldrick, 2004) and perceptual learning (Kraljic and Samuel, 2006), the result in the current study provides support for the notion of abstract featural representations in

phonetic imitation. Our results did not show a significant interaction between the degree of imitation and tested segment (i.e., /p/ and /k/) in either the main or the post-hoc analyses. This indicates that the degree of imitation for the two segments was the same, suggesting that the locus of the imitation observed in the current study is at a feature level (*cf.* Kraljic and Samuel, 2006). Note, however, that these results do not necessarily provide support for distinctive features per se. It is entirely possible that the imitated unit was a gesture (Browman and Goldstein, 1989; Shockley et al., 2004; Fowler et al., 2003) which specifies degree and duration of glottal opening. These two theoretically contrastive views (features vs. gestures) cannot be distinguished in the current study.

2.3.2 Word Specificity and Episodic Traces

Goldinger (1998, 2000) revealed word-level specificity in spontaneous imitation, and argued that detailed episodes constitute the basis of the mental lexicon. In addition to our finding that phonetic imitation was generalized at a feature level, our data also showed word-level specificity of phonetic imitation, replicating Goldinger's findings. As described in Chapter 1, exemplar theories (at least those assuming lexical representation) predict a greater degree of imitation among low-frequency words than high-frequency words, because low-frequency words have a smaller number of exemplars associated with them, and thus the weight of each new exemplar is relatively strong. As expected, then, the post-hoc analysis in Experiment 1 revealed that, for speakers who imitated, low frequency words which had been heard in the listening block showed a greater shift in VOT than high frequency words which had also been heard in the listening block. Exemplar theories also predict a greater degree of imitation among the target words (which had been heard in the listening block) than the novel words

(which had *not* been heard in the listening block), because the exemplar view predicts that the modeled feature (i.e., greater degree of [+spread glottis]) is more strongly associated with the target words than the novel words. As expected, a significant interaction between degree of imitation and presence of exposure (target vs. novel) was found in the post-hoc analysis. These specific patterns suggest that the participants' shift in their VOT cannot solely be an artifact of some global effect, and that traces of episodic memory were at least in part responsible for the imitation of VOT. Pardo (2006) argued that episodic traces could not have caused the phonetic convergence that persisted into her post-task session, because the delay would increase the influence of long-term memory traces on repeated words, predicting reduced phonetic convergence. However, the result of word-level specificity found in the current study shows that episodic traces are indeed present in *non-shadowing* post-exposure production, which is similar to Pardo's post-task session in terms of its timing, suggesting a role for the exemplar-based lexicon in phonetic imitation.

It should be noted, however, that although the effect of imitation was observed in the main analysis, the effect of word specificity was observed only in the post-hoc analysis (i.e., no interaction was observed between degree of imitation and lexical frequency or exposure in the main analysis), suggesting a relatively weak role of episodic traces in phonetic imitation. Several factors may have contributed to this result. First, as described above, the degree of imitation varied considerably among participants: Some did not show any change between baseline and test recordings, and thus could not show any specificity. Second, the current study included a warm-up phase, which was absent in Goldinger (1998, 2000). Previous studies have reported interactions of word-repetition and frequency (e.g., Scarborough et al, 1977; Norris, 1984; Goh and Pisoni, 1998): If high and low frequency words are presented

multiple times in a perceptual task, the frequency effect is steadily reduced. Recently experienced exemplars are assumed to have a stronger echo, and thus the presence of a warm-up phase might have reduced the effective difference between high and low frequency words. Third, it is also noteworthy that Goldinger (1998, 2000) manipulated the number of training exposures from none to twelve, while the current study compared none (= novel words) and two (= target words). Although there was a clear interaction between the strength of imitation (i.e., correct AXB judgment) and the number of repetitions in Goldinger (1998, 2000), the observed difference between zero and two exposures was rather small in those studies. Lastly, the current study estimated the strength of imitation from physical measurements of VOT, as opposed to perceptual judgment (i.e., AXB judgment). Among many acoustic features in the modeled speech signal (which were all available for the imitation participants to imitate, as well as for the AXB judgment participants to evaluate), VOT is the only feature manipulated and measured in this study. It is conceivable that measuring one acoustic feature is not as sensitive as an overall perceptual assessment with regard to showing word-level specificity.

2.3.3 Gender Difference in VOT and Imitation

As discussed in the Results section, our baseline data showed longer mean VOT by female participants, supporting previous claims in the literature that female English speakers tend to have longer VOT for voiceless stops than male speakers (e.g., Ryalls *et al.*, 1997, Whiteside *et al.*, 2004). At the same time, we observed a larger number of male participants who demonstrated clear imitation than female participants (see the Results section for post-hoc analysis of word-specificity). With regard to the gender difference of phonetic

imitation/convergence, conflicting observations have been reported in the literature: Namy et al. (2002) found women to be more likely than men to imitate in single-word shadowing, while Pardo (2006) found greater phonetic convergence on the part of male talkers.

Given that female speakers produced higher VOT values in general, and that the model speaker in the current study was male, there are at least two possible accounts for the gender difference observed in our data. First, the longer baseline VOT of female participants could have caused the observed gender difference. Their speech was more similar to the manipulated speech stimuli (i.e., lengthened VOT), and thus the modeled feature could have been perceived as less salient, receiving smaller attentional weight. Second, it is also possible that listeners are more likely to imitate/accommodate speech input produced by a speaker of the same gender (*cf*. Giles et al., 1999) and thus the gender of the model speaker caused the observed gender difference. In this case, the mechanism underlying phonetic imitation would need to have the capacity to accommodate paralinguistic factors such as gender. However, these two accounts are indistinguishable in the current study.

2.4 Interim Summary

The results from Experiment 1 demonstrated that spontaneous phonetic imitation is simultaneously across-the-board *and* word-specific. Imitation of lengthened VOT on the modeled phoneme /p/ was generalized to new instances of /p/ as well as a new phoneme /k/, suggesting that an abstract featural representation is involved in phonetic imitation. At the same time, lexical frequency as well as presence of exposure significantly influenced the degree of imitation, indicating that the mental lexicon is (at least in part) comprised of detailed episodic traces, as predicted by exemplar models. These findings suggest that

multiple levels of phonological representations are simultaneously at work (i.e., lexical and featural representations) which can differ in their degree of abstractness and episodicity. In addition, a suggestive gender difference in *likelihood* of imitation was observed, indicating that social factors such as gender may modulate our perception and/or imitation.

Chapter 3: Experiment 2 - Imitation of Reduced VOT

The results from Experiment 1 showed that artificially lengthened VOT on the modeled phoneme /p/ was imitated, and that the change in participants' speech was further generalized to novel items with initial /k/, revealing that phonologically irrelevant, fine phonetic detail of the modeled speech (i.e., greater degree of [+spread glottis]) was coded at a feature level. The current chapter describes Experiment 2, which examined the interaction between phonetic imitation and the presence of a phonological boundary, by testing whether phonetic imitation is still observed for target speech with *shortened* VOT on the same phoneme /p/.

The phonological contrast of voicing (the contrast between [+voice] and [-voice]) for stops can be realized in a number of phonetic/acoustic features, including VOT, closure duration, duration of vowel, F1 offset frequency, fundamental frequency (F0), release burst intensity, and so on (Lisker, 1986). Among them, VOT has been regarded as an effective measure for differentiating stops with different laryngeal actions in a wide variety of languages, including English (e.g., Lisker and Abramson, 1964). Figure 3-1 illustrates the distribution of utterance-initial VOT of English /b/ and /p/ [reprinted from Figure 3 in Keating (1984)]. The figure shows that VOT effectively differentiates the two categories, and that the phonological boundary (i.e., /b/-/p/ boundary in English) lies around 30 ms. Similar values for a /b/-/p/ boundary have been reported, e.g., Allen and Miller (2001), in which the boundary value was 34 ms for *beace-peace* stimuli, and 40 ms for *beef-peef* stimuli. Note that a categorical boundary is not a point on a continuum but rather a range, and other factors (e.g., the lexical status of stimuli, Ganong, 1980; speaking rate, Allen and Miller, 2001) are

known to influence the categorical decision.

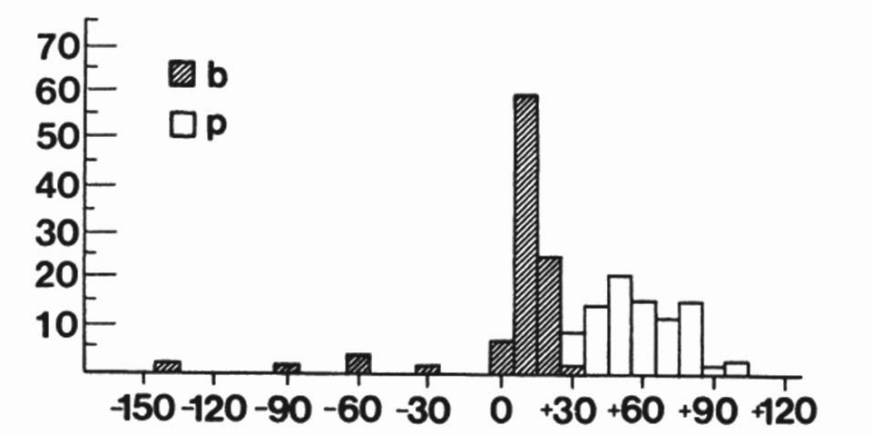

Figure 3-1: [Reprinted from Keating (1984), part of Figure 3] VOT values for English utterance-initial bilabial stops before twelve vowels
It shows that the category boundary between /b/ and /p/ on the VOT continuum is around +30 ms.

Linguistically speaking, imitation of shortened VOT is not parallel to imitation of lengthened VOT: For voiceless stops in English, increasing VOT (as in Experiment 1) has no *phonological* consequences, since there is no phonemic category along the direction of increasing VOT. On the other hand, shortening VOT might impair phonemic contrast with the voiced category (e.g., /p/ vs. /b/) which could further introduce lexical/semantic ambiguity (e.g., *pear* vs. *bear*). Contrast preservation is an essential part of phonological grammar (Flemming, 2001; Lubowicz, 2003), and thus shortening VOT should be phonologically highly marked. If phonetic imitation is a process which is sensitive to linguistic structure/grammar (as assumed by the abstractionist view), we would expect attenuated phonetic imitation for modeled speech with shortened VOT.

In contrast, if the mechanism of phonetic imitation is purely episodic and directly

reflects perceived episodic traces, we would expect the process to be rather automatic. In this case, exposures to the target speech should change the distribution of episodic traces regardless of the manipulation of the target speech, and thus should induce phonetic imitation of shortened VOT.

3.1 Method

The experimental method in Experiment 2 was identical to Experiment 1, except that the acoustic manipulation for the target speech (i.e., listening stimuli) was done in the opposite way, namely the word-initial VOT was *shortened* by 40 ms from the original tokens.

3.1.1 Participants

Seventeen monolingual native speakers of American English (9 females) with no reported history of speech or hearing deficits nor extensive contact with foreign languages served as participants for Experiment 2. They were recruited from the UCLA student population, and were paid or received course credit for their participation.

3.1.2 Stimuli Selection

The same lists as in Experiment 1 (both production and listening) were used in Experiment 2.

3.1.3 Stimulus Construction

The same original recordings of normally produced target words (used in Experiment 1) were used to construct listening stimuli for Experiment 2 (mean VOT =72.46 ms, SD=12.14 ms). In Experiment 2, VOT was *shortened* by eliminating the most stable part of the aspiration. Figure 3-2 and 3-3 illustrate examples of the word “pass” *before* and *after* VOT manipulation, respectively: As seen in Figure 3-2, the naturally produced VOT of the word was 68 ms. The

most stable portion of aspiration (① in Figure 3-2) was *deleted* so that the resulting token's VOT was reduced by exactly 40 ms from the original (Figure 3-3). Similar to Experiment 1, the *maximal* VOT threshold was set at 40 ms to ensure that VOTs of manipulated stimuli were substantially shorter than those of the normally produced tokens, and thus further shortening was conducted for the nineteen words whose original VOTs exceeded 80 ms. As a result, the average VOT and the standard deviation of the target stimuli were 30.36 ms and 8.95 ms, respectively.

Of eighty target words in the listening list, ten were monosyllabic words which presented potential lexical ambiguity (i.e., confusion with /b/ was lexically possible, as in *pall-ball, peck-beck*). To ensure that these tokens still sounded like the intended target words (i.e, initial phoneme perceived as /p/ as opposed to /b/, despite having VOTs in the region of the presumed category boundary), two native English speakers listened to the target words and recorded what they thought they heard. Every word was perceived as a /p/-initial word by both listeners.

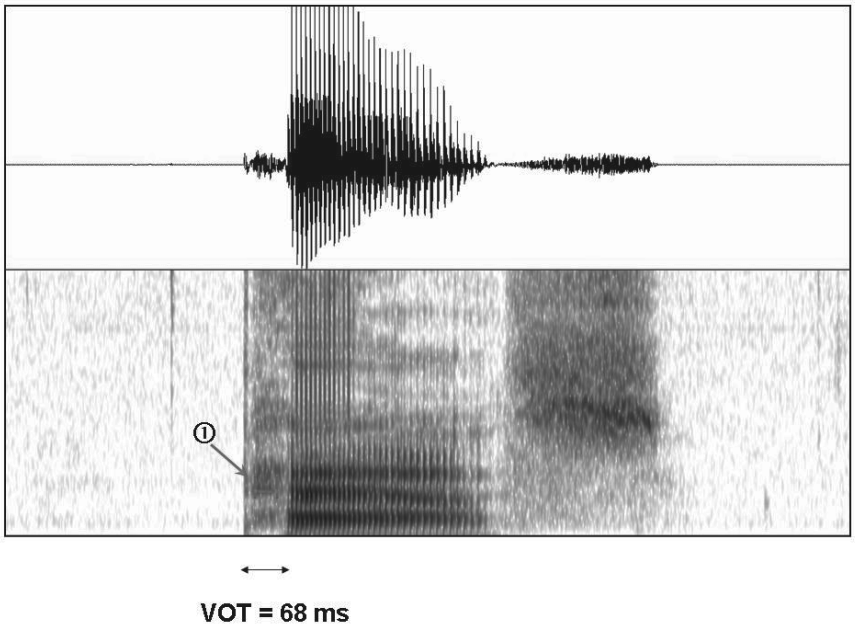

Figure 3-2: Example waveform and spectrogram of word "pass" before VOT manipulation
The original VOT was 68 ms. The most stable portion of aspiration (①) was deleted so that VOT was shortened by 40 ms.

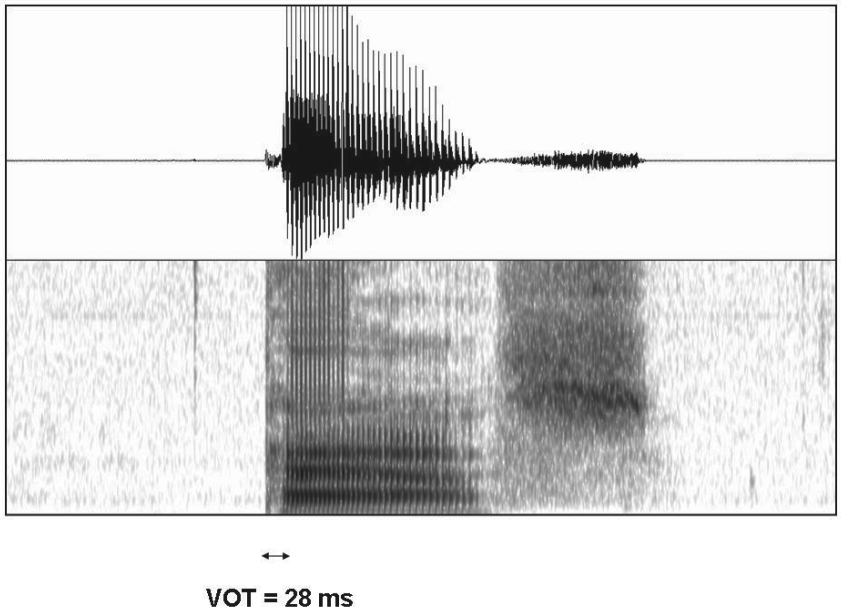

Figure 3-3: Example waveform and spectrogram of word "pass" after shortening VOT by 40 ms., which was used as one of the listening stimuli in Experiment 2
The new token's VOT was 28 ms.

3.1.4 Procedure

The same experimental procedure as in Experiment 1 was used in Experiment 2. Each session was divided into four blocks: 1) warm-up silent reading, 2) baseline recording, 3) target listening, and 4) post-exposure recording. Participants' tokens were digitally recorded onto a computer at a sampling rate of 22100 Hz, and word initial VOT was measured by a phonetically-trained research assistant, using both waveforms and spectrograms in PCQuirer (Scicon R&D, Inc.). No perceptual assessment (i.e., AXB testing) of the baseline versus post-exposure productions was conducted.

3.1.5 Statistical Analysis

The same within- and between-subjects factors as in Experiment 1 were used in Experiment 2: Within-subjects factors were 1) Production Type (Baseline vs. Post-exposure), 2) Lexical Frequency (High vs. Low), 3) Presence of Exposure (Target vs. Novel), and 4) Segment (/p/ vs. /k/), and between-subjects factor was Gender. In addition, in order to compare the results of Experiments 1 and 2, Listening Stimulus [extended VOT vs. reduced VOT] was added as the second between-subjects factor. VOT was the dependent variable. Statistical by-subjects analysis of the data was based on analysis of variance tests (ANOVA) performed with the software package SPSS.

3.2 Results

Table 3 shows a summary of the results from Experiment 2: The mean, median, standard error and standard deviation of measured VOT (ms) values are shown by stimulus type.

Table 3: Summary of Experiment 2 Results. The mean, median, standard error and standard deviation of VOT (ms) are shown by stimulus type

Stimulus Type	Production Type	Mean (ms)	Median (ms)	Std. Error of the Mean (ms)	Std. Deviation (ms)
Target /p/ low	**Baseline**	61.54	63.50	1.83	9.51
	Post-exposure	64.08	63.81	2.28	11.87
Target /p/ High	**Baseline**	63.58	65.91	1.65	8.57
	Post-exposure	64.03	63.31	2.31	12.01
Novel /p/ Low	**Baseline**	62.74	64.51	1.79	9.28
	Post-exposure	62.70	63.23	2.22	11.55
Novel /k/ Low	**Baseline**	77.00	75.26	1.24	6.45
	Post-exposure	76.51	76.62	1.50	7.78

As can be seen, the standard deviations are large in general, due to the individual variability of VOT. On the other hand, the standard errors quite small, showing that the lack of imitation was consistent across subjects.

3.2.1 Phonetic Imitation

A repeated-measures ANOVA analysis of target words, with Production Type (Baseline vs. Post Exposure) and Gender as within- and between-subjects factors, respectively, revealed no significant effect of either factor (Production Type [$F<1$, $p>0.1$]; Gender [$F<1$, $p>0.1$]) nor of the interaction between them [$F<1$, $p>0.1$]. Figure 3-4 summarizes the results in Experiment 2, and plots VOT in milliseconds for three types of stimuli. As seen under *Target /p/* in Figure 3-4, contrary to Experiment 1, the participants' VOT for target words did not decrease after they were exposed to the target speech with shortened VOT. That is, shortened VOT in the target speech was not imitated by the participants. Note that although the shift was not statistically significant, participants' mean VOT for the target words actually *increased* after being exposed to listening stimuli with shortened VOT.

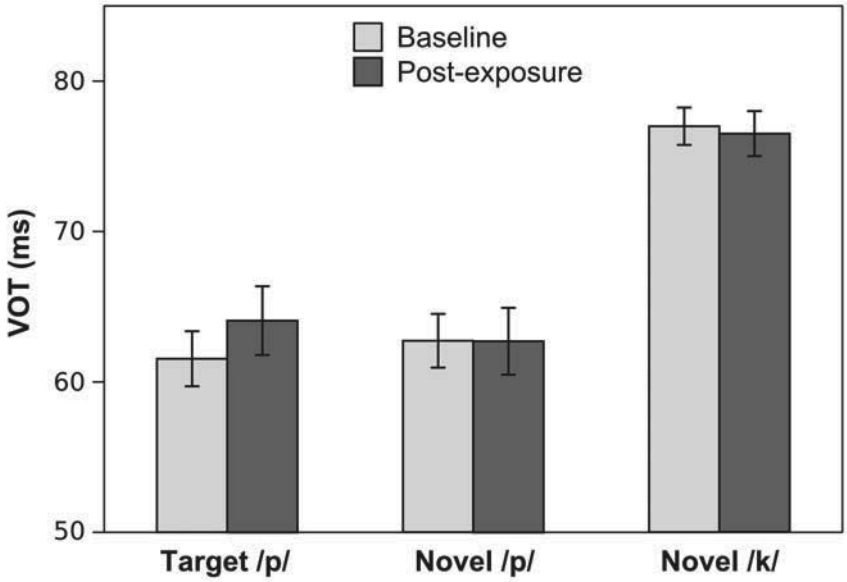

Figure 3-4: Absence of Short VOT imitation
Mean VOT and standard error of the mean in Experiment 2 (Shortened VOT), presented separately for the baseline and post-exposure productions, and for the three types of stimuli. The participants heard only Target /p/. As seen, shortened VOT was not imitated.

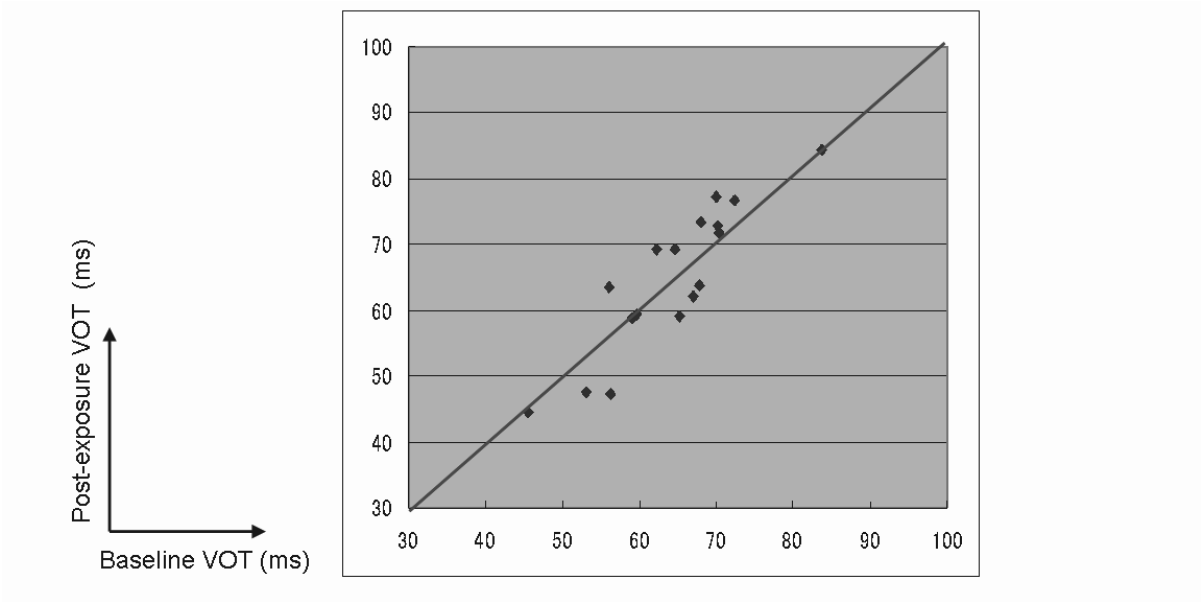

Figure 3-5: Individual variability in VOT and phonetic imitation
Mean VOT of individual participants in Experiment 2 (Shortened VOT). The baseline VOT is plotted on the x-axis, and post-exposure VOT is plotted on the y-axis

(R^2=0.88). The diagonal line's slope is one and indicates no imitation (baseline = post-exposure).

