THE INTERACTION OF LANGUAGE PROFICIENCY AND TALKER VARIABILITY IN LEARNING

by

Andréa Katharine Davis

Copyright © Andréa Katharine Davis 2015

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2015
As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Andréa Katharine Davis entitled The Interaction of Language Proficiency and Talker Variability in Learning and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 10 April 2015
Keith Johnson

Date: 10 April 2015
Andrew Lotto

Date: 10 April 2015
Diane Ohala

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College. I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 10 April 2015
Dissertation Director: LouAnn Gerken
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED:  Andréa Katharine Davis
The makings of a dissertation is allegedly a solitary experience, by I have never found it to be so. One of the greatest pleasures of doing the research for this dissertation has been meeting, being advised by, and being supported by incredibly smart, interesting, and giving people. I am deeply grateful:

My advisor: LouAnn Gerken, for your suggestions and inspiration throughout the process of designing, running, and analyzing my experiments, and in writing and theorizing. You have made me a better experimenter, writer, and thinker.

My committee: Keith Johnson, for your comments and suggestions and support while at Berkeley; Andrew Lotto for inspiring me with discussions about speech perception; Diane Ohala, for your thoughts on my experiment with children and general encouragement.

For helping me to run my experiments: Lauren Clough, Michelle Guernon, Sara Karabashlieva, Sara Knight, ChenTsung Yang, thank you! I could not have collected the data without your help.

Jane Tsay, for hosting me in your lab and for your suggestions on my Taiwanese experiment.

Yaya Tsai, for welcoming me into your house in Taiwan, and Joanne Chang, for taking care of me in Taiwan.

Kara Hawthorne and Carolyn Quam, for comments on experimental design, general moral support, and stress-relieving lab yoga and floor session.

Jessamyn Schertz, for much patient answering of my phonetics questions, and sharing of tea and M&M’s in the phonetics lab.

For further comments and discussion: Sarah Bakst, Emily Cibelli, Clara Cohen, Shinae Kang, Stephanie Shih (also for baked goods, i.e. brain sustenance).

My parents, for your unending support, even before you understood what this dissertation was about, and for then taking the trouble to understand.

Siddharth Jain, for asking me to not make this dissertation too long, so that you wouldn’t have to read quite so much and thus reminding me to make it readable.

Finally, thank you to the Arizona Linguistics Department and the Berkeley Linguistics Department. These past 5 years have been thoroughly thoroughly enjoyable, made possible by the supportive, encouraging environment that exists in both departments, in droves.

Research was supported by NSF Grant #0950601, NSF Doctoral Dissertation Research Improvement Grant #BCS-1348451, and NSF East Asia Pacific Summer Institute Fellowship.
TABLE OF CONTENTS

LIST OF FIGURES ................................................................. 8
LIST OF TABLES ................................................................. 9
ABSTRACT ............................................................ 10

CHAPTER 1 Introduction ....................................................... 11
1.1 The central question ................................................. 12
1.2 Learning to understand seems difficult, yet human performance is remarkably good ................................... 14
1.3 What is helpful about multiple talkers? .......................... 16
1.4 A secondary question: the relationship between perception and production ................................................. 19
1.5 Summary and explanation of terminology ......................... 20
1.5.1 Terminology ...................................................... 22

CHAPTER 2 An Overview of Previous Research on Learning from Variability ........................................... 24
2.1 Introduction ............................................................ 25
2.2 The importance of phonology ......................................... 25
2.2.1 The role of phonemes and phonology ....................... 26
2.2.2 First language learners and second language phonological learning .................................................. 31
2.3 Finding consistency in an inconsistent world ...................... 33
2.3.1 The search for invariance ..................................... 34
2.3.2 Speech categories: invariant and innate or variant and learned? .................................................... 37
2.4 Variation promotes generalization .................................... 43
2.4.1 Perception ....................................................... 44
2.4.2 Production ....................................................... 51
2.5 Word form learning ..................................................... 53
2.5.1 Infants and young children .................................... 53
2.5.2 Native speakers and second language learners ............ 57
2.5.3 Words as sound categories ..................................... 60
2.6 Is variation always useful? ............................................ 61
2.6.1 Learning to cope with variation .............................. 62
2.6.2 Continuing to cope: adults also have difficulties .......... 64
2.7 Situating the present work in context .............................. 68
### TABLE OF CONTENTS – Continued