Similar to Experiment 1, the degree of imitation varied considerably across participants. Figure 3-5 illustrates the individual variability of VOT and pattern of imitation in Experiment 2. As seen, the mean VOT ranged from 44 ms to 84 ms, and the magnitude of the changes also varied from participant to participant. Only eight participants decreased their VOT after listening to shortened VOT in the target speech.

3.2.2 Word Specificity

Another repeated-measures ANOVA analysis with two within-subject factors (Production Type and Lexical Frequency) and a between-subjects factor (Gender) was performed to determine the effect of lexical frequency on the degree of imitation. The result revealed that there is no significant effect of Production Type [$F<1$, $p>0.1$] (= no imitation), Lexical Frequency [$F<1$, $p>0.1$], or Gender [$F<1$, $p>0.1$]. The interaction between the two within-subject factors (Production Type and Lexical Frequency) was not significant [$F<1$, $p>0.1$], (= no word-level specificity), while the interaction between Lexical Frequency and Gender was significant [$F(15,1)=6.29$, $p<0.05$*]. The latter effect is attributed to the fact that the male speakers' average VOT was greater for high-frequency words than low-frequency words (collapsed across baseline and post-exposure productions) while female speakers' VOT did not vary across the frequency groups. Figure 3-6 illustrates this gender difference of lexical frequency effect on VOT. Given that this effect was not observed in Experiment 1, and is beyond the scope of the current study, it will not be discussed further.

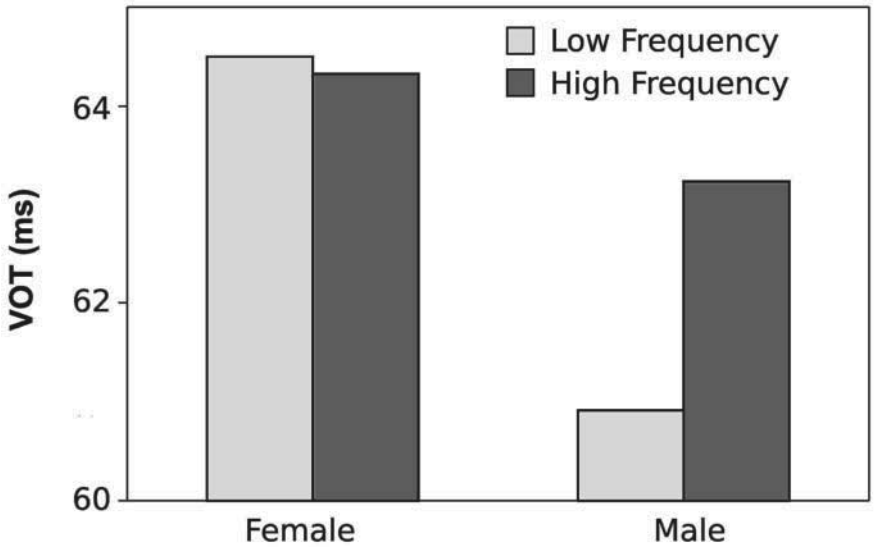

Figure 3-6: Gender difference of lexical frequency effect on VOT
Male participants' VOT was longer for high-frequency words.

Given that word-level specificity was found only in the post-hoc analysis in Experiment 1 (2.2.4), the same post-hoc analysis was performed here, including only those participants whose average VOT in the post-exposure phase *decreased* by more than 5% from their baseline production. Among seventeen participants in Experiment 2, only five participants (2F and 3M) met this criterion. A repeated-measures ANOVA analysis with two within-subjects factors (Production Type and Lexical Frequency) and a between-subjects factor (Gender) showed a significant main effect of Production Type [$F(3,1)=27.72$, $p<0.05$*], while the effect of neither Lexical Frequency nor Gender was significant [$F<1$, $p>0.1$]. The interaction between Production Type and Lexical Frequency was not significant [$F(3,1)=7.88$, $p>0.05$], and neither were the interactions between Gender and the two within-subject factors [$F<1$, $p>0.1$]. That is, even for the few subjects who showed imitation, there was no lexical frequency effect.

Similarly, the post-hoc analysis revealed that Presence of Exposure did not affect the degree of imitation. A significant main effect was observed only for Production Type [F(3,1)=57.21, p<0.01*], but not for Presence of Exposure [F(3,1)=3.52, p>0.1] or Gender [F(3,1)=1.34, p>0.1]. The interaction between Production Type and Presence of Exposure was not significant [F(3,1)=2.46, p>0.1], and neither was the interaction between Gender and Production Type [F(3,1)=1.46, p>0.1] nor between Gender and Presence of Exposure [F(3,1)=1.36, p>0.1]. Taken together, neither lexical frequency nor presence of exposure affected the degree of imitation in Experiment 2, even among the participants who imitated shortened VOT.

3.2.3 Generalizability

In order to test the generalizability of imitation, two more repeated-measures ANOVA analyses were performed (within-subjects factors: Production Type, Presence of Exposure, Segment; between-subjects factor: Gender). The results revealed no significant difference in any comparison [F<1, *p*>0.1] except for the effect of Segment [F(15,1)=70.43, p<0.01*]. As mentioned in Chapter 2 (2.2.3), this difference in VOT between places of articulation is well documented in the literature (e.g., Lisker and Abramson, 1964; Cho and Ladefoged, 1999) and thus was expected in Experiment 2. In short, the participants did not significantly change their speech with respect to their VOT after being exposed to target speech with shortened VOT, such that generalization of phonetic imitation could not be evaluated.

For this reason, the same post-hoc analysis (3.2.2) was performed, including only those participants whose average VOT in the post-exposure phase *decreased* by more than 5% from their baseline production. As described above in 3.2.2, the effect of Production Type

was significant [F(3,1)=57.21, p<0.01*] while there was no interaction between either Production Type and Presence of Exposure [F(3,1)=2.45, *p*>0.1] or Production Type and Segment [F<1, *p*>0.1]. There was no effect of Gender [F<1, *p*>0.1], or interaction regarding Gender [F<1, *p*>0.1]. In sum, for the participants who actually imitated the shortened VOT, the degree of imitation was the same across three types of production stimuli, namely, target words, novel words with initial /p/, and novel words with initial /k/. Figure 3-7 summarizes the results of the post-hoc analysis regarding the generalizability of imitation in Experiment 2: As seen, the degree of imitation did not vary across three types of production stimuli, showing that when shortened VOT was imitated, the change was generalized at both phoneme and feature levels.

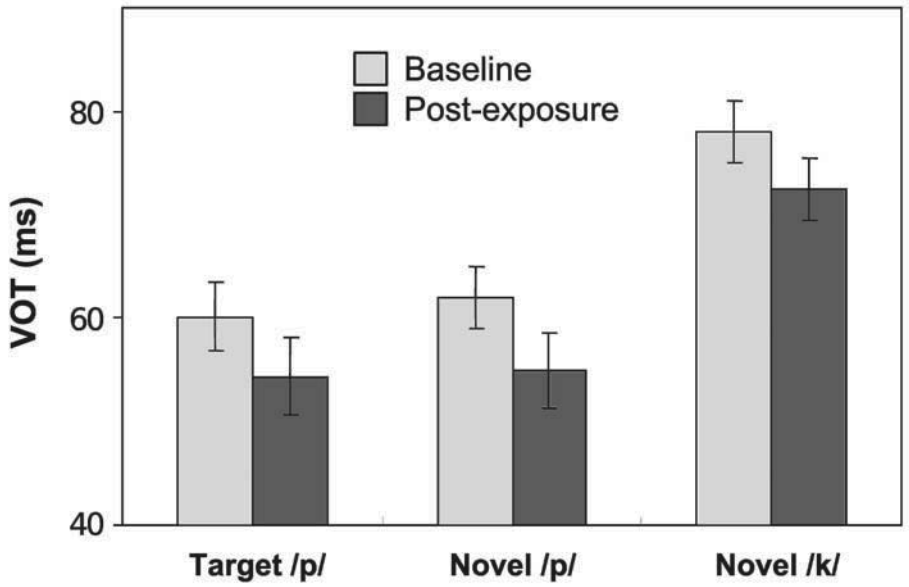

Figure 3-7: Generalization of Short VOT imitation
Figure 13: Mean VOT and standard error of the mean, for participants whose average VOT decreased by more than 5% in Experiment 2 (target = shortened VOT), presented separately for the baseline and post-exposure productions, and for the three types of production stimuli. The participants heard only Target /p/. As seen, the imitation was generalized to novel stimuli.

3.2.4 Long vs. Short VOT (Experiment 1 & 2 Comparison)

A repeated-measures ANOVA analysis with Production Type (within-subjects factor), Listening Stimulus, and Gender (between-subjects factors) revealed a significant interaction between Production Type and Listening Stimulus [$F(1,40)=6.56$, $p<0.05$*], indicating that the degree of imitation differed significantly depending on the listening stimuli. Figure 3-8 illustrates the different patterns of imitation between Experiment 1 and 2: As seen, post-exposure VOT values in the two experiments were clearly different, while the baseline VOT values were equivalent [$F<1$, $p>0.1$]. The main effect of Production Type (= imitation) was significant [$F(1,40)=10.56$, $p<0.01$*], and so was the effect of Gender [$F(1,40)=6.78$, $p<0.05$*], although there was no interaction between the two [$F<1$, $p>0.1$]. As described in section 2.2.2, there was a significant gender difference in VOT values in Experiment 1 [$F(1, 1)=12.60$, $p<0.01$*], while the effect was not observed in Experiment 2 ([$F<1$, $p>0.1$]) (see 3.2.1). This cross-group difference in the Gender effect was shown to be significant [$F(1, 1)=4.41$, $p<0.05$*]. After combining the two experiments, the gender effect on VOT was still significant, indicating that female participants had longer VOT in general for word-initial /p/.

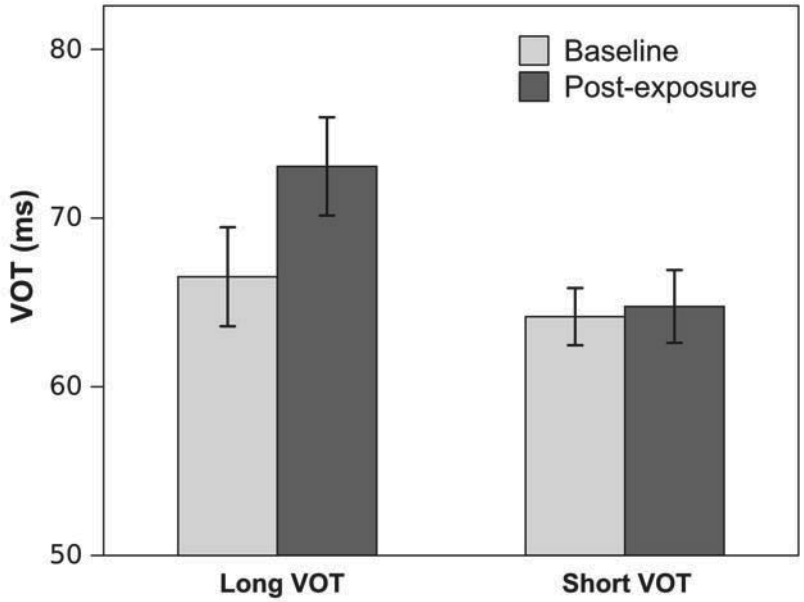

Figure 3-8*: *Comparison of Experiments 1 (Long VOT) and 2 (Short VOT)
Mean VOT and standard error of the mean for words with initial /p/, presented separately for the baseline and post-exposure productions, and for the types of listening stimuli. Long VOT was imitated, whereas short VOT was not.

3.3 Discussion

3.3.1 Asymmetry of Imitation

Contrary to the results in Experiment 1, the participants in Experiment 2 (who listened to shortened VOT) did not show a significant difference between baseline and test productions in terms of their VOT, as seen in Figures 10 and 12. The baseline data from the two experiments showed that the distributions of VOT for /p/ were equivalent: Both were normally distributed, ranging from around 20 ms to 130 ms (the average baseline VOTs in Experiment 1 and 2 were 66.52 ms and 64.16 ms, respectively). The combined VOT distribution of the baseline recordings was as follows: mean=65.1 ms; median=64.0 ms;

standard deviation=19.57 ms; skewness [5] =0.76; standard error of skewness=0.04; kurtosis=4.2. Figure 3-9 shows the histogram of VOT values for word-initial /p/ produced by the forty-four participants in the two experiments. As can be seen, the participants' VOT is normally distributed, centered around 65 ms. Assuming that Figure 14 adequately represents the real-life distribution of /p/ VOT, and that raw percepts are stored as episodic traces, we would expect similar patterns of imitation to occur in both directions. If anything, the shortened VOT values should be more likely to be imitated, as they are more frequent in this figure than the lengthened VOT values are. The clear asymmetry of phonetic imitation in the opposite direction found in this study, namely the absence of reduced VOT imitation, suggests that spontaneous phonetic imitation is *not* an automatic process which reflects the collection of raw percepts, but rather a process which is modulated by other factors. Note, however, that most exemplar theories in fact do not assume raw percepts to be episodic traces, and assert the necessary role of selective attention in shaping echo content; thus this point does not necessarily challenge the exemplar views.

[5] Skewness $\gamma 1$ = [n / (n -1) (n - 2)] sum[(x_i - mean)^3] / std^3. The standard error of skewness = sqrt (6/n).

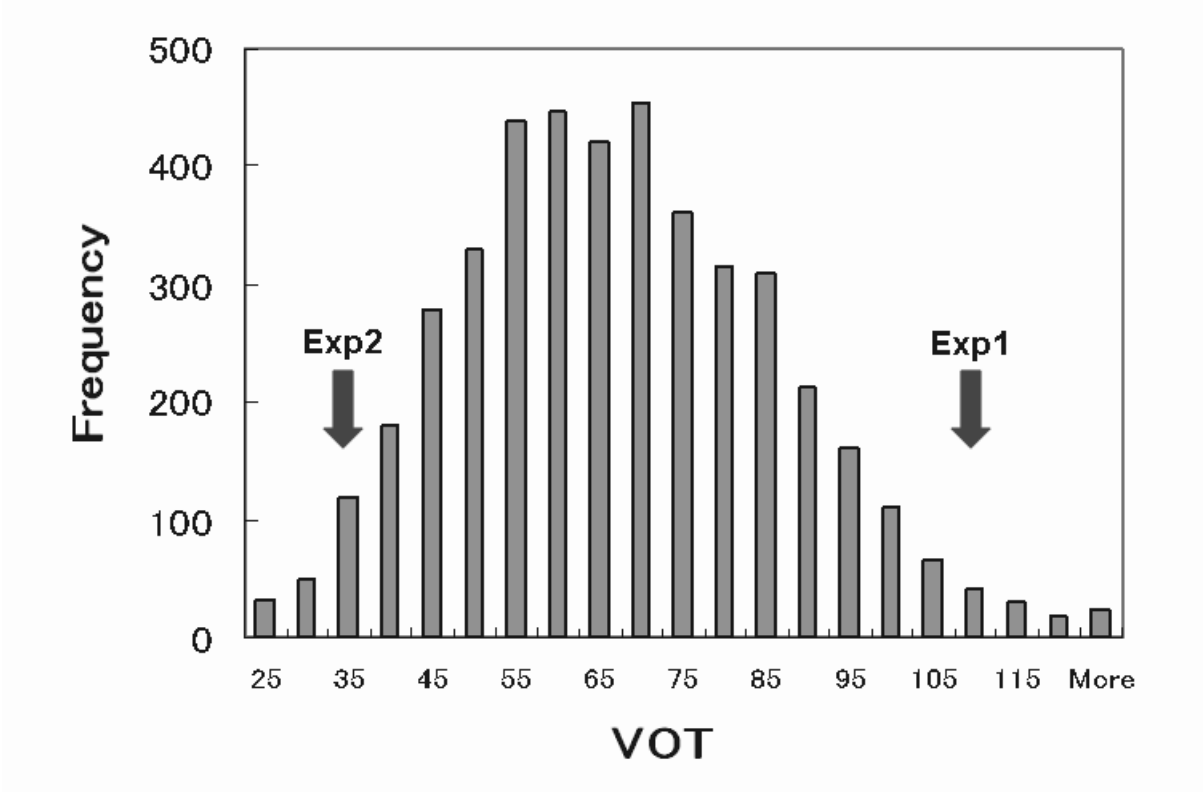

Figure 3-9: Histogram of VOT distribution of word-initial /p/ in the baseline recordings (two experiments combined)
The mean and median were 65.1 ms; 64.0 ms, respectively. Arrows were added to indicate average VOT values of the target speech in Experiments 1 & 2.

Linguistically speaking, the difference between the two conditions is clear: Imitating long VOT does not endanger the voiceless stop category, while imitating short VOT can introduce phonological ambiguity with the voiced stop, which could further cause lexical ambiguity. A similar asymmetrical pattern can be found in the VOT goodness-rating study by Allen and Miller (2001), which showed that different types of contextual factors influence internal category structure differently, such that a change in speaking rate affects both the upper and lower edge of the best-exemplar range along a VOT continuum, while a change in lexical status affects the location of the best-exemplar range only at the category boundary

location (*cf.* Ganong, 1980). The point of relevance here is that regardless of which factor is manipulated, the goodness rating curves are positively skewed (i.e., not symmetrical, but have an elongated tail towards the right), with the skewness being more prominent for slow speaking rate. That is, the rating goes down steeply as the stimuli approach the phonemic boundary (i.e., /p/ vs. /b/), while the rating goes down gradually as the stimuli approach the opposite end of the category. Note that it is not possible to infer goodness ratings of the target stimuli used in the current study from Allen and Miller (2001), given that VOT is sensitive to multiple factors, such as phonetic context and synthesis parameters. However, given that the target speech in Experiment 2 was slower than the *slow* series stimuli in Allen and Miller (2001)[6], we can at least assume that the target stimuli in Experiment 1 (with lengthened VOT) would have been rated higher than those in Experiment 2 (with reduced VOT). If the participants in Experiment 2 perceived the target stimuli as bad exemplars, it is conceivable that they were not committed to long-term memory, and thus failed to influence the production. As the asymmetrical distribution of the phonemic goodness curves suggests, phonemic goodness appears to be affected by the presence of phonemic contrast, rather than simply reflecting the frequency distribution in the category. Further, this perceived goodness might have contributed to the asymmetry of phonetic imitation.

The notions of phonemic goodness and its effect on phonetic imitation are both compatible with exemplar-based models, as long as episodes do not store "raw data", but reflect attended cognitive percepts (*cf.* Pierrehumbert, 2002, Goldinger, 2007). Bad or ambiguous exemplars receive small (or possibly no) attentional weight, subsequently leaving no traces in the exemplar space of the category. Even when the traces are present, the entire

[6] The average word duration of the speech stimuli in Allen and Miller (2001)'s *slow* series was 463 ms, as compared with 512 ms in Experiment 2 in the current study (/p/ initial monosyllabic words).

exemplar cloud of a given unit (e.g., word or phoneme) does not have to be the goal of production. For example, in order to produce a word with initial /p/, only the exemplars in *some area* of the /p/ region may be activated, serving as a production goal. In sum, perceived goodness of the target speech may account for the asymmetrical pattern of imitation between Experiments 1 and 2.

Alternatively, the observed asymmetrical pattern of imitation may be explained in terms of the depth of processing required by different experimental tasks. In the speeded categorization task in Miller (2001), response time was relatively long for poor voiceless exemplars with short VOT values (i.e., near the voiced-voiceless category boundary), while it was shorter for poor voiceless exemplars with long VOT values, indicating that categorization time is a function of the distance between the exemplar and the category boundary, not of the perceived category goodness. In fact, Goldinger (2000) reported that degree of imitation was negatively correlated with recognition time ($r = -.41$, $p < .01$). If the stimuli in Experiment 2 required longer processing time than those in Experiment 1, that would allow more time for the underlying phonemic representation to develop, and possibly override the surface information (i.e., reduced VOT), resulting in no imitation effect. McLennan et al. (2003) provide a similar account for the presence of allophonic and phonemic specificity effects. They conducted a series of long-term repetition-priming experiments, and found both allophonic and phonemic specificity in different tasks: When the task was an easy lexical decision resulting in shorter reaction times, only allophonic level priming effects were observed (i.e., [æɾəm] primed [æɾəm], but not [ædəm] nor [ætəm]). When the task was made more difficult resulting in longer reaction times, phonemic level

priming effects were found (i.e., [ærəm] primed [ærəm], [ædəm], and [ætəm]). They argue that the depth of processing (manipulated through ease of lexical discrimination) mediates the activation of underlying representations.

3.3.2 Gesture Theories of Speech Perception

Shockley et al. (2004) and Fowler et al. (2003) argued that gesture theories of speech perception (i.e., the Motor Theory of Liberman and Mattingly, 1985; the direct realist theory of Fowler, 1994) can account for the phonetic imitation of long VOT that they found in single-word shadowing. Unlike traditional acoustic theories, these theories assume listeners' percepts of speech signals to include motor information. Therefore, when participants are asked to shadow a speech signal, their responses are automatically guided by their perception of modeled gestures, thus resulting in spontaneous imitation.

Although it is not clear how long the memory of perceived gestures is sustained, our findings of phonetic imitation and its generalization at a feature level seem prima facie to be more compatible with both gesture theories than with traditional acoustic theories. However, the two gesture theories contrast in terms of their ability to account for the results in Experiment 2, namely the lack of phonetic imitation. The direct realist theory asserts that the objects of perception are distal (such as actual vocal tract movements), predicting phonetic imitation to be automatic. On the other hand, the motor theory asserts that the objects of perception are proximal (such as the perceiver's own knowledge of intended gestures) and speech signal and percepts are independent, and thus phonetic imitation should be non-automatic. Therefore, only the motor theory can account for our finding of non-automaticity of phonemic imitation without additional processes. Fowler et al. (2003)

attributed the discrepancy between modeled and shadowed VOT values to an influence of habitual action on speech production, as opposed to processes that intervene between perception and production. However, habitual action cannot account for the complete lack of imitation found in Experiment 2 in the current study, because the participants displayed a wide enough range of baseline VOTs to allow them to produce reduced VOT (as seen in Figure 14). Although the phenomenon of phonetic imitation suggests a close tie between speech perception and production, it is clearly not an automatic process, and other factors (such as linguistic contrast, sociostylistic register, and attentional weight) appear to play an important role.

3.4 Interim Summary

The results from Experiment 2 demonstrated that spontaneous phonetic imitation is not automatic: Contrary to Experiment 1, participants did not imitate *shortened* VOT in the target speech, suggesting that phonetic imitation was attenuated by the presence of a phonological boundary. A post hoc analysis revealed that for the subset of speakers who actually imitated shortened VOT, the change was generalized to new instances of /p/ as well as a new phoneme /k/, replicating the result of feature-level generalization observed in Experiment 1. On the other hand, the effect of word-specificity was not found in the post-hoc analysis. These results indicate that the effect of feature-level generalization is robust in phonetic imitation, while the effect of word specificity is only weakly manifested.

Chapter 4: Experiment 3 - Imitation of High Vowel Devoicing in Japanese

Experiment 1 showed that artificially lengthened VOT on the modeled phoneme /p/, which is not phonemically constrained, was imitated at two levels of representation: word and feature. Experiment 2 showed that artificially *shortened* VOT, which could endanger phonemic contrast with phoneme /b/, was not imitated, suggesting that the need to preserve phonemic contrast suppresses the effect of episodic traces.

The current chapter describes Experiment 3, which examined the extent to which phonological factors unrelated to contrast preservation also affect imitation of fine phonetic details. As mentioned in Chapter 3, contrast preservation is a fundamental and essential component of phonological grammar (e.g., Flemming, 2001; Lubowicz, 2003), and thus its violation should be highly marked. However, there are many phonological factors that are unrelated to contrast preservation and are weakly realized in grammar. If such factors constrain imitation of phonetic details, it will suggest that phonetic imitation occurs *only* in the absence of phonological factors. In this case, the episodicity previously observed in phonetic imitation might not have any implications/consequences in terms of *phonological representations.* On the other hand, if imitation is observed despite the presence of phonological factors, it will provide a stronger argument for the episodic account of phonological representations. To address this issue, Experiment 3 examined the phonetic imitation of Japanese vowel devoicing.

4.1 Background: Japanese Vowel Devoicing

In Tokyo Japanese, short high vowels [i, ɯ] tend to be devoiced or deleted when they occur between two voiceless consonants (Han, 1962; McCawley, 1968; Vance, 1987; Beckman and Shoji, 1984), e.g., [kɨ̥ʃa] *kisha* 'journalist', [koksai] *kokusai* 'international'.

Japanese vowel devoicing was traditionally considered a phonological assimilation of the feature [–voice]. For example, in McCawley (1968), the process is represented as the following: V [+high] → [–voice] / [–voice] __ [–voice]. Recently, alternative analyses were proposed by Tsuchida (1997) and Varden (1998) employing Optimality Theory (Prince and Smolensky, 1993) and Feature Geometry, respectively, in which Japanese voiceless vowels are specified as [+spread glottis] instead of [–voice], the [+spread glottis] having spread from the voiceless consonant context. Despite the difference in featural representations, the nature of Japanese vowel devoicing is considered to be a phonological feature assimilation or feature spreading in both analyses, and devoiced vowels are expected to be voiceless.

In contrast, Jun and Beckman (1993) proposed an alternative account of Japanese vowel devoicing as a gradient phonetic process involving overlap of glottal gestures. According to this view, Japanese (and also Korean) vowel devoicing is the result of extreme overlap and hiding of the vowel's (voiced) glottal gesture by the adjacent consonant's (voiceless) glottal gesture. This account predicts various phonetic factors to affect the rate (occurrence/application) of devoicing, such as segmental and prosodic context and speech tempo. In this view, complete vowel deletion is an endpoint of gradient gestural overlap.

As discussed in Tsuchida (1997) and Varden (1998), the precise phonetic implementation of Japanese vowel devoicing is not homogeneous. It ranges from shortened

(voiced) vowel, to devoiced/voiceless vowel, to pure deletion. In Beckman (1982), vowels in devoicing environments are normally deleted or shortened as opposed to 'devoiced', and similar results are also reported in Keating and Huffman (1984). Given that Japanese vowel devoicing appears not to typically involve devoiced/voiceless vowels, and that its realization ranges from (shortened) voiced vowel to vowel deletion, one possible way to quantify the degree of devoicing is to measure vowel duration.

Previous studies have shown that devoicing can be optional, with a number of both phonetic and phonological factors affecting the rate of devoicing. It is well known that devoicing is more common in fast speech than slow speech (Kuriyagawa and Sawashima, 1989). Consonantal context, especially manner of articulation**,** is also known to affect the devoicing rate (e.g., Maekawa and Kikuchi, 2005; Fujimoto, 2004). Vowel quality is also claimed to condition devoicing. For example, Han (1962) reported that [u] is more likely to devoice than [i]. Other phonological factors known to influence the likelihood of devoicing include syllable structure and position (Kondo, 1997, 2005), morphological/word boundary (Vance, 1987), moraic position (Kuriyagawa and Sawashima 1989), and location of pitch accent (Kitahara, 1999, 2001). In addition, Japanese vowel devoicing is particularly prominent in Tokyo dialect (= Standard Japanese), and thus there is a sociolinguistic register associated with the phenomenon (Imaizumi et al., 1999; Beckman, 1994; Morris, 2005).

Japanese vowel devoicing is suitable for our purpose because: 1) unlike shortened VOT, its imitation does not endanger phonemic contrast (Beckman and Shoji, 1984), 2) unlike lengthened VOT, its imitation in certain environments would violate well-established phonetic and phonological constraints just described, 3) it has been claimed to involve the feature [+spread glottis] (Tsuchida, 1997) which we have shown to be imitable in lengthened

VOT, 4) similar to VOT, Japanese devoicing is a gradient process and is quantifiable (Jun and Beckman, 1993), 5) high within- and cross-speaker variability has also been reported (e.g., Varden, 1998) which makes imitation more probable. Further, it does not have an overt association with a stylistic register such as the association of long VOT with "careful" speech: Imaizumi et al. (1995) showed that professional teachers of hearing-impaired children *reduced* their vowel devoicing in order to improve their listeners' comprehension, while Takeda and Kuwabara (1987) showed that it is not a fast or casual speech phenomenon, as it is found in most careful read speech (see also Kondo, 1997).

In addition to the pattern of phonetic imitation, the current experiment also aimed to examine the relationship between Japanese vowel devoicing and lexical frequency in light of the Probabilistic Reduction Hypothesis (Jurafsky et al., 2001), which predicts that more probable words are (phonetically) reduced more. Although Japanese vowel devoicing has been studied extensively from both phonetic and phonological viewpoints, the effect of lexical frequency is yet unknown. Vowel devoicing processes occur in many languages, and are generally considered to be part of the vowel reduction processes in which vowels are first reduced in duration and centralized in quality, then eventually devoiced and/or deleted (e.g., Dauer 1980, Kohler 1990, among others). However, as described above, earlier studies suggest that Japanese devoicing is not simply a vowel reduction process, but rather a process controlled by many linguistic factors. If it is indeed a reduction process, however, we would expect to see a higher instance of devoicing among high frequency words.

The current experiment also aims to examine gender differences in Japanese vowel devoicing. Previous studies have shown that there can be a gender differences in voice quality (e.g., Klatt and Klatt, 1990, in English; Gunzburger, 1991, in Dutch): Female

speakers were shown to have a breathier voice than male speakers. Given that both breathy voice and vowel devoicing are produced with a more open glottis, if Japanese female speakers have breathier voice, one could also expect a gender asymmetry in vowel devoicing.

Lastly, it is also our interest to examine the mechanism of Japanese devoicing, comparing the traditional feature assimilation account (e.g., McCawley, 1968) and the gradient phonetic account proposed by Jun and Beckman (1993).

4.2 Method

4.2.1 Participants

Twenty four native monolingual speakers of Tokyo Japanese (age 18–50, mean 28.25, 12 females) with normal hearing served as subjects for this experiment. None of the participants had lived outside of the greater Tokyo area (Tokyo, Chiba, Kanagawa, and Saitama), and none had daily exposure to foreign languages. They were recruited from the TUFS (the Tokyo University of Foreign Studies) population, and were paid for their participation.

4.2.2 Stimuli Selection

The production list consisted of 110 Japanese words: 1) 40 words containing devoiceable /u/ (20 target words which were played in the listening phase, and 20 test words which were not played during the listening phase), 2) 10 target words containing devoiceable mid vowels (five /e/ and five /o/), 3) 20 test words containing both devoiceable /u/ and /i/, 4) 20 test words containing devoiceable /i/, and 5) 20 filler words containing no devoicing environment. The listening list consisted of 50 Japanese words, including only the target and filler words (1a, 2, and 5 above). Table 4 illustrates the type and distribution of stimulus words used in Experiment 3: The production list contained all words in Table 4, while the listening list

contained words in the shaded cells. As seen, lexical frequency was varied within each stimulus group, except for the fillers which all had high frequency. Lexical frequency and familiarity were determined from the NTT database (Lexical properties of Japanese, Amano and Kondo, 2003). The thresholds for low- and high-frequency words were <500 and >1000, respectively (Low: average 239, standard error 20.3; High: average 17045, standard error 3563.7). All words had high familiarity (> 5.0 on the 7-point scale). Lexical neighborhood density, however, was not controlled in this experiment. For a complete list of words used in Experiment 3, see Appendix B.

Table 4: Type and distribution of Experiment 3 stimuli

Type	*Vowel*	*High Frequency*	*Low Frequency*	*Total Number*
(1a) Target u	*u*	**10**	**10**	**20**
(1b) Novel u	*u*	10	10	20
(2) Target mid	*e* or *o*	**4**	**6**	**10**
(3) Novel u&i	*u* and *i*	10	10	20
(4) Novel i	*i*	10	10	20
(5) Filler	*N/A*	**20**	0	**20**

Participants only listened to words in the listening list (i.e., Target and Filler, shown in shaded cells), while they produced all words in the baseline recording and post-exposure recording blocks.

4.2.3 Stimulus Construction

A phonetically trained female speaker of Tokyo Japanese (not the author) provided the recording of the 50 words (30 target and 20 filler) in the listening list. The speaker was first instructed to read the words aloud naturally at a comfortable speaking tempo. Her speech was recorded and was examined by the author. As expected, she produced very little voicing in most devoicing environments, with some regular exceptions such as word-final positions and

mid-vowels where she did produce voicing. Later, the speaker was asked to read the same list of words once again, and was explicitly instructed to devoice in certain phonological environments where she had failed to fully devoice in the earlier recording. Most vowels in devoicing environments were indeed fully deleted or devoiced in the second recording, including those environments that are presumably phonotactically highly marked, such as consecutive or word-final devoicing involving mid-vowels. The speaker deleted most devoiceable /u/s and rarely produced voiceless /u/: Of 20 target /u/ words (which contained 28 /u/s), the speaker produced only one [u̥] (in '*zokushutsu*' [zoku̥shtsu]). On the other hand, mid vowels (in 5 target /e/ words and 5 target /o/ words) were never deleted in devoicing environments and were indeed devoiced. Figure 4.1 shows the example spectrogram of the word '*taishokuteate*' (retirement bonus). As seen, the speaker produced voiceless mid vowels in both consecutive and word-final positions, while a high vowel /u/ following /k/ was deleted.

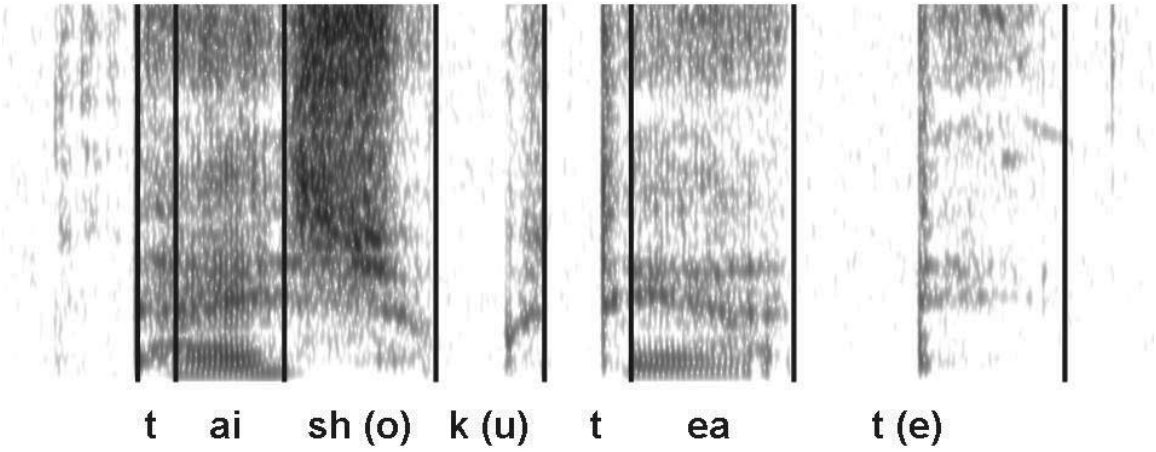

Figure 4-1: A spectrogram showing the word "taishokuteate" (retirement bonus)
In this example, the speaker produced complete devoicing in all the devoicing environments.

To make the listening stimuli, the remaining voiced portions in any devoicing environments were spliced out using Praat (Boersma and Weenink, 2007) so that the resulting tokens had no partial devoicing. Voiceless vowels (e.g., /o/ in Figure 4-1) were left in, although they were extremely rare in the speaker's production (i.e., except for the mid vowels, vowel devoicing was mostly realized as a deletion). Figure 4-2 illustrates examples of the word "tekagami" (hand mirror) *before* and *after* acoustic manipulation. As seen in Figure 4-2(b), the voiced portion of /e/ in the first mora (1 in Figure 4-2(a)) was deleted.

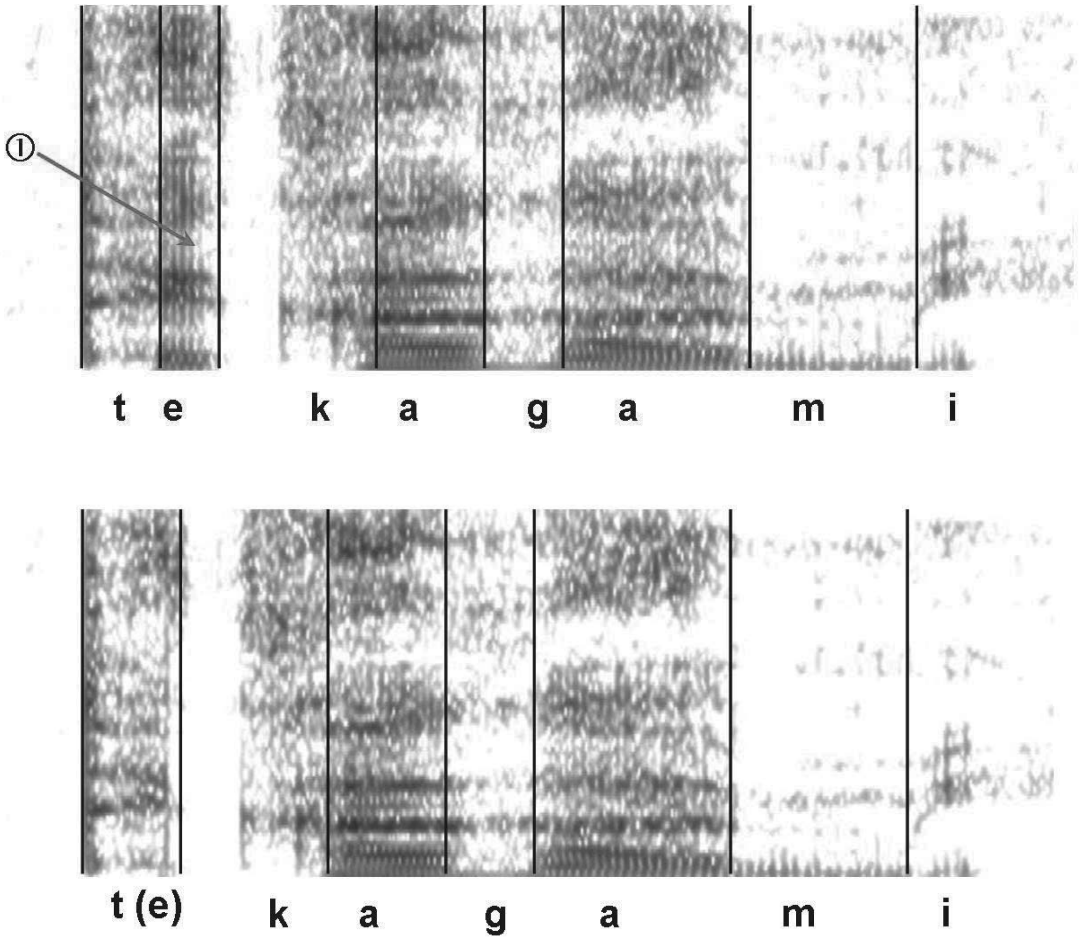

Figure 4-2: Spectrograms of the word "tekagami" (hand mirror) (a) before and (b) after the acoustic manipulation

There were two exceptional cases, where all vowels were devoiceable and thus it was not possible to eliminate all the voiced portions (i.e., '*fukutsu', 'kosokoso'*). For both cases, the speaker's natural productions were used: In '*fukutsu*', she only produced voicing in the word-final /u/, and in '*kosokoso*', she produced voicing in the first /o/. The durations of those two vowels were 112 ms and 52 ms, respectively. As a result, the devoicing rate of the target speech was 100%, with no voiced portion of any devoiceable vowel, excluding low vowel /a/ and the two exceptional cases mentioned above. The mean durations of voiceless

consonants preceding complete devoicing and (devoiceable) voiced vowels were 138 ms and 114 ms, respectively. The average whole-word duration was 509 ms.

4.2.4 Procedure

Experiment 3 was conducted in the Research Institute for Languages and Cultures of Asia and Africa (ILCAA) at the Tokyo University of Foreign Studies in Japan. The procedure, much as in Experiments 1 and 2, employed the word-naming imitation paradigm as in Goldinger (2000), in which each session consisted of the following three blocks: 1) baseline recording, 2) listening, and 3) post-exposure recording. Unlike in Experiments 1 and 2, no warm-up block was included prior to the baseline recording. This modification was done in order to avoid potential drawbacks associated with using the warm-up phase, in response to the relatively weak frequency effect observed in Experiment 1. Although the warm-up block was originally added to avoid hyper-articulation (in VOT) observed in the pilot study, previous studies have reported an interaction of word repetition and lexical frequency (e.g., Norris, 1984; Scarborough et al, 1977), potentially confounding the frequency effect. By eliminating the warm-up block, it was hoped that Experiment 3 would be better able to detect any effect of lexical frequency than Experiments 1 and 2 did.

After a brief explanation of the procedure by the author, each participant was seated in front of a computer in a sound booth. The experimental stimuli were presented using Matlab. The visual instruction (in Japanese) read as follows: “Please read aloud the words presented on the computer screen. Please pronounce them as naturally as possible.” In the baseline recording block, the words in the production list were visually presented, one at a time, on a computer screen every 2.5 seconds. The participants’ production was recorded

through a Logitech A-0356A headset. In the listening block, the participants were asked to listen carefully to two repetitions of the words in the listening list (which was a subset of the production list), presented every 3.3 seconds through the headset. There was no additional task during this block. Finally, in the test recording block, the participants read the words (in the production list) for the second time, providing a post-exposure recording. Each session typically lasted fifteen minutes. Across the three blocks, the words were presented in random order for each subject. Each participant's speech was digitally recorded into a computer at a sampling rate of 22100 Hz for later acoustic analysis. No perceptual assessment of the baseline versus post-exposure productions was conducted.

4.2.5 Acoustic Analysis

Acoustic measurements of the data were made in Praat (Boersma and Weenink, 2007). Three types of duration measurements were taken: voiced vowel (target words), consonant (target words), and whole word (target and filler words). Voiceless vowels were treated as a part of the preceding consonant (e.g., '*sh(o)*' in Figure 4-1). However, consistent with the modeled speaker's speech, voiceless high vowels were extremely rare in the data and most cases of complete devoicing were indeed realized as complete deletion. In addition, intensity measurements of devoiceable (yet voiced) vowels were taken. In order to automate the measurements, a TextGrid with four tiers (corresponding to the four measurement objects above) was created for each recording, and speech signals were segmented by the author or a research assistant using both waveforms and spectrograms. Figure 4-3 illustrates the acoustic measurement of the word '*shishimai*' (lion dance): As seen, the boundaries for consonants and vowels were marked (by the vertical lines with no break) in the tiers C and V,

respectively, and each segment was labeled. The middle point of each devoiceable vowel was marked in the Int tier, and word boundaries were marked and labeled in the W tier. The duration of each labeled interval (in the tiers C, V and W) was calculated using a Praat script, as was the intensity of each marking point in the third tier.