**CHAPTER 3** Experiment 1: English-Acquiring Children ........................................... 70

3.1 Introduction ........................................................................................................... 71

3.2 Experiment 1a: Perception .................................................................................. 71

3.2.1 Participants .................................................................................................... 71

3.2.2 Stimuli .......................................................................................................... 72

3.2.3 Procedure .................................................................................................... 72

3.2.4 Hypotheses .................................................................................................. 75

3.2.5 Results ......................................................................................................... 75

3.2.6 Discussion ................................................................................................... 79

3.3 Experiment 1b: Production ................................................................................ 79

3.3.1 Participants ................................................................................................... 79

3.3.2 Stimuli .......................................................................................................... 80

3.3.3 Procedure .................................................................................................... 80

3.3.4 Hypotheses .................................................................................................. 81

3.3.5 Results ......................................................................................................... 82

3.3.6 Discussion ................................................................................................... 83

3.4 General Discussion for Experiment 1a and 1b ................................................... 84

3.5 Pilot experiment: a second test of perception ..................................................... 85

3.5.1 Participants ................................................................................................... 86

3.5.2 Stimuli .......................................................................................................... 86

3.5.3 Procedure .................................................................................................... 86

3.5.4 Results and discussion ................................................................................ 88

3.6 General Discussion of Results .......................................................................... 89

3.6.1 Suggestion for another experiment ............................................................... 90

**CHAPTER 4** Experiment 2: Native English-Speaking Adults ............................... 92

4.1 Introduction ........................................................................................................... 93

4.2 Experiment 2a: Perception ................................................................................ 94

4.2.1 Participants ................................................................................................... 94

4.2.2 Stimuli .......................................................................................................... 94

4.2.3 Procedure .................................................................................................... 95

4.2.4 Hypotheses .................................................................................................. 97

4.2.5 Results ......................................................................................................... 97

4.2.6 Discussion ................................................................................................... 100

4.3 Experiment 2b: Production ................................................................................ 100

4.3.1 Participants ................................................................................................... 100

4.3.2 Stimuli .......................................................................................................... 101

4.3.3 Procedure .................................................................................................... 101

4.3.4 Results ......................................................................................................... 101
TABLE OF CONTENTS – Continued

4.3.5 Discussion .................................................. 103
4.4 General Discussion ............................................. 103

CHAPTER 5  Experiment 3: Second Language English Learners ........... 105
  5.1 Introduction ................................................... 106
  5.2 Experiment 3a: Perception ..................................... 106
    5.2.1 Participants .............................................. 106
    5.2.2 Stimuli .................................................. 107
    5.2.3 Procedure ............................................... 108
    5.2.4 Determining English proficiency .......................... 110
    5.2.5 Hypotheses .............................................. 111
    5.2.6 Results .................................................. 112
    5.2.7 Discussion ............................................... 116
  5.3 Experiment 3b: Production .................................... 117
    5.3.1 Participants .............................................. 117
    5.3.2 Stimuli .................................................. 117
    5.3.3 Procedure ............................................... 117
    5.3.4 Hypotheses .............................................. 118
    5.3.5 Results .................................................. 119
    5.3.6 Discussion ............................................... 121
  5.4 General discussion ........................................... 122

CHAPTER 6  Conclusion ................................................ 124
  6.1 Summary of main findings ..................................... 125
  6.2 Perception and production: resolving an inconsistency ............. 128
  6.3 Implications to theories of speech representation .................. 131
    6.3.1 If phonemes exist, when are they used? Perhaps for learning 134
  6.4 Final words .................................................. 135