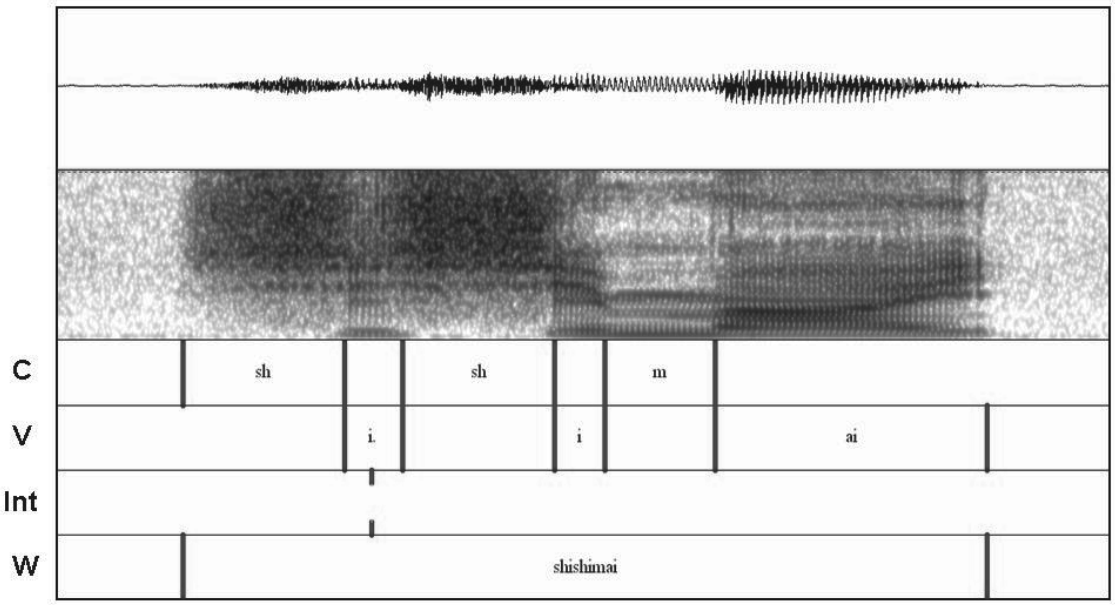

Figure 4-3: Example of acoustic measurement
An example of acoustic measurement for the word '*shishimai*' (lion dance). Participants' speech was first segmented using both spectrogram and waveform, and duration of each segment and whole word, as well as the intensity of each devoiceable vowel, were measured.

Note that in some cases, it was difficult to determine a segment boundary, particularly in the case of complete vowel deletion (= consecutive fricatives) as seen in Figure 4-4. For these cases, the intensity reading in Praat (the thin line in the spectrogram in Figure 4-4) was used.

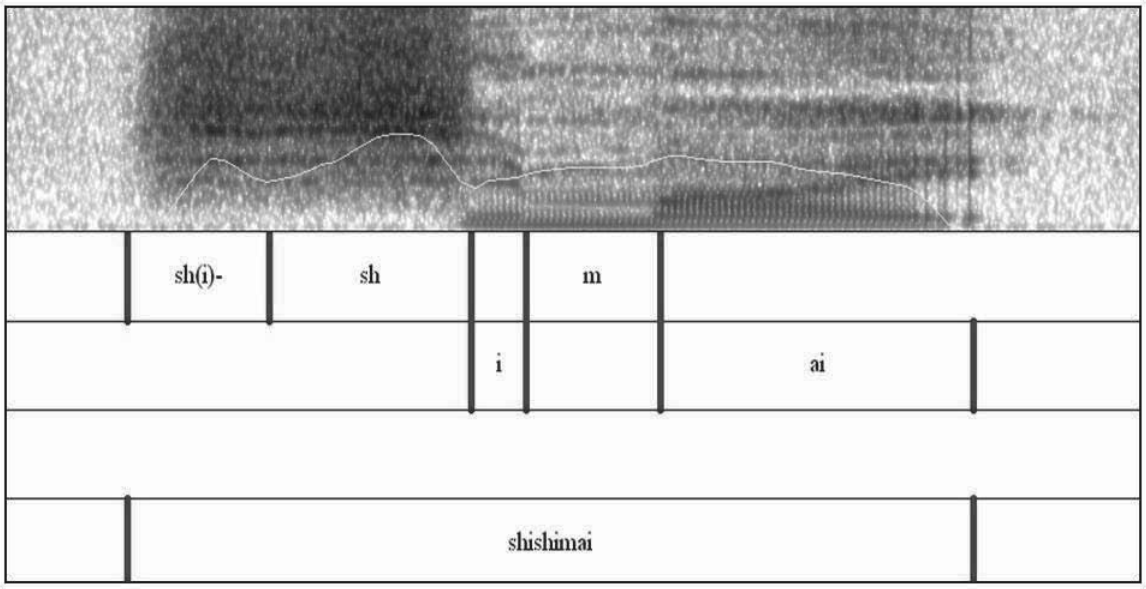

Figure 4-4: Example of acoustic measurement using intensity reading
Intensity (thin line in the spectrogram) was used when it was difficult to determine the segmental boundary.

In order to examine the results phonetically as well as phonologically, the degree of devoicing was quantified in two ways: (1) A *gradient* measure was obtained by measuring the duration of devoiceable (yet voiced) vowels (e.g., [i.] in the V tier of Figure 4.3), and (2) a *binary* measure was obtained by calculating the fraction of devoiceable vowels that were completely deleted (e.g., [sh(i)-] in Figure 4-4) or devoiced.

4.2.6 Statistical Analysis

Statistical by-subjects analysis of the data was based on repeated-measures analysis of variance tests (ANOVA) performed with the software package SPSS. As described in 4.2.5, the degree of devoicing was analyzed according to both gradient and categorical measures in

order to examine the results from phonetic and phonological perspectives. Unless specified otherwise, the dependent variable in the following gradient analyses was (voiced) vowel duration, and the dependent variable in the following categorical analysis was rate of complete vowel devoicing and deletion, which is hereafter called rate of complete devoicing.

Among the independent variables in Experiment 3, the within-subject factors were 1) Type of Production (Baseline vs. Post-Exposure), 2) Vowel (/u/, /i/, /e, o/), 3) Segment (vowel vs. consonant), 4) Lexical Frequency (High vs. Low), and 5) Presence of Exposure (Target vs. Novel), and the between-subjects factor was Gender.

4.3 Results

4.3.1 Phonetic Imitation

The effect of phonetic imitation was examined using repeated-measures ANOVA analyses with the following independent variables: Production Type, Vowel (/u/, /i/, and /e,o/), and Gender. As described in 4.2.3, the target speech did not contain any voiced vowels in devoicing environments (except for two cases described earlier), and thus the targets of imitation with regard to the duration of devoiceable vowels (gradient) and the rate of complete devoicing (categorical) were 0ms and 100%, respectively.

4.3.1.1 Gradient Analysis

The extreme degree of devoicing in the target speech was imitated by the participants: A significant main effect of Production Type [$F(1,22)=45.72$, $p<0.01$*] (= imitation) was found in the gradient analysis. As expected, the mean duration of devoiceable vowels in post-exposure productions was significantly shorter (by 9 ms) than in the baseline productions. Figure 4-5 illustrates the effect of phonetic imitation on average vowel duration.

As seen, participants' vowel duration decreased in the post-exposure productions. The effect of Vowel was also significant [F(1,22)=3.76, $p<0.05$*], while the interaction between the two variables did not reach significance [F(1,22)=2.95, $p>0.05$]. The effect of Gender was not significant, and did not interact with either within-subject factors [F<1, $p>0.1$].

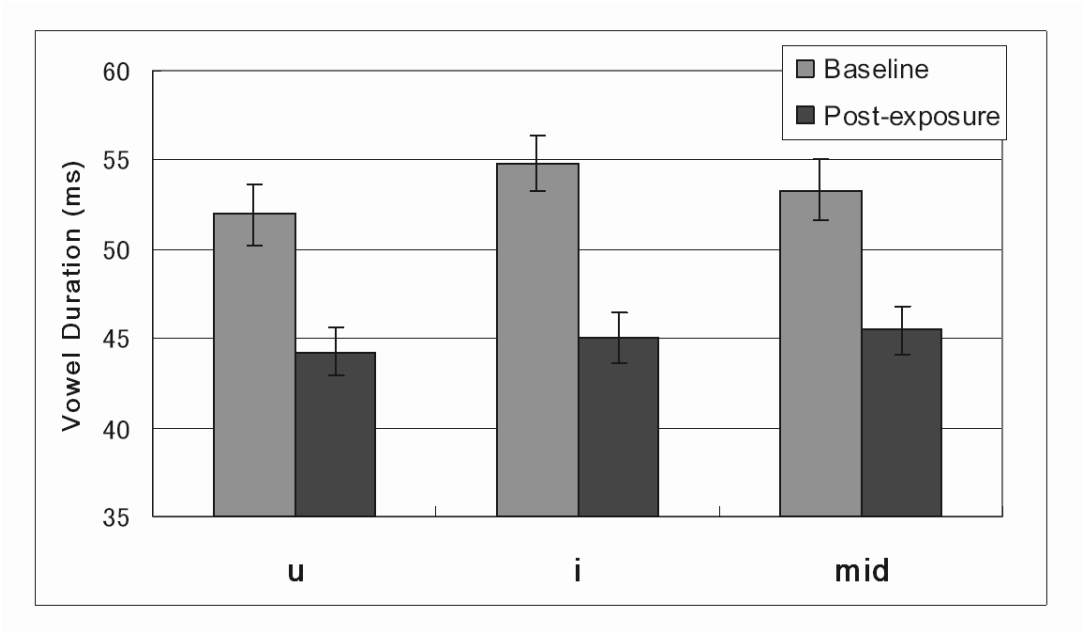

Figure 4-5: Phonetic imitation of Japanese vowel devoicing in vowel duration
Mean duration of devoiceable vowels (ms) and standard error of the mean, presented separately for the baseline and post-exposure productions, and for the three types of vowels tested. After listening to the target speech, participants' vowel duration (in devoicing environments) decreased.

A significant effect of Production Type was also found on both whole word duration (target words only) [F(1,22)=10.55, $p<0.01$*], and intensity [F(1,22)=17.33, $p<0.01$*], where word duration became shorter, and intensity became lower post-exposure. The effect of Gender was not significant in either analysis (word-duration, intensity), and did not interact with either factor [F<1, $p>0.1$]. Given the whole-word duration result, in order to determine

whether the shorter vowel duration in the post-exposure was actually due to phonetic imitation, rather than to overall shortening, another repeated-measures ANOVA analysis was performed with Production Type and Segment (vowel vs. consonant) as independent variables. In addition to significant main effects of the two variables (Production Type [$F(1,22)=10.01$, $p<0.01$*], Segment [$F(1,22)=887.5$, $p<0.01$*]), the result revealed a significant interaction between the two [$F(1,22)=18.41$, $p<0.01$*]: While the target vowels became shorter, the target consonants became longer after being exposed to the target speech. This result provides an argument against the interpretation that the imitation is simply due to global changes in the participants' speech style or speech rate.

4.3.1.2 Categorical Analysis

Consistent with the results from the gradient analysis, the categorical analysis (independent variables: Production Type, Vowel, and Gender) showed that target speech was imitated. The result of a repeated-measures ANOVA analysis revealed a significant effect of Production Type [$F(1,22)=23.98$, $p<0.01$*] and Vowel [$F(1,22)=51.83$, $p<0.01$*]. The interaction between the two variables was significant [$F(1,22)=5.67$, $p<0.05$*], due to the low degree of imitation among mid-vowels. As seen in Figure 4-6, the rate of complete devoicing increased in post-exposure production by 5%. The low rate of mid-vowel devoicing was expected. The clear difference between /u/ and /i/ in the baseline is consistent with the previous literature (e.g., Han, 1962). The current experiment was not designed to compare the baseline difference between the high vowels, and this point is not discussed further.

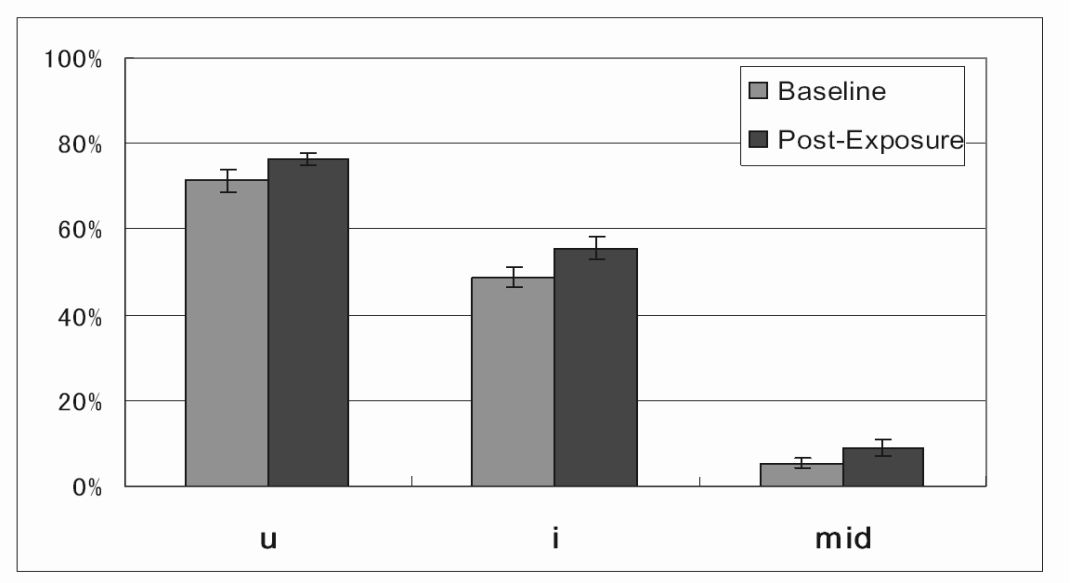

Figure 4-6: Phonetic imitation of Japanese vowel devoicing measured in rate of complete devoicing
Mean occurrence of complete devoicing and standard error of the mean, presented separately for the baseline and post-exposure productions, and for the three types of vowels tested. After listening to the target speech, participants produced more cases of complete devoicing.

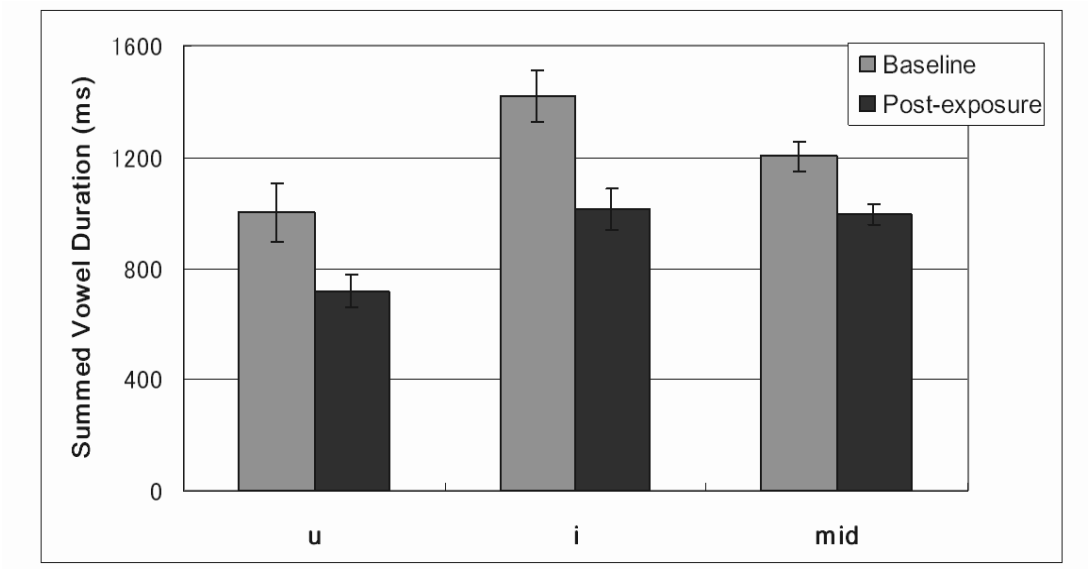

Figure 4-7: Phonetic imitation of Japanese vowel devoicing in summed vowel duration
Mean of summed vowel duration and standard error of the mean, presented separately for the baseline and post-exposure productions, and for the three types of vowels tested.

To see a more holistic effect of devoicing imitation, another repeated-measures ANOVA analysis was performed with mean duration of summed voiced portions of vowels in devoicing environments as the dependent variable, and Production Type and Vowel as independent variables. The vowel duration measure here was the sum of the durations of all instances of a given vowel (i. u, mid vowels) for each speaker, including 0 as the duration value for a fully-deleted vowel. This analysis revealed a significant effect of Production Type [$F(1,22)=63.29$, $p<0.01$*] and Vowel [$F(1,22)=46.20$, $p<0.01$*], and the interaction between the two variables was significant [$F(1,22)=10.09$, $p<0.01$*]. Figure 4-7 illustrates the effect of phonetic imitation on summed vowel duration. As seen, summed duration decreased in all the vowel types tested, although the mid vowels showed the smallest change, due to its low rate of complete devoicing. Similarly, the low summed duration of /u/ reflects its high rate of complete devoicing (see Figure 4-6), not necessarily its intrinsic vowel duration. The effect of Gender was not significant, and did not interact with either within-subject factor [$F<1$, $p>0.1$].

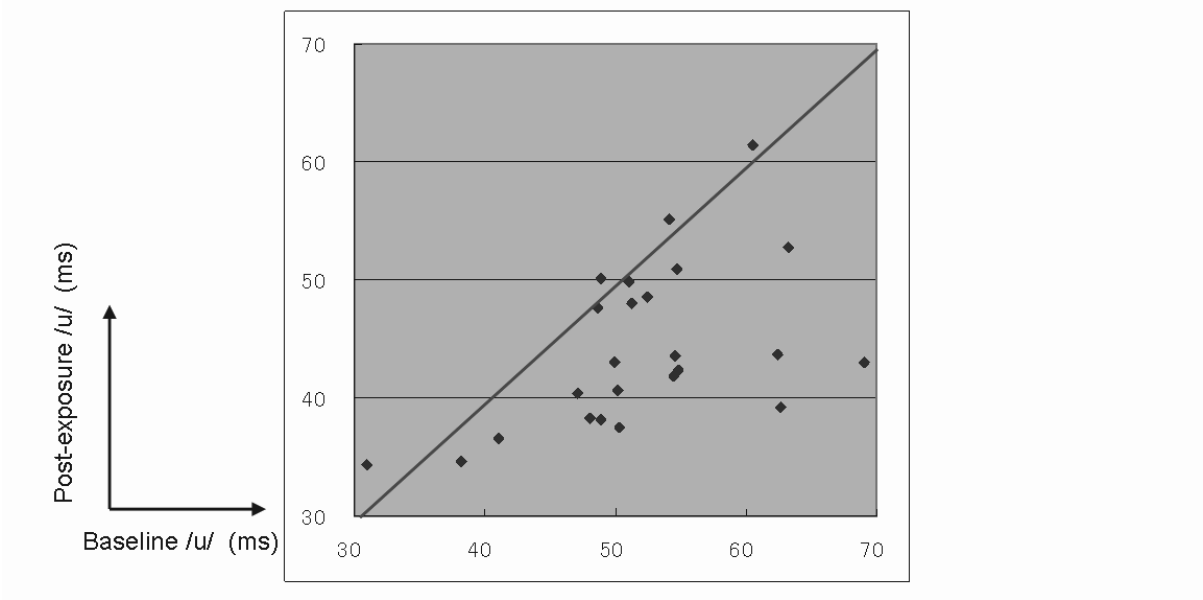

Figure 4-8: Individual variability in vowel duration and phonetic imitation
Mean voiced /u/ duration (in devoicing environments) of individual participants in Experiment 3. The baseline /u/ duration is plotted on the x-axis, and post-exposure /u/ duration is plotted on the y-axis (R^2=0.26). The diagonal line's slope is one and indicates no imitation (baseline = post-exposure).

Consistent with Experiments 1 and 2, a large individual variability was observed in the magnitude of imitation. Figure 4-8 illustrates the speaker variability of vowel duration and degree of imitation observed in Experiment 3. As seen, although most speakers decreased devoiceable (yet voiced) /u/ duration after listening to the target speech with extreme vowel deletion/devoicing, the degree of change (or imitation) varied greatly across participants.

4.3.2 Generalizability

The generalizability of phonetic imitation was examined at two levels: phoneme level (Target /u/ vs. Novel /u/) and feature level (Novel /u/ vs. Novel /i/).

4.3.2.1 Phoneme-level Generalization

Phoneme-level Generalization (Target /u/ vs. Novel /u/ comparison) was examined by two repeated-measures ANOVA analyses: gradient and categorical. The independent variables for both analyses were Production Type, Presence of Exposure, and Gender. The dependent variables for the gradient and categorical analyses were the duration of voicing and the rate of complete devoicing of /u/ in Target words, respectively.

According to the gradient analysis, the effect of Production Type was significant [$F(1,22)=24.37$, $p<0.01$*], revealing that phonetic imitation occurred in both Target /u/ and Novel /u/. The effect of Presence of Exposure was not significant [$F<1$, $p>0.1$], while the interaction between the two was significant [$F(1,22)=4.78$, $p<0.05$*], indicating that imitation was greater for actually heard items. The effect of Gender was not significant, and did not interact with either within-subject factor [$F<1$, $p>0.1$]. Figure 4-9 (a) illustrates the result of the gradient analysis for Target vs. Novel comparison: As seen, the duration of voicing was reduced in *both* Target and Novel stimuli. In other words, the extreme degree of devoicing on /u/ in the target speech was imitated, and the imitation was generalized across new words containing the same vowel.

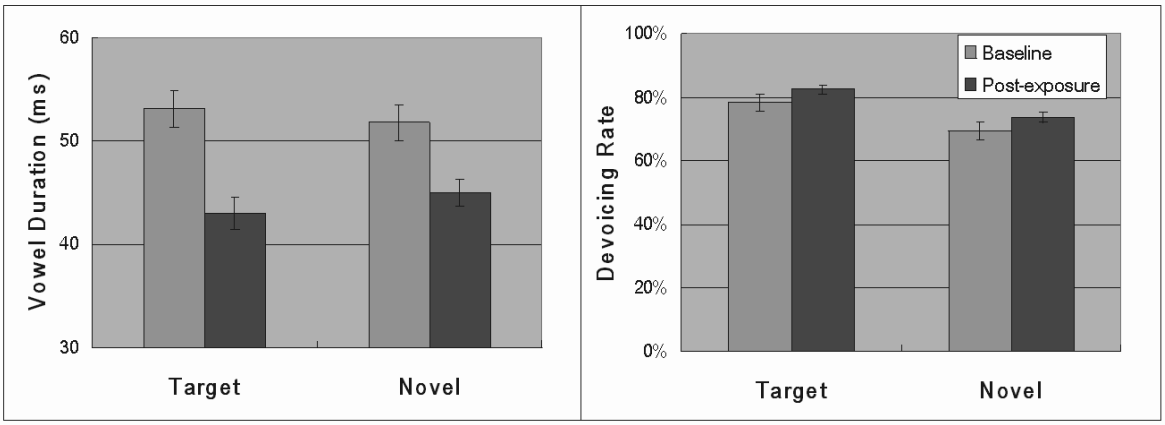

Figure 4-9: Generalization of phonetic imitation at phoneme level
Mean (a) vowel duration and (b) devoicing rate and standard errors of the means, presented separately for the baseline and post-exposure productions, and for Target and Novel words. Both figures show that imitation was generalized to Novel stimuli containing /u/.

The categorical analysis revealed significant main effects of Production Type [$F(1,22)=6.04$, $p<0.05$*] and Presence of Exposure [$F(1,22)=67.81$, $p<0.01$*]. The interaction between the two factors was not significant [$F<1$, $p>0.1$], revealing that the degree of imitation did not differ between Target and Novel items. That is, the categorical analysis showed a stronger pattern of generalization than the gradient analysis. The effect of Gender was not significant, and did not interact with either within-subject factor [$F<1$, $p>0.1$]. Figure 4-9(b) illustrates the results of the Target vs. Novel comparison in the categorical analysis, showing that phonetic imitation was generalized to Novel items. The observed baseline difference between Target and Novel items was not expected. However, this pattern (i.e., the lower baseline rate of complete devoicing for Novel words) does not affect the finding of generalization.

4.3.2.2 Feature-level Generalization

Generalization at the feature level was examined by comparing the imitation patterns of Novel /u/ and Novel /i/. The independent variables were Production Type, Vowel (Novel /u/ and Novel /i/), and Gender.

Phonetic imitation of extreme devoicing was generalized at a feature level: The results of the gradient analysis revealed significant main effects of both Production Type [$F(1,22)=47.30$, $p<0.01$*] and Vowel [$F(1,22)= 7.31$, $p<0.01$*]. The interaction between the two variables was also significant [$F(1,22)=4.91$, $p<0.05$*]. The effect of Gender was not significant. Figure 4-10(a) illustrates the results of the Novel /u/ and Novel /i/ comparison in the gradient analysis, showing that after listening to the target speech, participants' vowel duration decreased for both Novel /u/ and Novel /i/. Note, however, that the observed interaction between the degree of imitation and vowel quality (i.e., Novel /u/ and Novel /i/) does not show phoneme specificity in imitation (i.e., stronger imitation for /u/), because the degree of imitation was actually larger for /i/, as seen in Figure 4-10(a).

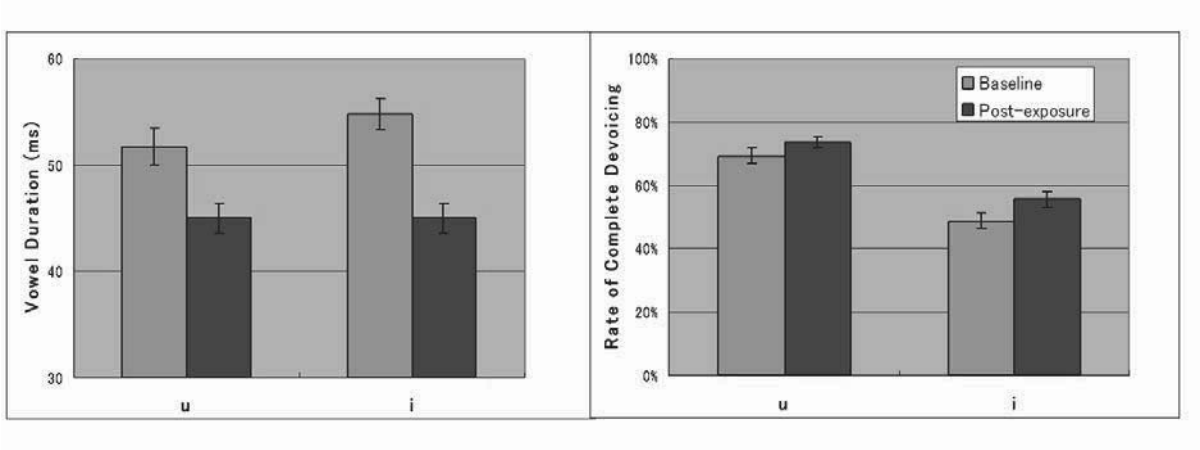

Figure 4-10: Generalization of phonetic imitation at a feature level
Mean (a) vowel duration and (b) devoicing rate and standard errors of the means, presented separately for the baseline and post-exposure productions, and for Novel /u/ and /i/. Both figures show that imitation was generalized to the new vowel /i/.

The results of the categorical analysis also revealed generalization of phonetic imitation at the feature level: The effects of Production Type [F(1,22)=22.51, p<0.01*] and Vowel [F(1,22)= 388.69, p<0.01*] were both significant, while the interaction between the two variables was not [F(1,22)=1.123, p>0.1]. There was no effect of Gender [F(1,22)=3.36, p>0.05], nor interaction with other variables. Figure 4-10(b) illustrates the results of the Novel /u/ and Novel /i/ comparison in the categorical analysis: As seen, the rate of complete devoicing increased for both Novel /u/ and Novel /i/. As described in 4.3.1.2, the baseline vowel difference was not controlled, however, that difference does not challenge the feature-level generalization observed here.

4.3.3 Word Specificity

Lexical Frequency and Presence of Exposure were controlled as independent variables in Experiment 3 in order to examine the effect of word-level specificity on the degree of imitation. As described earlier, word specificity would mean that low frequency words as

well as words with more exposure show a stronger imitation effect.

4.3.3.1 Lexical Frequency

In order to examine the effect of Lexical Frequency on the degree of imitation, two repeated-measures ANOVA analyses were performed. The independent variables for both analyses were Production Type, Lexical Frequency, and Gender. The dependent variable was the duration of target vowels for the gradient analysis, and the rate of complete devoicing for the categorical analysis.

The gradient analysis found significant main effects of both Lexical Frequency [$F(1,22)=4.61$, $p<0.05$*] and Production Type [$F(1,22)=45.87$, $p<0.01$*], while the interaction between the two was not significant [$F<1$, $p>0.1$]. Figure 4-11(a) illustrates the effect of lexical frequency and phonetic imitation on vowel duration: As seen, there was a difference between the two frequency groups on vowel duration (i.e., vowels were slightly shorter in low-frequency words than high-frequency words), but they did not differ in the degree of imitation. The effect of Gender was not significant, and did not interact with either within-subject factor [$F<1$, $p>0.1$].

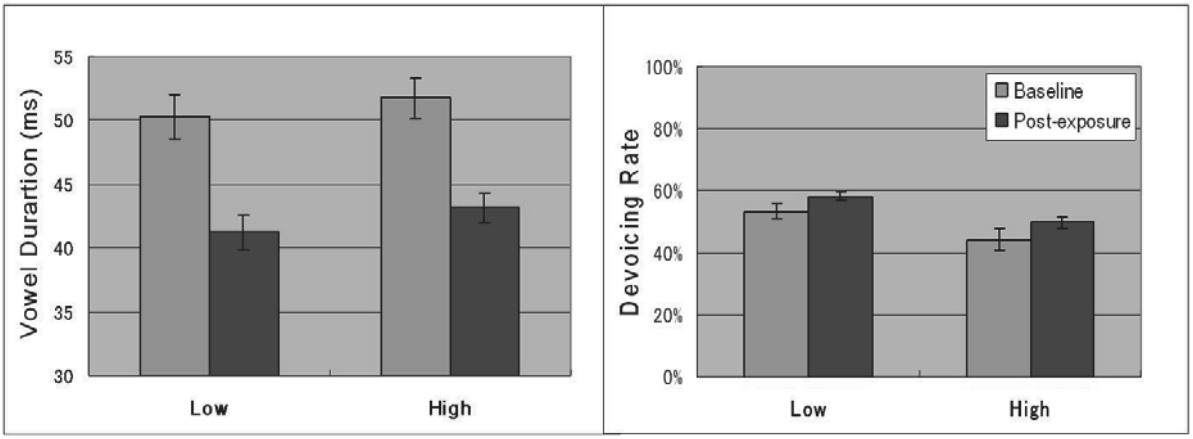

Figure 4-11: Effect of lexical frequency
Mean (a) vowel duration (ms) and (b) complete devoicing rate (%) and standard errors of the means, presented separately for the baseline and post-exposure productions, and for the Low and High frequency words. Vowels in low frequency words were slightly shorter, and devoiced more often than vowels in high frequency words. The degree of imitation was not affected by lexical frequency.

The categorical analysis also revealed significant main effects of Lexical Frequency [$F(1,22)=24.89$, $p<0.01$*] and Production Type[$F(1,22)=5.13$, $p<0.05$*], while there was no interaction between the two factors [$F<1$, $p>0.1$]. Again, the effect of Gender was not significant, and did not interact with either within-subject factors [$F<1$, $p>0.1$]. Figure 4-11(b) illustrates the effect of lexical frequency and phonetic imitation on the rate of complete devoicing occurrence for the target vowels, which the participants listened to during the experiment. As seen, the rate of devoicing was higher for low frequency words than for high frequency words. However, the degree of imitation did not vary across the two frequency groups. In sum, regarding lexical frequency, there was no word specificity found in Experiment 3.

4.3.3.2 Presence of Exposure

Next, the effect of Presence of Exposure (comparison between Target and Novel words) on

the degree of imitation was examined. The independent variables for both analyses were Production Type, Presence of Exposure, and Gender. The dependent variables for the gradient and categorical analyses were the duration and rate of complete devoicing of /u/ in Target words, respectively.

As described in 4.3.2.1 (phoneme-level generalization), the effect of Presence of Exposure was not significant [F<1, *p*>0.1], while that of Production Type was significant [F(1,22)=24.37, *p*<0.01*]. The interaction between the two was significant [F(1,22)=4.78, *p*<0.05*], revealing the word specificity of phonetic imitation. Figure 4-12(a) shows the effect of Presence of Exposure and phonetic imitation on the duration of /u/: As seen, Target words showed stronger imitation, as expected on the exemplar view. The effect of Gender was not significant, and did not interact with either within-subject factor [F<1, *p*>0.1].

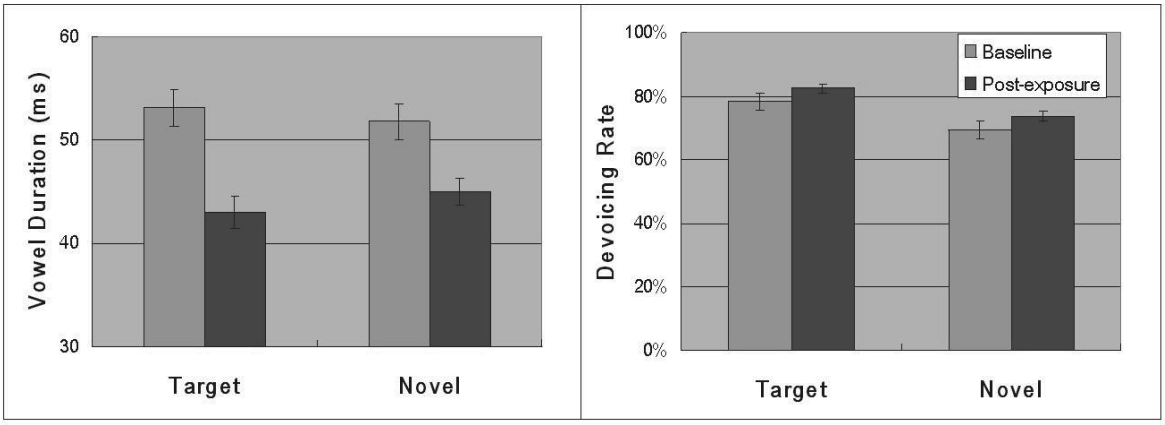

Figure 4-12 [repeated from 4-9]: Effect of target speech exposure
Mean (a) vowel duration and (b) devoicing rate and standard errors of the means, presented separately for the baseline and post-exposure productions, and for Target and Novel words. (a) shows that the degree of imitation measured by vowel duration was affected by the exposure to target speech.

Also as described in 4.3.2.1 (phoneme-level generalization), there were significant main effects of Presence of Exposure [$F(1,22)=67.81$, $p<0.01$*] and Production Type [$F(1,22)=6.04$, $p<0.05$*] in the categorical analysis, while there was no interaction between the two factors [$F<1$, $p>0.1$]. Again, the effect of Gender was not significant, and did not interact with either within-subject factors [$F<1$, $p>0.1$]. Figure 4-12(b) shows the effect of Presence of Exposure and phonetic imitation on the rate of complete devoicing of /u/ in Target words. As seen, the rate of devoicing was not higher for Target words than for Novel words, showing no word specificity.

In sum, word specificity was found only in the gradient analysis which examined Presence of Exposure. When the vowel durations were compared, the data revealed a greater degree of imitation for Target words than Novel words. On the other hand, the degree of imitation did not differ between low and high frequency words. As for the rate of complete devoicing, neither exposure to target speech nor lexical frequency showed evidence for word specificity. Note that unlike in Experiment 1, a post-hoc analysis on word specificity was not performed in Experiment 3, due to the fact that a lack of imitation was not an issue (i.e., of twenty-four participants, eighteen displayed over 5% imitation in the gradient analysis of /u/; see Figure 4-8).

4.3.4 Marked Environments

In order to examine the imitation pattern of highly marked environments, data from the following devoicing environments, described in the literature as showing low devoicing rates, were analyzed separately: 1) consecutive (Tsuchida, 1997), 2) mid-vowel, 3) pre /h/ (Maekawa and Kikuchi, 2005; Fujimoto, 2004), and 4) word-final (Martin, 1952). Highly

marked means that *complete devoicing* (categorical) is uncommon. Nonetheless, in our results the duration of mid-vowels decreased, as for high vowels (see 4.3.1.1) despite their highly marked status. For this reason, only the categorical results were examined in this section by repeated-measures ANOVAs. The independent variables used in all analyses were Production Type and Gender. The dependent variable was rate of complete devoicing. All types of stimuli (i.e., Target and Novel; Low and High frequency) were included in the analysis.

4.3.4.1 Consecutive Devoicing

Consecutive vowels are regard as a *phonologically* marked environment for devoicing: Kondo (2005) argued that consecutive devoicing is blocked by Japanese syllable structure, which bans triple consonant clusters. In Tsuchida's OT analysis (1997), where devoiced vowels are represented with the feature [+spread glottis], consecutive devoicing is completely blocked by the undominated constraint OCP [+spread glottis]. On the other hand, in the large corpus study of connected speech by Maekawa and Kikuchi (2005), consecutive devoicing was observed in 29% of all eligible cases.

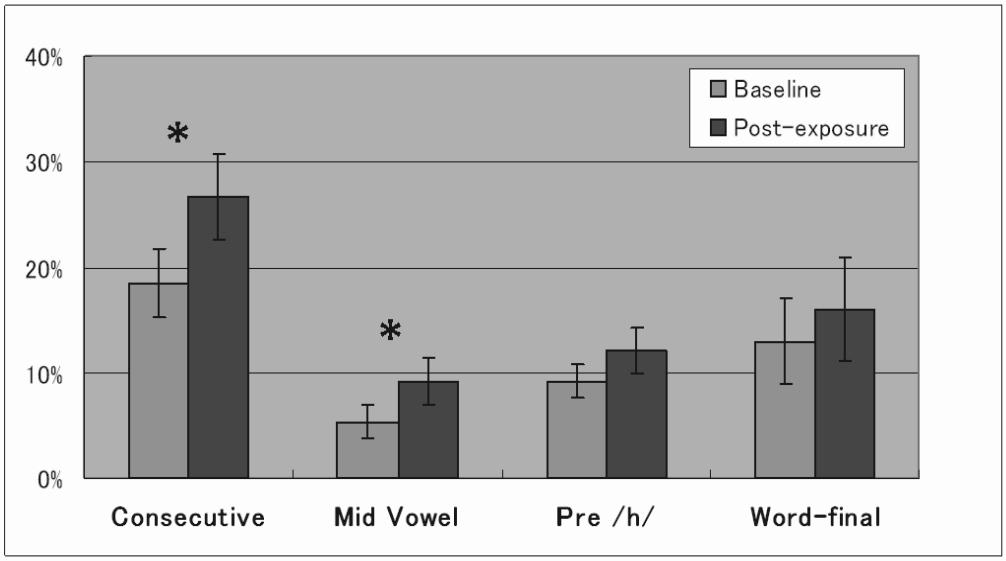

Figure 4-13: Phonetic imitation in phonologically marked environments
Mean occurrence of complete devoicing and standard error of the mean, presented separately for the baseline and post-exposure productions, and for the four phonologically marked environments tested: *1) Consecutive, 2) Mid-vowel, 3) Pre-H, and 4) Word-final.* The figure shows that 1) consecutive devoicing was the most common among the conditions tested, and 2) consecutive and mid-vowel devoicing were imitated.

Extreme devoicing in a (categorical) consecutive environment was imitated: A repeated-measures ANOVA analysis (independent variables: Production Type and Gender) revealed a significant main effect of Production Type [$F(1,22)=18.23$, $p<0.01$*]. There was no main effect of Gender, and no interaction with Production Type [$F<1$, $p>0.1$]. As seen in Figure 4-13 (*Consecutive*), the rate of consecutive devoicing increased from 18% (baseline) to 27% (post-exposure) in the word-naming task. Note that these rates are lower than in Maekawa and Kikuchi (2005), likely due to the differences between isolated words in a list vs. connected speech.

4.3.4.2 Mid-vowel Devoicing

After exposure to target speech with extreme (i.e., 100%) mid-vowel devoicing, the rate of

mid-vowel devoicing in participants' speech increased significantly (Production Type [$F(1,22)=5.14$, $p<0.05$*]) (see Figure 4-13, *Mid Vowel*). Again, neither a main effect of Gender nor its interaction with Production Type was observed [$F<1$, $p>0.1$]. As expected, the rate of mid-vowel devoicing was very low. Neither vowel quality (i.e., e vs. o) nor lexical frequency appears to affect the rate of mid-vowel devoicing. However, the presence of pitch accent does seem to block the devoicing of mid vowels, since there were no cases of devoicing in the data in which the vowel was mid and carried pitch-accent.

4.3.4.3 Pre /h/ Devoicing

In the feature classification of Halle and Stevens (1971), /h/ is assigned [+spread glottis] and thus a vowel which follows a voiceless obstruent and precedes /h/ is expected to devoice according to existing phonological analyses assuming [+spread glottis] (e.g., Tsuchida, 1997). However, Maekawa and Kikuchi (2005) showed that devoicing before /h/ is highly marked, with the lowest devoicing rate (33.88%) among all the voiceless obstruent contexts tested.

Our data showed that rate of the complete devoicing preceding the consonant /h/ did not change significantly after the target exposure (Production Type [$F(1,22)=2.77$, $p>0.1$]; Gender [$F<1$, $p>0.1$]), indicating that extreme devoicing in the target speech was not imitated. As seen in Figure 4-13 (*Pre /h/*), the devoicing rates (baseline - 9%, post-exposure - 12%) were notably lower than the rate reported in Maekawa and Kikuchi (2005). The elicitation style and small data size in the current study might contribute to this difference. In addition, the current analysis was based on only ten words containing a pre-/h/ vowel. In sum, our data clearly showed that pre-/h/ devoicing is highly marked and resisted imitation, confirming the finding by Maekawa and Kikuchi (2005).

4.3.4.4 Word-final Devoicing

Word-final devoicing has been treated as optional in previous studies (e.g., Sugito, 1988; Hasegawa, 1999), although the actual devoicing pattern is less studied. As seen in *Figure 4-13* 4-13 (*Word-final*), the current analysis revealed that the rate of devoicing is quite low (baseline - 13%, post-exposure - 16%). According to a repeated measures ANOVA, the main effect of Production Type was not significant [$F(1,22)=3.4$, $p>0.05$]. On the other hand, the effect of Gender was significant [$F(1,1)=10.49$, $p<0.01$*]. The interaction between the two factors was not significant [$F<1$, $p>0.1$]. Figure 4-14 shows this difference in word-final devoicing in Japanese: As seen, male participants showed a higher rate of complete devoicing in word-final position, both Baseline and Post-exposure.

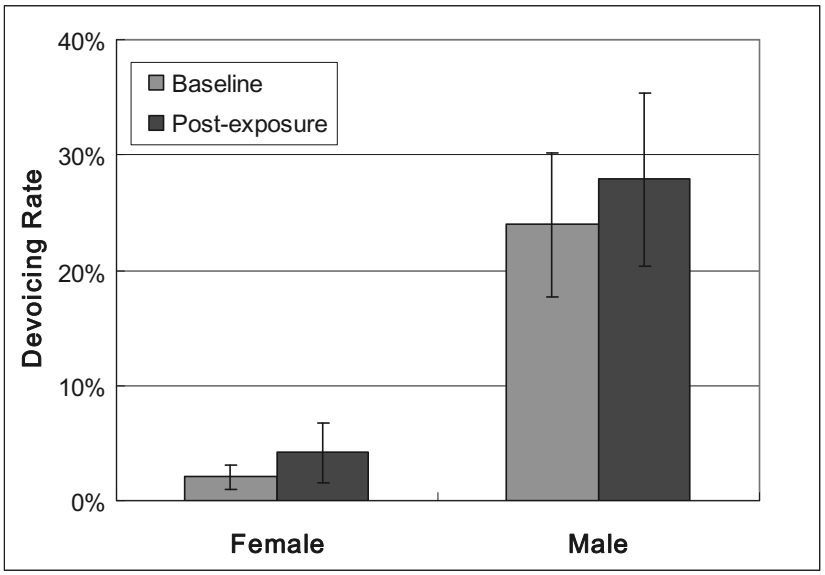

Figure 4-14: Gender difference of word-final devoicing in Japanese
Mean occurrence of complete devoicing in word-final positions and standard error of the mean, presented separately for the baseline and post-exposure productions, and for gender. Male participants showed a higher degree of complete word-final devoicing than female participants.

In addition to the gender difference, large speaker variability was found in the data, as seen from the error bars in Figure 4-14. In fact, half the participants did not produce *any* word-final devoicing. Among the "devoicers" who produced final devoicing more than once (11 out of 24 participants, 9 males), the average rates of devoicing in baseline and post-exposure were 34% and 39%, respectively (as opposed to 14% and 17% in Figure 4-13, *Word-final*).

Martin (1952) pointed out that a high vowel in word-final position is devoiced only if one of the preceding syllables is accented. Although there were some counterexamples (i.e., word-final devoicing with no preceding pitch accent), a trend consistent with this claim was observed in the current data: A t-test comparing words with and without preceding pitch accent showed that their difference in the rate of word-final devoicing is significant [$t=2.09$, two-tailed $p<0.01$*]. Taken together, then, word-final devoicing appears to be a process which is potentially controlled by phonological, phonetic, and socio-linguistic factors.[7]

4.4 Discussion

4.4.1 Phonetic Imitation

The results from Experiment 3 revealed that the extreme degree of Japanese vowel devoicing/deletion in the target speech was imitated by the participants, and that the imitation was generalized to new instances of the modeled vowel /u/ as well as to a new vowel /i/ in both categorical and gradient analyses. In post-exposure production, the rate of complete devoicing increased from 62% (baseline) to 67% (post-exposure), and the duration of vowels in devoicing environments decreased from 53 ms (baseline) to 44 ms (post-exposure).