REFERENCES ....................................................... 138
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Mispronunciation detection accuracy was not higher in the multiple talker condition</td>
<td>76</td>
</tr>
<tr>
<td>3.2</td>
<td>Mispronunciation detection accuracy was not higher in the multiple talker condition</td>
<td>78</td>
</tr>
<tr>
<td>3.3</td>
<td>Fewer pronunciation errors were made when trained with multiple talkers</td>
<td>82</td>
</tr>
<tr>
<td>3.4</td>
<td>Performance was about chance</td>
<td>88</td>
</tr>
<tr>
<td>4.1</td>
<td>Accuracy was not higher in the multiple talker condition</td>
<td>98</td>
</tr>
<tr>
<td>4.2</td>
<td>Reaction time not predicted by training condition</td>
<td>99</td>
</tr>
<tr>
<td>4.3</td>
<td>Time to production onset not faster in the multiple talker condition</td>
<td>102</td>
</tr>
<tr>
<td>5.1</td>
<td>Only less experienced learners benefited from multiple talkers in perceptual accuracy</td>
<td>113</td>
</tr>
<tr>
<td>5.2</td>
<td>A comparison of L1 and L2 speakers</td>
<td>115</td>
</tr>
<tr>
<td>5.3</td>
<td>Second language learners differed in production accuracy</td>
<td>120</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1.1 Does variability help? .................................................. 21
2.1 Theories of speech perception in the face of high variability .... 43
5.1 Participants in Experiment 3 ........................................... 107
6.1 Does variability help? .................................................. 127
Previous studies have shown that multiple talkers help learners make more robust word representations when the learner is not very experienced with the language (Richtsmeier et al., 2009; Rost & McMurray, 2009, 2010). This is likely because exposure to variation allows the learner to observe which acoustic dimensions vary unpredictably across talkers, and which acoustic dimensions vary predictably. However, this predicts that only learners who are less experienced with a language will benefit from multiple talkers, as more experienced learners should be able to use their previous knowledge about the language’s speech sounds.

Three word-learning experiments, with participants who were expected to have different levels of experience in the language, were performed to test this prediction. In the first experiment, English-acquiring children did benefit from multiple talker in the production but not perception of newly learned words. In the second experiment, native English-speaking adults did not benefit from learning from multiple talkers in either the perception or production of new words. Finally, second language-learning adults benefited from multiple talkers if they were less proficient speakers, but not if they were more proficient. Collectively, these results suggest that learning from multiple talkers is only beneficial for less experienced language learners.
CHAPTER 1

Introduction
1.1 The central question

Learning to speak and to understand speech is one of the greatest accomplishments of our early life. The complexity of the task is often not appreciated until one attempts to study a second language, or to understand how it is that infants and young children accomplish it. While we hear even before birth, we must learn to understand.

One difficulty in learning to understand, is that speech is inherently variable. A well-studied source of that variability is talker variability: different talkers pronounce words differently. An infant may learn a word from one speaker, but not be able to recognize that word when it is pronounced by a different speaker. Learning from multiple talkers seems to help infants and young children with this task (Houston, 2000). In particular, learning from multiple talkers helps learners with both the perception and production of new speech sounds, as well as new, similar-sounding words. Rost & McMurray (2009, 2010) found that infants learned to pair similar-sound nonce words with new referents when they learned from multiple talkers, but not when they learned from a single talker. Richtsmeier et al. (2009) found that young children produced words more accurately when they learned the words from multiple talkers.

But the benefit of multiple talkers is not limited to infants and children. Adults learning foreign sounds also benefit from multiple talkers. In one of the earliest studies, Lively et al. (1993) found that monolingual Japanese speakers were able to discriminate English [i] from [l] in new speakers only when trained on the distinction with multiple talkers, but not when trained with a single talker. Kingston (2003)
likewise found that American English listeners were better able to identify German vowels differing in tenseness when trained with multiple talkers. Holt & Lotto (2006) were able to explicitly manipulate two acoustic cues for a sound category, where only one acoustic cue could be reliably used to distinguish between the two sound categories. They showed that high variability of one cue and low variation in a second cue caused learners to weight the second cue more highly in distinguishing the categories. These studies suggest that high variation of the less reliable cues and low variation of the more reliable cues helps learners to know which parts of the acoustic signal reliably distinguish between sounds.

Assuming this explanation for the benefit of variation is correct, a question arises. What happens when a learner is already very familiar with the sound system of her language? An experienced learner already knows which acoustic cues are reliable. If talker variability is useful mainly for helping learners determine the relevant acoustic cues for a language, only learners who are less experienced in a language should benefit from high variability learning. However, if being exposed to multiple talkers serves some other function in the experimental tasks in which it has been used, all learners should continue to benefit, regardless of their experience. Past studies have focused exclusively on inexperienced learners. In this dissertation, experienced learners will be compared with less experienced learners: native English adult learners, English-acquiring children, and higher and lower proficiency L2 learners. In all cases, both perception and production were tested.
1.2 Learning to understand seems difficult, yet human performance is remarkably good

As mentioned, one thing that makes the task of understanding speech so difficult is that the acoustics of a given speech sound vary. For example, the way a baby’s mother says the word *bear* is different from the way a baby’s father says the word, which is in turn different from the way the parents’ friends might say the word. This is what makes it possible to distinguish one speaker from another, but it makes identifying the word *bear* very difficult, because the way the word is realized changes across talkers.

Compounding this difficulty, is the fact that many words in a language sound very similar. The word *bear* is very close to the word *pear*, for example. Adult native speakers have no trouble distinguishing these two words across different talkers, but it is not actually a trivial ability. A language user must realize that the differing acoustics of *bear* and *pear* make two different words, but that the differing acoustics of one person’s pronunciation of *bear* and another person’s pronunciation do not make different words. A language learner must learn which parts of the acoustic signal vary, how much they vary, and what that variation means, in terms of differentiating one word from another. This seem like a Herculean task, but every normally developing child learns to do this, and also to produce understandable words, as well as perceive them.

Further, speech perception is remarkably flexible, and adjusting to new speakers is not the limit to its flexibility. Not only can adult native speakers adjust to different speakers, but they can perceive speech even when it is greatly distorted. Sine-wave
speech, which reduces the speech signal to the center frequencies of the first three formants, can be understood (Remez et al., 1981). Listeners with cochlear implants also learn to understand speech, despite the low frequency resolution and lack of harmonic detail (Friesen et al., 2001). Perhaps most pointedly, as noted by Holt & Lotto (2010), people understand phone conversations, which are between 300 and 3,000 Hz; yet, Lippmann (1996) found that people can also categorize consonants, with 90% accuracy, when the signal is filtered to contain information only below 800 and above 4,000 Hz. In other words, proficient speakers must adjust what information they use to categorize speech.

How speakers accommodate for the variation in speech and listening conditions has been much studied. In early work, speech sounds were thought of as being represented abstractly and categorically in a language user’s mind (McClelland & Elman, 1986; Best, 1995; Eimas et al., 1971). This makes intuitive sense, since, as previously mentioned, speech is perceived as constant, while the acoustics of speech are variable. The variability of the acoustics could only distract a listener’s understanding of speech. An adult native speaker thus would have learned to discard irrelevant information, such as talker identity, in her representation of a speech sound unit - a phoneme. Variation would then only be relevant to those still learning the language.

However, later work challenged this abstract, strictly categorical idea of a phoneme. Even for adult native speakers, speech perception is affected by the familiarity of a voice (Nygaard & Pisoni, 1998), and the ability to remember a word is also affected by the voice (Goldinger, 1996, 1998). It would seem that even adult
native speakers’ representations of speech includes variation.

It seems like a paradox that speech sounds could both be consistently divided into categories, and include the variation present within those categories. How can a listener recognize /ba/ as being the same as a different instance of /ba/, when the two are acoustically different? If there is in fact no invariance in a speech category, this is all the more surprising (Lindblom, 1996). Somehow, people are both sensitive to variation within a category, and yet treat that category as constant.

Because speech sound categories differ between languages, the ability to accommodate variation must be learned. Further, adult L2 learners must learn new sound categories, and thus must learn to accommodate variation in a different way in the second language.