[7] Potentially other factors, not included in the present experiment, are also involved: For example, Maekawa and Kikuchi (2005) reported extremely high rates of word-final devoicing for Japanese copulas (*i.e., desu* and *masu*), suggesting an effect of lexical properties (e.g., function vs. content words).

Further, a word-specific pattern of imitation was observed in a gradient analysis comparing Target and Novel items. Vowels in the words that were heard by the participants exhibited a greater degree of imitation (i.e., shorter vowel durations) compared to vowels in the words that were not heard (Figure 4-9(a)). On the other hand, the expected effect of lexical frequency on the degree of imitation (i.e., greater imitation for low-frequency words) was not observed.

These findings are broadly consistent with the results in Experiment 1, namely: 1) the mechanism of phonetic imitation possesses abstract and episodic components, 2) both featural and lexical representations contribute to phonetic imitation, and 3) the effect of feature-level generalization is robust in phonetic imitation, while the effect of word specificity is weak. By obtaining the same pattern of results for two different types of allophonic variability (VOT and vowel devoicing) in two different languages, we have demonstrated the robustness of these findings.

More importantly, Experiment 3 showed that phonetic imitation of allophonic variability occurs in environments where various phonetic/phonological factors control the output, and that imitation affects the categorical phonetic (or, allophonic) realization (i.e., presence/absence of a vowel) in addition to the gradient phonetic realization (i.e., shorter vowel duration). Further, in the analyses that examined imitation patterns in phonologically marked environments, the rate of complete consecutive devoicing and mid-vowel devoicing increased significantly. These results suggest that details of perceived allophonic variability indeed have consequences for phonological representations, contrary to the assumption by the abstractionist view. The relationship between (phonologically irrelevant) fine phonetic details and phonological representations is a central issue in this study, and is discussed

further in the general discussion (Chapter 5).

In addition, a significant gender difference was observed in the analysis examining the rate of word-final devoicing. The difference could be due to low-level phonetic factors such as physiologically based variability in voice quality (N.B. regardless of impressionistic overall voice quality, many participants displayed word-final creakiness); it could also be due to sociolinguistic factors such as perceived femininity/masculinity or politeness (*cf.* Klatt and Klatt, 1990; Ito, 2004). Given that a gender difference in devoicing rate has not been reported in the previous literature, further investigation with a larger sample size is required in order to elucidate its nature.

4.4.2 Lexical Frequency & Japanese Devoicing

Experiment 3 revealed that vowels in low-frequency words are devoiced to a greater extent: The gradient analysis showed shorter vowel durations in low frequency words, and the categorical analysis showed higher rates of complete deletion among low frequency words. These rather counterintuitive findings are indeed consistent with Maekawa and Kikuchi (2005), who reported a slight negative correlation between lexical frequency and devoicing rate when a high vowel is preceded by a voiceless constant and followed by a *voiced* consonant.

If Japanese vowel devoicing is a process of phonetic reduction, we would expect a higher instance of devoicing among high frequency words, as shown in English (e.g., Jurafsky *et al.*, 2001; Fidelholz, 1975). Further, Ernestus et al. (2006) found a positive correlation between frequency and degree of voice assimilation in Dutch, suggesting that higher frequency leads not only to acoustic reduction but also to more assimilation.

The reverse lexical frequency effect observed in Experiment 3 suggests that Japanese vowel devoicing is not a process of phonetic reduction or assimilation. The effect of a lexical property on coarticulation found by Scarborough (2004) might provide an account for our findings. She observed a higher degree of coarticulation among words with low frequency and high neighborhood density (= low *relative frequency*[8]), and interpreted this result in light of the idea that coarticulation is a perceptually useful source of linguistic information. Table 5 illustrates the effects of frequency and neighborhood density on lexical difficulty (taken from Table 12 in Scarborough, 2004). Although an effect of lexical frequency on coarticulation has not previously been found (= the missing cells in Table 5), Japanese vowel devoicing would be such a case, if vowel devoicing can be construed as coarticulation.

Ogasawara (2007) provides support for this interpretation (specifically, that devoicing provides perceptual cues to facilitate word recognition). In her word recognition tasks, listeners performed better when vowels were devoiced in the environment where vowel devoicing was expected. Given that low frequency words are more difficult in both word recognition (perception) and retrieval (production) (e.g., Howes, 1957; Jescheniak and Levelt, 1994), the higher degree of devoicing observed in Experiment 3 suggests that Japanese vowel devoicing is not due to assimilation or reduction, which is perceptually less salient. Rather, it is more likely a coarticulation which provides perceptual cues to facilitate word recognition.

[8] Calculated by dividing the log frequency of the word by the sum of the log frequencies of the word and all of its neighbors.

Table 5: The effects of frequency and neighborhood density on lexical difficulty (from Table 12 in Scarborough, 2004)

word type	**listener** (recognition)	speaker (retrieval)	effect on hyperarticulation	effect on coarticulation
low frequency	**hard**	hard	**less reduced**	?
high frequency	easy	easy	more reduced	?
low density	easy	hard	more reduced	less coartic.
high density	**hard**	easy	**hyperarticulated (less reduced)**	**more coarticulation**

Note that the finding in Imaizumi et al. (1995) (i.e., reduced degree of devoicing in speech directed toward hearing-impaired children) does not challenge this interpretation, but rather helps us to understand the relationship between coarticulation, which arises from (*temporal*) gestural overlap between neighboring segments, and hyperarticulation, which can be represented as an increase in the size (or *amplitude*) of a gesture. As seen in Figure 4-15 (from Figure 2 in Scarborough, 2004), which illustrates hyperarticulation and coarticulation of two adjacent segments, the two mechanisms can work independently: If both the amplitude of a gesture and the speed of an articulator are increased, the resulting hyperarticulation will have a reduced degree of coarticulation (left panel). On the other hand, if the amplitude of a gesture is increased *without* changing the overall speed of articulation, the resulting hyperarticulation will have an increased degree of coarticulation (right panel). The observed effect of lexical frequency on Japanese devoicing, assuming that it is construed as coarticulation, could be parallel to the scenario in the right panel: greater amplitude and duration of consonant gestures due to hyperarticulation, giving rise to greater C-V overlap and thus to more prominent coarticulatory cues. It is conceivable that speakers are (subconsciously) aware that increased coarticulation may benefit listeners with normal

hearing, while it does not benefit hard-of-hearing listeners, and that they are capable of making adjustments in both amplitude and velocity of articulators according to their speech environments.

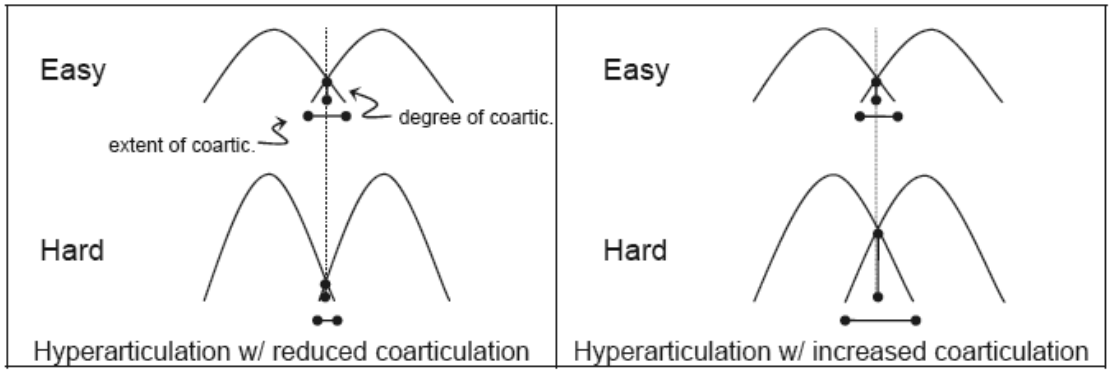

Figure 4-15: Hyperarticulation with reduced or increased coarticulation (From Figure 2 in Scarborough, 2004)
'Easy' refer to words with high relative frequency, and 'Hard' refers to words with low relative frequency.

This is clearly a very important issue in both speech perception and production, and should be investigated further. However, given that certain important factors (e.g., neighborhood density) were not fully counter-balanced between the two frequency groups in the current study, future investigation with carefully selected stimuli is required to address this issue.

4.4.3 On the Mechanism of Japanese Devoicing

As described in 4.1, the traditional account of Japanese vowel devoicing is based on a phonological feature-changing (or assimilation) rule (e.g., McCawley, 1968; Vance, 1987).

Alternatively, Jun and Beckman (1993) argued that Japanese devoicing is a gradient phonetic process, and proposed a gestural overlap account of devoicing that is based on Browman and Goldstein (1989). Our data provide support for the latter account from two points of view, each of which will be discussed below.[9]

4.4.3.1 Effect of Lexical Frequency

As described in 4.4.2, previous studies have shown that high frequency words are more likely to be reduced or assimilated (e.g., Jurafsky et al., 2001; Ernestus et al., 2006), while words with low relative frequency show a higher degree of *coarticulation* (Scarborough, 2004). Our finding of a lexical frequency effect (i.e., higher rate of devoicing among low frequency words) suggests that the mechanism underlying Japanese vowel devoicing is not assimilation or reduction, but rather coarticulation, which provides perceptual cues to facilitate word recognition. While assimilation is more consistent with a feature-changing account, (non-reduction) coarticulation is more consistent with a gradient phonetic account.

4.4.3.2 Segmental Duration

Kondo (2005) showed that consonants preceding completely devoiced vowels were longer than consonants preceding (devoiceable) voiced vowels. Similarly, Davidson (2006) examined pretonic schwa elision in English, and showed that consonants preceding complete schwa deletion were longer than consonants preceding schwa. Our data on both consonant and vowel durations are consistent with these previous findings on segmental duration: The

[9]Although our data suggest that Japanese vowel devoicing is a gradient phonetic process as proposed by Jun and Beckman (1993), they do not necessarily provide exclusive support for a gestural overlap account. That is, the data can also be explained by other accounts such as adjustment in acoustic segmental duration.

average duration of voiceless obstruents preceding complete devoicing was significantly longer than that of voiceless obstruents preceding voiced vowels (141 ms and 113 ms, respectively), and the average duration of vowels in devoicing environments was significantly shorter than the average duration of all vowels (53 ms and 97 ms, respectively).

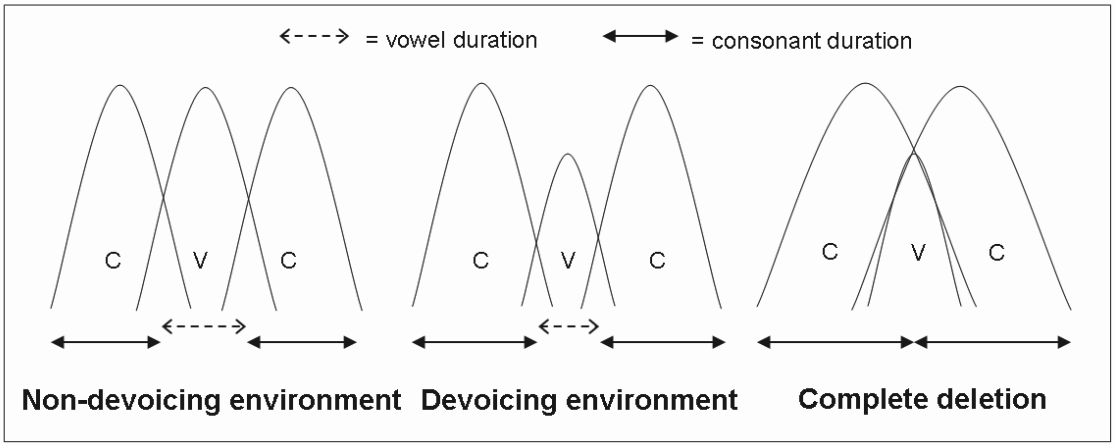

Figure 4-16: Schematized Japanese vowel devoicing as a phonetic process

A gradient phonetic account would readily predict these effects, namely consonant lengthening with devoiceable vowel reduction. Figure 4-16 shows Japanese vowel devoicing as a schematized phonetic process. As seen in the left panel, vowels are not reduced or shortened in non-devoicing environments. In contrast, in devoicing environments (middle panel) vowels are shorter and reduced because the wide glottal gesture of the surrounding voiceless consonants causes the consonants to lengthen. The right panel shows how further consonant lengthening and overlap may result in complete or near-complete deletion, with no voiceless vowel.

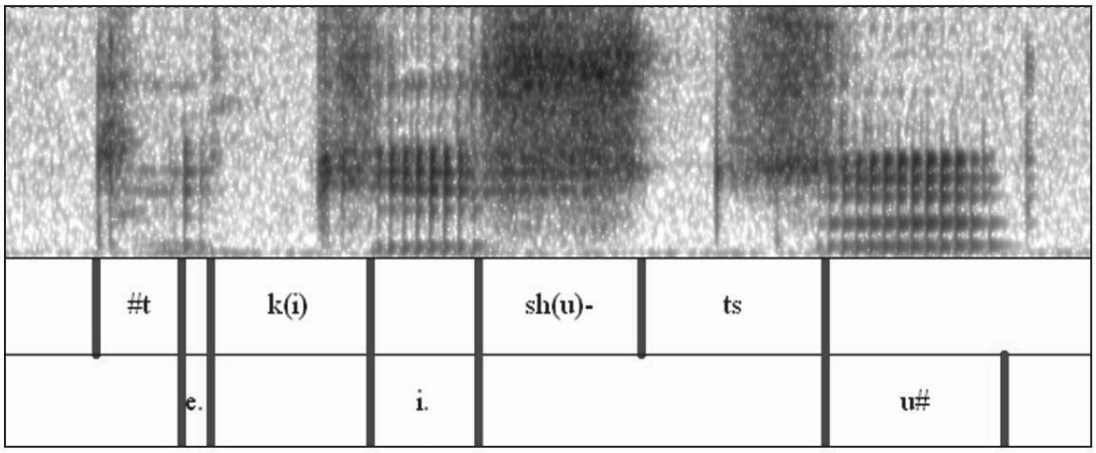

Figure 4-17: Example of extreme partial devoicing
The voicing of /e/ (duration = 20ms) was not audible.

Further, many cases of extreme (yet) partial devoicing (or voiced vowel shortening) were observed in the current data, in which the devoiceable vowels were not audible, yet were visible in the acoustic display. Figure 4-17 shows an example of such *partial* devoicing: The vowel [e] was classified as "voiced" in the categorical analysis due to the presence of glottal pulses (voice bars) (*cf.* Jun and Beckman, 1994) despite the fact that the vowel voicing was not audible. These observations show that Japanese vowel devoicing is indeed a gradient process rather than a categorical one (Jun and Beckman, 1993), and suggest that apparent vowel deletion can in fact be the endpoint on a continuum of shortening.

Taken together, our data suggest that Japanese vowel devoicing is a gradient phonetic process: A phonological voice assimilation process (i.e., V [+high] → [–voice] / [–voice] __ [–voice]) would predict the presence of visible devoiced vowels, which was absent in most cases in our data. Further, our data showed many cases of partial devoicing (acoustic vowel shortening) as well as longer consonant durations preceding complete

devoicing, neither of which would be predicted by an assimilation account. In contrast, these patterns are in agreement with the gradient phonetic account proposed by Jun and Beckman (1993).

4.5 Interim Summary

The results from Experiment 3 showed that the extreme degree of Japanese vowel devoicing (realized as vowel deletion) of the modeled vowel /u/ was imitated in environments where various phonetic/phonological factors control the output, according to both categorical and gradient measures. Further, the change was generalized to novel items with /i/, suggesting that the change in speech signal was coded at a feature level. On the other hand, the *degree* of imitation was shown to be at least in part word-specific. Compared to Novel items (which participants did not listen to), a slightly greater degree of imitation was observed for Target items (which participants listened to during the experiment) in the gradient analysis. These findings confirm our previous observations in Experiment 1 that 1) featural and lexical representations simultaneously contribute to phonetic imitation, and 2) the effect of feature-level generalization is robust in phonetic imitation, while the effect of word specificity is more subtle. These results suggest that perceived allophonic information is retained after lexical access, which has consequences for phonological representations. As for the mechanism of Japanese vowel devoicing, our results were more compatible with the gradient phonetic account than the phonological feature assimilation account.

Chapter 5: Discussion and Conclusion

5.1 Summary of Results

This dissertation investigated the nature of phonological representations by examining how variations in perceived speech signals are reflected in speech production. Using a new experimental design based on the imitation paradigm (Goldinger, 1998), three experiments were conducted to investigate how manipulated phonetic features/allophonic variations in target speech were imitated by speakers. The data were analyzed in light of the following research objectives: 1) the imitability of allophonic variability, 2) the generalizability of phonetic imitation at the phonemic and featural levels, 3) word-level specificity of phonetic imitation, and 4) the interaction between phonetic imitation and the presence of a phonemic boundary or other phonetic/phonological factors. The phonetic feature manipulated in the experiments was [+spread glottis] (aspiration/voicelessness). Experiments 1 and 2 examined the phonetic imitation of *lengthened* and *shortened* VOT in English, respectively, in which the *degree* of [+spread glottis] in the target speech was manipulated by changing VOT values on the modeled phoneme /p/. Experiment 3 examined the imitation of Japanese high vowel devoicing, in which allophonic realization of glottal gesture in the target speech was changed by eliminating any voicing on the modeled phoneme /u/ in devoicing environments (i.e., deletion of a segment with [-spread glottis]).

The results from Experiment 1 showed that artificially *lengthened* VOT was imitated by the participants. Furthermore, the shift in their speech was generalized to

new instances of the target phoneme /p/ and the new phoneme /k/, showing that the imitation was generalized at a feature level. At the same time, a word-specific imitation pattern was observed in two post-hoc analyses, which compared the degree of imitation for high- vs. low-frequency words, as well as for words that were in the target speech vs. words that were not.

The results from Experiment 2 showed that artificially *shortened* VOT, which can endanger phonemic contrast with the phoneme /b/, was *not* imitated, suggesting that the need to preserve phonemic contrast suppresses phonetic imitation. Among the participants who did demonstrate imitation, the change was generalized to the new phoneme /k/.

The results from Experiment 3 showed that the (artificially created) extreme degree of Japanese vowel deletion of the phoneme /u/ was imitated, despite the fact that this violated certain phonetic and phonological constraints (that are unrelated to contrast preservation) in certain environments. Further, the change was generalized to novel items containing /i/. Consistent with the finding in Experiment 1, the *degree* of imitation was shown to be partially word-specific.

Taken together, these results provide the following observations, which relate to the research objectives described above: 1) the acoustic properties (or, fine phonetic details) of allophonic variability, that are critical for decoding the linguistic message, are retained in listeners' memory and subsequently affect their speech production, 2) multiple levels of phonological representations (i.e., word- and feature-level units) simultaneously contribute to the pattern of phonetic imitation, 3) words are not made up purely of discrete abstract units, but contain rich information of fine phonetic details that

are related to specific episodes, and 4) phonetic imitation is not an automatic process which reflects the collection of raw percepts, but rather a process which is modulated by other factors, such as the presence of linguistic contrast.

Observations 1, 2, and 3 are based on results obtained in both English and Japanese, which strengthens their credibility. Also, the results regarding word specificity (Experiments 1 and 3) provide an argument against the interpretation that the observed imitation effect is simply the result of global short-term adaptation to the target speech, rather than a reflection of long-term memory for specific exemplars.

The experiments were specifically designed to investigate the role played by three levels of phonological representations assumed in generative linguistics (i.e., word, phoneme, and feature) in phonetic imitation. Although evidence for word-level and feature-level representations was found through word-specific patterns of imitation and its generalizability from /p/ to /k/, the data did not show a phoneme-level difference of imitation (i.e., /p/ vs. /k/), and hence this study does not provide evidence that phoneme-level representations play a role in phonetic imitation.

In addition, large speaker variability in the degree of imitation was observed in all three experiments, as well as a gender difference in VOT, in likelihood of imitation (Experiment 1), and in rate of word-final devoicing (Experiment 3).

5.2 General Discussion

5.2.1 Phonological Representations: How episodic and abstract?

Goldinger's findings on spontaneous imitation and its word specificity (1998, 2000) provide evidence for the view that detailed episodes constitute the basis of the mental lexicon. The

current dissertation replicated these findings through acoustic measurements of one phonetic dimension, which provided a more precise measure of the imitation itself as well as quantitative information about the variability among speakers and lexical items.

In addition, the current study demonstrated that phonetic details of more *allophonic* variability are also retained in word-level representations, challenging the invariant view of phonological representations. As described in the Introduction (1.2), the specificity effect has previously been found mostly in indexical variability (e.g., voice, Mullennix et al., 1989; intonation and F0, Church and Schacter, 1994; speech style, McLennan, 2003; similar voices, Goldinger, 1996), and little was known about the specificity of allophonic variability.

Our results revealed that both gradient and categorical allophonic variations were imitated by the participants, providing a stronger argument (than that offered by previous studies on indexical specificity) against the view that perceived speech signals are *reduced* to sequences of discrete, abstract representations.

More importantly, however, our data also demonstrated that the imitation was systematically *generalized* at a feature level across words that participants did not listen to during the experiment, and that the effect of generalization was more robust than the observed word specificity. These findings are broadly consistent with Carlson et al. (under review), who found a robust effect of systematic learning of Glaswegian /r/ (or transfer of an existing allophone to a new phoneme) and small lexical effects.

As for the episodicity of phonological representations, our finding of word-level specificity provides positive evidence for the episodic nature of the word form. Note that although the finding of phonetic imitation does challenge the abstract and invariant view of phonological representation, it is our finding of word-specificity which lends strong support

for the exemplar view, because the concept of a variant representation is compatible with other accounts such as statistical learning or accommodation. As described above, however, the effect of specificity was seen only weakly in the data compared to the effect of feature-level generalization, indicating that the mental lexicon not only consists of detailed word-size episodes, but also of feature-size units.

Our data were inconclusive with regard to the role played by phoneme-size representations and their episodicity. The finding of feature-level generalization hints at the "abstractness" of this representation. Note, however, that it neither proves the abstractness of featural representations nor challenges exemplar theories, so long as sub-phonemic exemplars are assumed. Our experiments were not designed to differentiate between episodic and abstract featural representations. However, if we assume *multiple levels* (or sizes) of exemplars, the seemingly large distinction between abstractionist and episodic views will diminish. In fact, the key contribution of the current dissertation is exactly this point: Mental representations of sound structures are multi-leveled, and as a consequence they can be both episodic and systematic.

5.2.2 Size/Unit of Representation

As just described, our results provide evidence in support of word-level and feature-level representations. Lexical frequency as well as the presence/absence of target exposure influenced the way words were produced, demonstrating that words contain rich phonetic information.

Although previous studies have indicated the presence of featural representation in perceptual and phonotactic learning experiments (Kraljic and Samuel, 2006; Goldrick, 2004),

little was known about its production counterpart (except for the feature accounts of speech errors, e.g., Fromkin, 1973). Given the previous findings on dissociation between speech perception and production (e.g., Labov, 1994), our finding of feature-level generalization in phonetic imitation reveals its psychological reality in speech production, strengthening the argument for the existence of featural representations. More importantly, the observed pattern of imitation (i.e., word-specificity and feature-level generalization) can only be explained by assuming that two levels of representations were simultaneously at work.

Note, however, that these *featural* representations do not necessarily correspond to the *distinctive features* that are widely used in phonology. In fact, the sub-phonemic features which were responsible for the generalization in Experiments 1 and 2 cannot be described in terms of categorical features but have to be more fine-grained, since what we manipulated was the *degree* of aspiration, not the categorical value of the feature 'spread glottis'. Although the categorical *value* of the feature [spread glottis] was changed in the target speech in Experiment 3, the imitation was observed in both categorical and gradient analyses. Furthermore, the data indicated that complete devoicing is indeed an endpoint of gradient shortening or gestural overlap, suggesting the underlying mechanisms of categorical and gradient imitation to be the same. In other words, in order to account for the results found in the current study, the featural representation (which was imitated and generalized) has to be expressed in a gradient, not categorical manner. This is consistent with our observations regarding the pattern of word specificity, namely that our phonological representations have to include rich phonetic information.

In fact, a growing body of research now focuses on the gradient nature of phonological representations. The P-map proposed by Steriade (e.g., 1999, 2001) considers

perceptual cues as conditioning factors for certain phonological alternations (instead of categorical features). Flemming (2001) provides a unified model of phonetics and phonology, which integrates the two disciplines that are often regarded as separate components of grammar in order to capture perceptual contrasts that are phonetically grounded.[10] His approach does not refute the existence of phonological categories, but rather assumes that categories are *derived* from phonetically rich representations. This assumption by Flemming was broadly born out by Lin (2005), who demonstrated that phonological categories such as features and segments can arise from unsupervised statistical learning by a computational model, which takes episodic acoustic data as its input.

As described earlier, our finding of generalization between /p/ and /k/ is in agreement with Goldrick (2004) and Kraljic and Samuel (2006), who observed generalizations between /f/ and /v/, and between /d/–/t/ continua and /b/–/p/ continua, respectively. In contrast, Wilson (2006) did not observe generalization between /k/ and /g/ in his phonotactic learning experiment, hinting at the complexity of feature-level generalization. However, this apparent discrepancy can be easily explained once we assume that *perceptual* distance and contrast are important factors in organizing features and subsequently in driving phonotactics. Perceptually speaking, the salience of a given contrast is not consistent across features or phonemes. For example, confusion matrices (e.g., Miller and Nicely, 1955; Benki, 2003) show that voicing contrasts are more salient (= less confusable) for stops than fricatives, and stop place contrasts are less salient than their voicing contrasts. These observations are in agreement with the cross-experimental discrepancy of feature-level generalizations described above, and provide further support for the view that the salience of

[10] Except for Articulatory Phonology, in which the two are already unified.

gradient perceptual contrasts determines how phonetic representations are organized, and further drive phonological categories and phonotactics.[11]

Taken together, it appears that both word-level and gradient feature-level representations play an essential role in phonetic imitation. They both contain rich phonetic information, and are modular and thus could function independently. The current study did not provide evidence for phonemic representation, and further investigation is needed in order to elucidate its role.

5.2.3 Implications for Models of Speech Perception and Production

Our results suggest that models of speech perception and production should include word-level *and* gradient feature-level representations, as well as a capacity to account for other external factors such as socio-stylistic effects. The two levels of representations need to be functionally independent, and potentially vary in their degree of episodicity. Further, the featural representations are expected to be sensitive to perceptual linguistic contrast. Some models of speech perception and production that have recently been proposed appear to be capable of predicting our main findings, namely the imitation and generalization of extended VOT and increased devoicing, as well as the absence of reduced VOT imitation.

For example, the hybrid model of speech perception and production proposed by Carlson et al. (2008) includes both neo-generative and exemplar components, and is equipped with mechanisms to index various factors such as attentional weight and sociostylistic variations. The neo-generative and exemplar components would readily predict

[11] In addition to the trend toward more gradient phonological representations, *Gradient phonotactics*, in which a speaker's internal grammar is derived from a statistical distribution of sound patterns over the lexicon (e.g., Frisch et al., 2004; Coetzee and Pater, 2006, to appear; Hayes and Wilson, 2008), captures the gradient nature of phonological grammar (also morphological grammar; e.g., Hay and Baayen, 2005).

the effects of systematic generalization and word specificity, respectively. There are three levels of processing representations assumed in its perception system: perceptual encoding, phonological parsing, and lexical access. A direct connection between perceptual encoding and lexical access is permitted, which predicts word-specific patterns of phonological changes as well as effects of lexical properties such as frequency. Its phonological parsing appears to represent phoneme-size units, predicting phoneme-level generalization (but not feature-level generalization). However, incorporating additional levels of representation into the multi-leveled model seems feasible.

The complementary learning systems (CLS) approach developed by Norman and O'Reilly (2003) and implemented by Goldinger (2007) provides a neurologically realistic account of phonological representations and learning. In this model, speech input is initially processed in the cortical system, where it undergoes abstraction, and streams of input from the cortex form episodic traces in the hippocampus. These accumulated episodes are slowly consolidated back into the cortex to create cortical representations. This cycle enables gradual learning, predicting that more abstraction takes place for more frequent input. That is, more abstraction is predicted to take place at the feature level than at the word level, and more specificity is predicted for low-frequency words than for high-frequency words.

According to Grossberg's adaptive resonance theory (ART) of speech perception (1980, 1986, 2003), speech input activates *items* (which are composed of feature clusters) in working memory, which in turn activate *list chunks* in short-term memory. These chunks then resonate with the input, and the resonance between input and chunk constitutes the percept (Grossberg, 1986). List chunks correspond to possible groupings of items that may vary in size, such as word, phoneme, or even feature, and could refer to tiered processing levels of

representations. In McLennan et al. (2003), list chunks were specified as surface forms and abstract forms (which they called *allophonic* and *phonemic*, respectively), while in Vitevitch and Luce (1999), list chunks were composed of sub-lexical and lexical units. McLennan et al. (2003) proposed that speech input takes the form of specific and veridical surface representations that preserve articulation style, and when speech input is perceived, its surface forms resonate with chunks that correspond to each level of representation. To interpret the result of the current study within the ART framework, we could assume list chunks to include lexical, phonemic, and featural units. We could also assume that in the absence of external factors such as attentional weight, the lexical chunk receives more activation than sub-lexical chunks by default. The hypothesized scenario for our results within this framework is the following. Suppose the extended VOT in speech input was salient to listeners, activating a featural chunk. The chunk then resonated with the input, constituting the percept which included extended VOT. The percept was subsequently reflected in production, resulting in phonetic imitation and its generalization at a feature level. On the other hand, suppose the manipulated feature (i.e., reduced VOT) was not salient to listeners. Instead of creating a strong resonance between a featural chunk and the input (with reduced VOT), lexical or phonemic chunks (associated with the prototypical/default VOT value) were activated and received stronger resonance than the featural chunk, resulting in no imitation. Even if the reduced VOT was salient to listeners, if it was perceived as a bad exemplar, the "badness" can serve as an inhibitory force on the resonance at the feature level, leading other chunks (associated with the prototypical VOT value) to receive stronger resonance (*cf.* phonemic restoration, Samuel, 1981; Warren, 1984). The resulting percepts thus did not contain the reduced VOT, causing no change in production (= no phonetic imitation).

Although it is not a fully developed model, Coleman (2002) provides an account of phonological representations as an alternative to the traditional abstractionist account. This account is unique in that it does not assume any abstract representations such as phonemes. Instead, phonological structure consists of phonetic, statistical and semantic knowledge. In this view, phonological representations are "essentially phonetic, rather than symbolic-phonological" (p.96). Phonological constituents are statistical regularities over the phonetically rich representations, and contrasts among the categories are represented in two dimensions, a continuous physical scale and continuous probability scale. The combination of these scales is able to capture sociolinguistic variation, effects of lexical frequency, all the imitation patterns observed in the current study, as well as some phonological processes that require gradient representations as described in 5.2.2. This account appears to be linguistically informed, given that statistical regularities have already been shown to provide a rich source of information in language acquisition (e.g., Juczyk et al., 1994; Saffran et al., 1996).

Lastly, the Bayesian network model of multi-level phonetic imitation (Wilson and Nielsen, forthcoming) provides a detailed, quantitative account of the imitation effect for each of the participants in Experiment 1. By employing Bayesian learning techniques, our model assumes that listeners *update* their internal phonetically fine-grained probabilistic models in response to perceived target speech (i.e., lengthened VOT), which subsequently generates imitation. The multi-leveled model can correctly predict the level of generalization from /p/ to /k/ and the relatively weak effects of word specificity.

Note that some connectionist models in the field of spoken word recognition (e.g., TRACE by McClelland and Elman, 1986; Shortlist/Merge by Norris, McQueen, and Cutler,

2000) also possess multiple levels of representation, including a feature level. However, the purpose of these models is lexical access, and thus these models are not explicitly designed to account for factors such as the effects of sociolinguistic variations or lexical properties (e.g., frequency, neighborhood density) on speech production, observed in the literature as well as in the current study.

5.3 Directions for Future Work

Research on phonetic imitation is still in its early stage, and it could be extended in many ways in future work.

The results of the current dissertation have already inspired two collaborative projects: One is the Bayesian modeling of the speech production system (Wilson and Nielsen, forthcoming) described above in 5.2.4. Another project concerns sociolinguistic aspects of phonetic imitation (Graff et al., forthcoming). The goal of this project is to examine and elucidate the effect of gender and age group on imitation, in particular, to test whether listeners are more likely to imitate/accommodate speech input produced by a speaker of the same gender or age group (*cf*. Giles et al., 1999). Spontaneous imitation in and of itself may be an important component of language learning in general, and understanding when it will or will not happen not only informs socio-phonetic theories, but could also make a contribution to theories of first and second language acquisition.

The effect of lexical properties on imitation also requires further investigation. In particular, examining the interaction between neighborhood density and patterns of phonetic imitation is a natural next step. Neighborhood density is known to play a role in word recognition, making some words harder to process than others (e.g., Luce, 1986; Luce and

Pisoni, 1998; Vitevitch, 2002). Recent studies have shown the effect of relative frequency (see 4.4.2) on speech production, such as vowel space (Wright, 1997), VOT (Baese and Goldrick, submitted), and coarticulation (Scarborough, 2004). In Experiments 1 and 2 of the current study, neighborhood density was balanced between two frequency groups, making the overall relative frequency of low-frequency words lower (="harder") than high-frequency words. If "hard" words require more time in processing, the surface form may subsequently receive less attention, resulting in a reduced degree of imitation and specificity (which might have confounded the lexical effect in Experiment 1). On the other hand, allocating more attentional weight to hard words could result in increased imitation and specificity. Future work may determine which of these views best describes the true nature of imitation.

In addition, phonetic imitation provides an appropriate tool with which to examine the nature of phonological representations in a bilingual's mind. For example, an investigation on whether phonetic imitation can be generalized to novel stimuli in a bilingual's other language (at a phoneme or feature level) will likely to help determining how the phonological structures of the two languages are represented, and whether they are connected.

Developmental change in linguistic representations is another area of research in which the imitation paradigm can contribute. By using the imitation paradigm, one could investigate units and episodicity of phonological representations for small children (both pre and post literacy), and how they differ from adults', as well as how they develop over the course of language acquisition.

5.4 Conclusion

The results presented in this dissertation demonstrated that allophonic variability is imitable, indicating that perceived fine phonetic details are retained in listeners' memories, and can subsequently affect their speech production. Multiple levels of phonological representations (i.e., word and sub-phonemic unit) were shown to simultaneously contribute to the pattern of phonetic imitation, revealing that words are not purely made up of discrete abstract units, but can be episodic and abstract at the same time. Further, phonetic imitation was shown to be not an automatic process which reflects the collection of raw percepts, but rather a process which is sensitive to the presence of linguistic contrast. These observations are crucial for constructing realistic models of speech perception and production, and further contribute to advancing theories not only within phonetics and phonology, but also within a variety of related fields.

Appendix A: English Test and Filler Words with Lexical Frequency

word	Phoneme	K&F	Denity	Celex	Syll #	Hi-Lo	Read	Listen
pain	p	91	32	1505	1	1	1	1
pair	p	58	26	1201	1	1	1	1
park	p	94	18	1275	1	1	1	1
pass	p	89	24	4687	1	1	1	1
past	p	281	17	1871	1	1	1	1
peace	p	327	19	1596	1	1	1	1
pick	p	55	34	3386	1	1	1	1
point	p	395	8	7785	1	1	1	1
post	p	84	15	1535	1	1	1	1
pull	p	51	16	3330	1	1	1	1
paper	p	157	9	4039	2	1	1	1
parent	p	15	3	5679	2	1	1	1
Paris	p		10	1090	2	1	1	1
partly	p	49	4	1286	2	1	1	1
party	p	216	5	8069	2	1	1	1
passage	p	49	3	828	2	1	1	1
pattern	p	113	3	2120	2	1	1	1
payment	p	53	1	852	2	1	1	1
people	p	847	5	26543	2	1	1	1
perfect	p	58	0	1131	2	1	1	1
period	p	265	1	3716	2	1	1	1
permit	p	77	0	773	2	1	1	1
person	p	175	4	4705	2	1	1	1
picture	p	162	2	3113	2	1	1	1
pocket	p	46	7	1343	2	1	1	1
power	p	342	9	6159	2	1	1	1
public	p	438	1	4595	2	1	1	1
publish	p	3	1	1545	2	1	1	1
purchase	p	47	0	717	2	1	1	1
purpose	p	149	0	2355	2	1	1	1
permanent	p	40	0	686	3	1	1	1
personal	p	196	1	2613	3	1	1	1
policy	p	222	0	4063	3	1	1	1
politics	p	69	1	1543	3	1	1	1
popular	p	98	0	1440	3	1	1	1
positive	p	74	1	796	3	1	1	1
possible	p	373	0	6034	3	1	1	1
possibly	p	61	0	1642	3	1	1	1
poverty	p	20	0	1030	3	1	1	1

word	Phoneme	K&F	Denity	Celex	Syll #	Hi-Lo	Read	Listen
powerful	p		0	1462	3	1	1	1
pall	p	42	24	32	1	0	1	1
pave	p	2	18	41	1	0	1	1
paw	p	3	19	108	1	0	1	1
peal	p	4	33	25	1	0	1	1
peck	p	5	20	99	1	0	1	1
perk	p	1	22	33	1	0	1	1
pod	p	3	23	63	1	0	1	1
poke	p	1	27	254	1	0	1	1
pout	p	1	20	57	1	0	1	1
pulp	p	5	4	71	1	0	1	1
pallet	p	1	5	17	2	0	1	1
pamper	p	1	3	40	2	0	1	1
panther	p	1	3	149	2	0	1	1
parcel	p	1	3	213	2	0	1	1
parrot	p	1	6	71	2	0	1	1
pastry	p	4	0	163	2	0	1	1
peacock	p	2	0	69	2	0	1	1
pebble	p	1	4	156	2	0	1	1
pedal	p	5	8	55	2	0	1	1
pendant	p	1	2	26	2	0	1	1
perish	p	2	3	118	2	0	1	1
pester	p	1	7	38	2	0	1	1
piping	p	5	1	18	2	0	1	1
pirate	p	4	2	93	2	0	1	1
pivot	p	2	4	24	2	0	1	1
pointer	p	3	1	56	2	0	1	1
polo	p	4	7	62	2	0	1	1
poncho	p	3	0	37	2	0	1	1
poodle	p	2	7	20	2	0	1	1
puppy	p	2	6	123	2	0	1	1
panama	p	4	0	19	3	0	1	1
parachute	p	1	0	65	3	0	1	1
parasol	p	3	0	43	3	0	1	1
parody	p	4	1	80	3	0	1	1
patio	p	2	1	37	3	0	1	1
pitiful	p	4	0	66	3	0	1	1
polarize	p	1	0	49	3	0	1	1
polio	p	1	3	21	3	0	1	1
politic	p	4	2	18	3	0	1	1
porcelain	p	2	0	108	3	0	1	1
pamphlet	p	3	0	268	2	0	1	0
pantry	p	2	3	51	2	0	1	0

word	Phoneme	K&F	Denity	Celex	Syll #	Hi-Lo	Read	Listen
parsley	p	1	2	118	2	0	1	0
parting	p	4	0	127	2	0	1	0
patter	p	3	20	38	2	0	1	0
patty	p	1	17	62	2	0	1	0
pelvis	p	1	0	160	2	0	1	0
pickle	p	1	12	64	2	0	1	0
pidgin	p	5	1	6	2	0	1	0
pillar	p	2	9	285	2	0	1	0
pitfall	p	3	0	43	2	0	1	0
ponder	p	1	5	174	2	0	1	0
poplar	p	1	1	57	2	0	1	0
poppy	p	2	7	63	2	0	1	0
porridge	p	1	1	95	2	0	1	0
portal	p	3	5	29	2	0	1	0
possum	p	2	2	44	2	0	1	0
puddle	p	1	10	79	2	0	1	0
pumpkin	p	2	1	38	2	0	1	0
pungent	p	4	0	49	2	0	1	0
cabbage	k	4	1	188	2	0	1	0
candid	k	3	2	41	2	0	1	0
canine	k	1	0	31	2	0	1	0
canker	k	1	5	7	2	0	1	0
capsule	k	5	3	86	2	0	1	0
catchy	k	1	7	9	2	0	1	0
cater	k	4	6	249	2	0	1	0
caucus	k	2	1	36	2	0	1	0
cavern	k	1	1	54	2	0	1	0
coco	k	3	3	3	2	0	1	0
colon	k	2	0	14	2	0	1	0
comet	k	2	4	34	2	0	1	0
comrade	k	4	0	283	2	0	1	0
consul	k	3	3	55	2	0	1	0
coolant	k	4	0	6	2	0	1	0
corny	k	1	4	22	2	0	1	0
courtship	k	2	0	51	2	0	1	0
kinship	k	3	0	58	2	0	1	0
kitten	k	5	7	126	2	0	1	0
kosher	k	1	1	24	2	0	1	0
all	filler	3001	25	2628	1	1	1	0
art	filler	208	13	3006	1	1	1	0
else	filler	176	6	4875	1	1	1	0
look	filler	399	17	2371	1	1	1	0
nose	filler	60	18	1454	1	1	1	0

word	Phoneme	K&F	Denity	Celex	Syll #	Hi-Lo	Read	Listen
now	filler	1314	16	1290	1	1	1	0
risk	filler	54	9	1442	1	1	1	0
rock	filler	75	23	2081	1	1	1	0
use	filler	589	13	4151	1	1	1	0
wish	filler	110	13	3373	1	1	1	0
account	filler	117	2	2364	2	1	1	0
approach	filler	123	0	1597	2	1	1	0
explain	filler	64	0	3606	2	1	1	0
increase	filler	195	0	4140	2	1	1	0
marriage	filler	95	2	1877	2	1	1	0
neither	filler	141	2	1120	2	1	1	0
only	filler	1747	0	2518	2	1	1	0
united	filler	482	0	3358	2	1	1	0
welfare	filler	53	0	1169	2	1	1	0
yourself	filler	67	1	2437	2	1	1	0
attitude	filler	107	3	2246	3	1	1	0
election	filler	77	2	1620	3	1	1	0
equipment	filler	167	0	1352	3	1	1	0
interest	filler	330	1	5150	3	1	1	0
memory	filler	76	2	1584	3	1	1	0
natural	filler	156	2	2814	3	1	1	0
occasion	filler	58	1	1877	3	1	1	0
recognize	filler	62	0	2467	3	1	1	0
religion	filler	119	1	1030	3	1	1	0
wonderful	filler	53	0	1048	3	1	1	0
awe	filler	5	20	148	1	0	0	1
itch	filler	5	18	75	1	0	0	1
lark	filler	2	11	69	1	0	0	1
lobe	filler	3	15	69	1	0	0	1
mend	filler	2	18	151	1	0	0	1
muff	filler	1	18	15	1	0	0	1
nap	filler	4	20	98	1	0	0	1
noun	filler	1	8	29	1	0	0	1
ooze	filler	2	10	68	1	0	0	1
owl	filler	2	15	128	1	0	0	1
adorn	filler	1	1	75	2	0	0	1
eject	filler	1	3	60	2	0	0	1
evade	filler	1	1	135	2	0	0	1
irate	filler	1	9	20	2	0	0	1
lily	filler	1	5	206	2	0	0	1
melon	filler	1	1	61	2	0	0	1
neuter	filler	1	4	32	2	0	0	1
oblong	filler	1	0	28	2	0	0	1

word	Phoneme	K&F	Denity	Celex	Syll #	Hi-Lo	Read	Listen
outgrowth	filler	1	1	9	2	0	0	1
unrest	filler	5	0	94	2	0	0	1
allergy	filler	1	1	68	3	0	0	1
audition	filler	3	2	38	3	0	0	1
emerald	filler	3	0	31	3	0	0	1
immoral	filler	5	2	94	3	0	0	1
irrigate	filler	1	2	38	3	0	0	1
lavender	filler	5	0	28	3	0	0	1
lullaby	filler	5	0	18	3	0	0	1
Norwegian	filler	2	0	69	3	0	0	1
nominate	filler	3	0	96	3	0	0	1
numeral	filler	1	0	25	3	0	0	1
elm	filler	3	7	176	1	0	0	1
mop	filler	3	16	102	1	0	0	1
niche	filler	3	16	83	1	0	0	1
waltz	filler	1	3	27	1	0	0	1
yawn	filler	2	12	135	1	0	0	1
elector	filler	2	1	61	3	0	0	1
indulgence	filler	5	0	84	3	0	0	1
librarian	filler	5	0	90	3	0	0	1
reassure	filler	1	0	272	3	0	0	1
octopus	filler	1	0	27	3	0	0	1

Appendix B: Japanese Devoicing Test and Filler Words with Lexical Frequency

word	Japanese	vowel	freq	frequency	listening
fukaku	不覚	u	0	167	1
fukutsu	不屈	u	0	61	1
Futsukayoi	二日酔い	u	0	212	1
ikusa	戦(いくさ)	u	0	46	1
kusai	臭(くさ)い	u	0	307	1
kushakusha	くしゃくしゃ	u	0	6	1
kutakuta	くたくた	u	0	101	1
shukushaku	縮尺(しゅくしゃく)	u	0	100	1
taishokuteate	退職手当	u	0	215	1
yakusou	薬草	u	0	287	1
fukuro	袋	u	1	5924	1
fukusuu	複数	u	1	13402	1
fusai	負債（ふさい）	u	1	3789	1
kaku	書く	u	1	47254	1
kokusai	国際	u	1	71948	1
kosumosu	コスモス	u	1	1407	1
kufuu	工夫	u	1	9924	1
kusuri	薬	u	1	9715	1
supai	スパイ	u	1	3786	1
zokushutsu	続出	u	1	3309	1
kosokoso	コソコソ	o	0	56	1
kosui	湖水	o	0	162	1
sokujitsu	即日	o	0	618	1
kokoro	心	o	1	1493	1
sokai	疎開	o	1	1293	1
sekoi	せこい	e	0	34	1
setsubun	節分	e	0	185	1
tekagami	手鏡	e	0	50	1
sekaiichi	世界一	e	1	3419	1
tehon	手本	e	1	1385	1
asuka	飛鳥	u	0	242	0
fuhou	訃報	u	0	262	0
fuka	不可	u	0	375	0
futa	蓋（ふた）	u	0	165	0
gakushoku	学食	u	0	23	0
hakusen	白線	u	0	131	0
hinshuku	ひんしゅく	u	0	206	0
kokuho	国保	u	0	380	0

word	Japanese	vowel	freq	frequency	listening
sokutou	即答	u	0	384	0
tsutsuku	つつく	u	0	223	0
fuhai	腐敗	u	1	6396	0
futsuu	普通	u	1	6280	0
gakusei	学生	u	1	42414	0
kashu	歌手	u	1	6788	0
keisatsu	警察	u	1	21274	0
kokuso	告訴	u	1	2308	0
kutsu	靴	u	1	4009	0
shukushou	縮小	u	1	11665	0
susumu	進む	u	1	62098	0
tsukusu	尽くす	u	1	6058	0
fukushikikokyuu	腹式呼吸	i, u,	0	27	0
kashitsuchishi	過失致死	i, u,	0	53	0
kikikaesu	聞き返す	i, u,	0	199	0
kishukusha	寄宿舎	i, u,	0	216	0
kuchikiki	口利き	i, u,	0	331	0
kushikatsu	串カツ	i, u,	0	2	0
shikitsumeru	敷き詰める	i, u,	0	260	0
sukihoudai	好き放題	i, u,	0	18	0
sukitouru	透き通る	i, u,	0	255	0
tsukisasu	突き刺す	i, u,	0	274	0
chikuseki	蓄積	i, u,	1	3386	0
fukushi	福祉	i, u,	1	20555	0
kashitsuke	貸し付け	i, u,	1	1971	0
kifukin	寄付金	i, u,	1	1781	0
kikuchi-san	菊池さん	i, u,	1	1531	0
kousakukikai	工作機械	i, u,	1	1413	0
shishutsu	支出	i, u,	1	12372	0
shouhishishutsu	消費支出	i, u,	1	1007	0
sutoraiki	ストライキ	i, u,	1	1571	0
tekishutsu	摘出	i, u,	1	1583	0
akisame	秋雨(あきさめ)	i	0	167	0
chichikata	父方	i	0	155	0
hikkoshi	引越し	i	0	21	0
hisan	悲惨	i	0	335	0
kikikomi	聞き込み	i	0	415	0
shikeru	湿気る（しける）	i	0	9	0
shikikin	敷金	i	0	359	0
shisai	司祭	i	0	423	0
shishimai	獅子舞	i	0	57	0
shiso	紫蘇(しそ)	i	0	187	0
asahi	朝日	i	1	11467	0

word	Japanese	vowel	freq	frequency	listening
denshikiki	電子機器	i	1	n/a	0
kashi	歌詞	i	1	1709	0
kihon	基本	i	1	25183	0
kisha	記者	i	1	36618	0
koshihikari	コシヒカリ	i	1	1195	0
shihai	支配	i	1	15913	0
shikai	歯科医	i	1	1041	0
shikichi	敷地	i	1	4686	0
shikisha	指揮者	i	1	2020	0
arayuru	あらゆる	filler	1	8275	1
au	合う	filler	1	8475	1
gimu	義務	filler	1	10640	1
ginkou	銀行	filler	1	61818	1
gyuunyuu	牛乳	filler	1	3157	1
inai	以内	filler	1	16275	1
inu	犬	filler	1	4207	1
iryou	医療	filler	1	35332	1
moyou	模様	filler	1	12574	1
nagare	流れ	filler	1	21680	1
naiyou	内容	filler	1	65992	1
nanmin	難民	filler	1	14708	1
nemuru	眠る	filler	1	3803	1
nougyou	農業	filler	1	27488	1
nouzei	納税	filler	1	5238	1
ondan	温暖	filler	1	5506	1
piano	ピアノ	filler	1	5447	1
renmei	連盟	filler	1	15563	1
seiji	政治	filler	1	91629	1
uriage	売り上げ	filler	1	11914	1

References

Abercrombie, D. (1967). *Elements of general phonetics.* Chicago: Aldine.

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics*, *63*, 798-810.

Allen, J. S. & Miller, J. L. (2004). Listener sensitivity to individual talker difference in voice-onset-time. *Journal of the Acoustic Society of America*, *115,* 3171-3183.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustic Society of America, 113,* 544–552.

Amano, S., Kondo, T., & Kato, K. (1999). Familiarity effect on spoken word recognition in Japanese. *Proceedings of the 14th ICPhS, 2,* 873-876.

Amano, S., & Kondo, T. (2003). Nihongo-no Goi-Tokusei (Lexical properties of Japanese) Vol. 1-7, CD-ROM version, Sanseido, Tokyo.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [distributor].

Baese, M. M. & Goldrick, M. (Under review). Mechanisms of interaction in speech production.

Beckman, M. E. (1982). Segment duration and 'mora' in Japanese. *Phonetica, 39,* 113-135.

Beckman, M. E. (1994). When is a syllable not a syllable? *Ohio State Working Papers in Linguistics No. 44, Papers from the Linguistics Laboratory,* 50-69. Columbus, OH: The Ohio State University.

Beckman, M. E. & Shoji, A. (1984). Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese, *Phonetica, 41,* 61-71.

Belin, P. & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport, 14,* 2105–2109.

Benkí, J. R. (2003). Analysis of English Nonsense Syllable Recognition in Noise. *Phonetica, 60*, 129–157.

Boersma, P. & Weenink, D. (2007). Praat: doing phonetics by computer (Version 4.5.25) [Computer program]. Retrieved May 7, 2007, from http://www.praat.org/

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6,* 201-251.

Carlson, K., German, J., & Pierrehumbert, J. B. (Under revision). Reassignment of the flap allophone in rapid dialect adaptation. *Journal of Phonetics*.

Cho, T., & Ladefoged, P. (1999). Variations and universals in VOT: Evidence from 18 languages. *Journal of Phonetics 27*, 207-229.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Church, B. A., and Schacter, D. L. (1994). Perceptual Specificity of Auditory Priming: Implicit Memory for Voice Intonation and Fundamental Frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20 (3)*, 521-533.

Coetzee, A. W. & Pater, J. (2006). Lexically ranked OCP-Place constraints in Muna, Ms., University of Michigan and University of Massachusetts, Amherst, ROA-842.

Coetzee, A. W. & Pater, J. (to appear). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory.*

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. Behavior Research Methods, Instruments, & Computers, *25*, 257-271.

Coleman, J. (2002). Phonetic representations in the mental lexicon. In J. Durand, & B. Laks (Eds.), *Phonetics, phonology and cognition* (pp. 96–130). Oxford: Oxford University Press.

Dauer, R. M. (1980). The reduction of unstressed high vowels in modern Greek. *Journal of International Phonetic Association, 10,* 17-27.

Davidson, L. (2006). Schwa elision in fast speech: Segmental deletion or gestrual overlap? *Phonetica, 63*, 79- 112.

Delvaux, V. & Soquet, A. (2007). The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica, 64,* 145-173.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67,* 224–238.

Ernestus, M., Lahey, M., Verhees, F., & Baayen, R. H. (2006). Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America, 120,* 1040–1051.

Fidelholz, J. (1975). “Word frequency and vowel reduction in English,” in R. E. Grossman, L. J. San, and T. J. Vance (Eds), *Papers from the 11th Regional Meeting of the Chicago Linguistic Society*, *11*, 200–213.

Flemming, E. (2001). ‘Scalar and Categorical Phenomena in a Unified Model of Phonetics and Phonology’, *Phonology, 18*, 7–44.

Fowler, C. A. (1994). Speech perception: Direct realist theory. *Encyclopedia of language and linguistics, Vol. 8*, 4199-4203. Oxford: Pergamon Press.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Languages*, *49*, 396-413.

Frisch, S. (1996). *Similarity and Frequency in Phonology*. Doctoral dissertation, Northwestern University.

Frisch, S., Pierrehumbert, J. B., & Broe M. (2004). Similarity Avoidance and the OCP. *Natural Language and Linguistic Theory, 22,* 179-228.

Fromkin, V. (Ed.). (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Frye, R. E., McGraw Fisher, J., Coty, A., Zarella, M., Liederman, J. & Halgren E. (2007). Linear Coding of Voice Onset Time. *Journal of Cognitive Neuroscience, 19 (9),* 1476-1487.

Fujimoto, M. (2004). Effects of consonant type and syllable position within a word on vowel devoicing in Japanese. *Proceedings of Speech Prosody 2006,* 625-628.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance, 6 (1)*, 110-125.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of Accommodation: Developments in Applied Sociolinguistics.* Cambridge University Press, Cambridge.

Goh, W. D., & Pisoni, D. B. (1998). Effects of lexical neighbourhoods on immediate memory span for spoken words: A first report. *Research on spoken language processing Progress Report. No. 22.* Bloomington, IN: Indiana University, Department of Psychology, Speech Research Laboratory.

Goldinger, S. D. (1996). Words and voices: Episodic Traces in Spoken Word Identification and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166-1183.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson and J.W. Mullennix (eds.) *Talker Variability in Speech Processing*. San Diego: Academic Press, 33-66.

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, *105*, 251-279.

Goldinger, S. D. (2000). The role of perceptual episodes in lexical processing. In Anne Cutler, James M. McQueen and Rian Zondervan (eds.), *Proceedings of SWAP Spoken Word Access Processes.* Nijmegen: Max-Planck Institute for Psycholinguistics, 155-159.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, 49-54.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 1211-1238.

Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language*, *51*, 586-603.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition 103*, 386–412

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1–51.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception.* New York: Academic Press, 187–294.

Grossberg, S. (2003). The resonant dynamics of speech perception. *Journal of Phonetics, 31,* 423-45.

Günzburger, D. (1991). Breathiness in male and female speakers. *Proceedings of the 12th International Congress of Phonetic Sciences, vol. 3,* 62-65.

Halle, M. (1985). Speculation about the representation of words in memory. In V. Fromkin (Ed.), *Phonetic Linguistics*. New York: Academic Press, 101-114.

Halle, M. & Stevens, K.N. (1971). A note on laryngeal features. *MIT Quarterly Progress Report, 101*, 198–212.

Han, M. S. (1962). "Unvoicing of Vowels in Japanese," *Onsei no Kenkyuu 10*, 81–100.

Harrington, J., Palethorpe, S., & Watson, C. I. (2000). Does the Queen speak the Queen's English? *Nature, 408,* 927–928.

Hay, J. & Baayen, H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences, 9*, 342–348.

Hayes, B., & Steriade, D. (2004). Introduction: the phonetic basis of phonological markedness. In Hayes, B., Kirchner, R., Steriade, D. (Eds.), Phonetically-Based Phonology. Cambridge University Press, Cambridge, pp. 1–33.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.3.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411-428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528-551.

Howes, D. H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, *29,* 296-305.

Imaizumi, S., Hayashi, A., & Deguchi, T. (1995). Listener adaptive characteristics of vowel devoicing in Japanese Dialogue. *J. Acoust. Soc. Amer. 98(2),* 768-778.

Imaizumi, S., Fuwa, K., & Hosoi, H. (1999). "Development of adaptive phonetic gestures in children: Evidence from vowel devoicing in two different dialects of Japanese." *Journal of the Acoustical Society of America, 106,* 1033-1044.

Ito, M. (2004). Politeness and voice quality – The alternative method to measure aspiration noise, Proc. *Speech Prosody 2004*, 213-216.

Jacoby, L.L., & Hayman, C. (1987). Specific visual transfer in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 456-463.

Jescheniak, J.D., & Levelt, W.J.M. (1994). Word frequency effects inspeech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 824–843.

Johnson, K. (1997). Speech perception without speaker normalization: an exemplar model. In K. Johnson and J.W. Mullennix (eds.) *Talker Variability in Speech Processing*. San Diego: Academic Press,145- 66.

Jun, S. & Beckman, M. E. (1993). A Gestural-overlap Analysis of Vowel Devoicing in Japanese and Korean. Paper presented at the 1993 Annual Meeting of the LSA, Los Angeles, 7-10 January, 1993.

Jun, S. & Beckman, M. E. (1994). "Distribution of Devoiced High Vowels in Korean," in the *Proceedings of the International Conference on Spoken Language Processing*, *Vol. 2,* 479-482

Jurafsky, D., Bell, A., Gregory, M. & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in Lexical Production. In J. Bybee and P. Hopper (eds.) *Frequency and the emergence of linguistic structure*. John Benjamins, Amsterdam, 229–254.

Jusczyk, P.W., Luce, P. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630–645.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language, 60,* 286-319.

Keating, P.A. & Huffman, M. K. (1984). Vowel Variation in Japanese. *Phonetica, 41*, 191-207.

Kitahara, M. (1999). Accentual contrast maintenance and vowel devoicing in Tokyo Japanese. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, 2363–2366, San Francisco, CA.

Kitahara, M. (2001). *Category structure and function of pitch accent in Tokyo Japanese*. Doctoral dissertation, Indiana University.

Klatt, D.H. & Klatt, L.C. (1990). Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America*, *87*, 820–856.

Kohler, K. J. (1990). Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In: W. J. Hardcastle and A. Marchal (eds.): Speech Production and Speech Modelling. Dordrecht: Kluver, 69-92.

Kondo, M. (1997). *Mechanisms of Vowel Devoicing in Japanese.* Doctoral dissertation, University of Edinburgh.

Kondo, M. (2005). Syllable structure and its acoustic effects on vowels in devoicing environments. In van de Weijer, J., Najo, K., Nishhara, T. (eds), *Voicing in Japanes,* Mouton de Gruyter, 229-245.

Kraljic, T. & Samuel, A.G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13,* 262–268.

Kreiman, J. & Van Lancker, D. (1988). Hemispheric specialization for voice recognition: Evidence from dichotic listening. *Brain & Language, 34,* 246-252.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling.* London: Sage University Press.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present day American English.* Providence, RI: Brown University Press.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606-608.

Kuriyagawa, F., & Sawashima, M. (1989). Word accent, devoicing and duration of vowels in Japanese. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, University of Tokyo, 23*, 85-108.

Labov, W. A. (1972). *Language in the Inner City: The Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, W. A. (1994). Principles of linguistic change. Vol. 1: *Internal factors*. Oxford & Cambridge, Mass.: Blackwell.

Ladefoged, P. (2000). *A course in phonetics (5th edition).* San Diego: Harcourt, Brace and Jovanovich.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: The MIT Press.

Liberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition, 21,* 1-36.

Lin, Y. (2005). *Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition.* Doctoral dissertation, UCLA.

Lisker, L. (1986). Voicing in English: A Catalogue of Acoustic Features Signaling /b/ versus /p/ in Trochees. *Language and Speech, 29,* 3-11.

Lisker, L. & Abramson, A. (1964). Cross-language study of voicing in initial stops: acoustical measurements. *Word, 20*, 384-422.

Lubowicz, A. (2003). *Contrast Preservation in Phonological Mappings.* Doctoral dissertation. University of Massachusetts, Amherst. Amherst: GLSA Publications.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon.* Doctoral dissertation, Indiana University, Bloomington, IN.

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

Luce, P. A., McLennan, C., & Charles-Luce, J. (2003). Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In J. Bowers & C. Marsolek (Eds.), *Rethinking implicit memory.* Oxford: Oxford University Press, 197-214.

Maekawa, K. & Kikuchi, H. (2005). Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In van de Weijer, J., Najo, K., Nishhara, T. (eds), *Voicing in Japanese*, Mouton de Gruyter, 205-228.

Martin, S. E. (1952). *Morphophonemics of Standard Colloquial Japanese*, supplement to *Language, Language Dissertation No. 47*.

Massaro, D. W. (1974). Perceptual units in speech perception. *Journal of Experimental Psychology, 102*, 199-208.

Maye. J., Aslin, R.N. & Tanenhaus, M.K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32,* 543–562.

McCawley, J. D. (1968). *The Phonological Component of a Grammar of Japanese*. Mouton, The Hague.

McLennan, C. T., Luce, P.A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 539-553.

McLennan, C. & Luce, P. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 306–321.

Miller, G. and Nicely, P. (1955). An Analysis of Perceptual Among Some English Consonants. *Journal of the Acoustical Society of America, 27*, 338-352.

Miller, J.L. (2001). Mapping from acoustic signal to phonetic category: Internal structure, context effects and speeded categorization. *Language and Cognitive Processes, 16,* 683-690.

Morris, M. Y. (2005). Devoicing and Its Environments in Perception: Kinki Japanese, or Tokyo? *Proceedings of the 30th annual meeting of the Berkeley Linguistics Society,* 280-291.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effect of talker variability on spoken word recognition. *Journal of the Acoustic Society of America*, *85*, 365-378.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception, *Journal of Language and Social Psychology*. *21*, 422–432.

Norman, K., & O'Reilly, R. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning-systems approach. *Psychological Review, 110,* 611–646.

Norris, D. G. (1984). The effects of frequency, repetition con and stimulus quality in visual word recognition. *CogQuarterly Journal of Experimental Psychology, 36A,* 507–518.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23(3)*, 299-370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47 (2)*, 204-238.

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology, General, 115*, 39-57.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report*, No. 10. Bloomington: Indiana University, Psychology Department, Speech Research Laboratory.

Ogasawara, N. (2007). *Processing of Speech Variability: Vowel reduction in Japanese.* Doctoral dissertation, University of Arizona.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119,* 2382–2393.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., & Poeppel, D. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience, 12 (6),* 1038–1055.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J and P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure*. John Benjamins, Amsterdam, 137-157.

Pierrehumbert, J. B. (2002). Word-Specific Phonetics. In C. Gussenhoven and N. Warner (eds.) *Laboratory Phonology VII*, Mouton de Gruyter, Berlin. 101-140.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullenix (Eds.), *Talker Variability in Speech Processing.* San Diego, CA: Academic Press, 9-32.

Prince, A. & Smolensky, P. (1993). Optimality Theory: Constraint Interaction in Generative Grammar. (ms). Rutgers University and University of Colorado, Boulder.

Ryalls, J., Zipprer, A., & Baldauff, P. (1997). ‘‘A preliminary investigation of the effects of gender and race on voice onset time,’’ *Journal of Speech, Language, and Hearing Research*, *40 (3),* 642–645.

Saffran, J.R., Aslin R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Sancier, M. & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25,* 421-436.

Samuel, A.G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110,* 474-494.

Scarborough, D. L., Cortese, C., & Scarborough, H. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 1–17.

Scarborough, R. A. (2004). *Coarticulation and the Structure of the Lexicon.* Doctoral dissertation, UCLA.

Semon, R. (1923). *Mnemic Psychology* . (B. Duffy, Trans.) London: George Allen & Unwin. (Original publication: 1909).

Shacter, D., Eich, J., & Tulving, E. (1978). Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior, 17,* 721-743.

Shepard, R. (1980). Multidimensional scaling, tree-fitting and clustering. *Science, 210,* 390-398.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66 (3),* 422-429.

Steriade, D., (2001). The phonology of perceptibility effects: the P-map and its consequences for constraint organization. Ms. UCLA.

Sugito, M. (1988). Utterance of voiceless vowels by speakers in eight cities in Japan, *Annual Reports (Osaka Shoin Women's College, Japan), 25*, 1–10.

Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills, 75,* 983–992.

Takeda, K. & Kuwabara, H. (1987). Analysis and Prediction of Devocalizing Phenomena (in Japanese). *Proceedings of Acoustic Society of Japan, 3-3-8,* 105-106.

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin and Review 2,* 339-363.

Tsuchida, A. (1997). *Phonetics and Phonology of Japanese Vowel Devoicing.* Doctoral dissertation, Cornell University.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory,* New York: Academic Press, 381-403.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford: Clarendon Press.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual review of psychology, 53,* 1-25.

Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany, N.Y.: State University of New York Press.

Varden, J. K. (1998). *On High Vowel Devoicing in Standard Modern Japanese: Implications for Current Phonological Theory*. Doctoral dissertation, University of Washington.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 735–747.

Vitevich, M., & Luce, P. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*, 325–329.

Vitevitch, M. & Luce, P.A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374-408.

Warren, R. M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin, 96*, 371-383.

Washington University in St. Louis Speech and Hearing Lab Neighborhood Database as a web-based implementation of the 20 000-word Hoosier Mental Lexicon. http://neighborhoodsearch.wustl.edu/Neighborhood/NeighborHome.asp

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The Linguistic Review, 23*, 247-274.

Whiteside, S.P., Henry, L., & Dobbin, R. (2004). Sex differences in voice onset time: a developmental study of phonetic context effects in British English. *Journal of the Acoustic Society of America*, *116 (2)*, 1179-1183.

Wilson, C. (2006). Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science, 30 (5)*, 945- 982.

Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. *Research on Spoken Language Processing Progress Report No. 21*. Bloomington, IN: Speech Research Laboratory.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, J., R. Ogden, and R. Temple (eds.). *Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press.

Zue, V. (1980). Acoustic characteristics of stop consonants: A controlled study. Bloomington, IN: Indiana Linguistics Club.