1.3 What is helpful about multiple talkers?

It might seem that learning from multiple talkers would be a hindrance, rather than a help. Learning from multiple talkers increases acoustic variation, which means that the learner has to keep track of more than if she were learning from a single talker. And indeed, a variety of studies have found that changing the talker during a speech processing task makes language processing more difficult (see Chapter 2, Section 2.6 for further discussion). Yet, other studies find that learning from multiple talkers helps with recognizing a phoneme or word when it is pronounced by a new talker (Lively et al., 1993; Kingston, 2003; Rost & McMurray, 2009, 2010).

If the world were simple, and speakers did not vary in their pronunciations, learning from a single talker probably would be easier. However, a learner cannot
always listen to the same talker. As previously mentioned, she must be able
to identify a word across talkers - she must generalize to a new pronunciation.
Remembering this makes it seem more intuitive that a learner would benefit from
learning from multiple talkers. If the world is full of variation, it makes sense that
it is best to learn with variation. Learning in the environment that more closely
resembles the environment in which you will be tested improves performance.

But if this were the full explanation, then learning a new speech category from
multiple talkers should improve learning of the category for all learners. However,
learning from multiple talkers is often assumed to specifically help learners determine
the relevant acoustic cues for the new speech category they are learning. If you
listen to a new word from multiple talkers, then you have information about the
within category variation for that word. In contrast, listening to a new word
pronounced by only one talker, you lack that information. Further, if you learn
two, similar-sounding words from multiple talkers, you get information about both
within category variation, and also between category variation. You are essentially
learning what kinds of variation as well as how much variation are allowed within a
category, and also the degree of change required before there is a change in category.
More specifically, if acoustic cue A varies in a way that is predictable, and acoustic
cue B varies in a way that is not predictable, then a learner may decide that acoustic
cue A is more reliable for deciding if a speech sound is, for example, a /p/ or a /b/.

Thus, learning from multiple talkers could be causing the formation of a more
general word representation. A more general representation, which would include
information on the expected degree of within-word variation, would allow a word to
be recognized across talkers. Such a representation thus solves the apparent paradox mentioned in Section 1.2: learners can learn the patterns of variation within and across speech categories, thus recognizing distinct categories while also retaining knowledge of the variation within them. In fact, in this scenario knowledge of speech sound variation is necessary for categorizing speech sounds.

Word forms are one kind of speech sound category. Phonemes are another. There has been some controversy as to the existence of phonemes (see Chapter 2, Section 2.2.1), but if it is possible for people to parse known words into phonemes, then it should be possible to do the reverse and build new word representations out of familiar phonemes. Supposing a speaker is learning a new word in a language that she is highly proficient in, then she should have a lot of knowledge about the phonemes that the word is made up of. In particular, she should know how these phonemes’ pronunciation varies in different contexts and when pronounced by different speakers. Thus, a very proficient speaker would not benefit from learning the new word from multiple talkers. She already has the knowledge that learning from multiple talkers would provide. Learning from a single talker would give her all the information she needed, while adjusting to different talkers might actually prove a hindrance. A less proficient speaker on the other hand would not have this extensive knowledge of the language’s phonemes as pronounced in varying contexts and by different speakers. Thus, less proficient speakers would get information they didn’t previously have, by learning a new word from multiple talkers. Past studies finding a multiple talker learning benefit all focused on less proficient speakers, such as infants, young children, and L2 learners; whether proficient speakers cease to
benefit from multiple talkers has not been tested.

1.4 A secondary question: the relationship between perception and production

The relationship between speech perception and speech production is a long-standing question in phonetics. While few would argue that the two are entirely unrelated, the degree to which one informs the other is uncertain. At one extreme, Motor Theory states that speech perception more or less is speech production: in order to perceive speech, a speaker uses her knowledge of motor commands to reconstruct the production of what she hears, and thus is able to perceive speech despite its acoustic variability (Liberman et al., 1962; Liberman & Mattingly, 1985). In this scenario, the representation of speech is the same for both perception and production. The views opposed to Motor Theory are less cohesive, but attempts to disprove Motor Theory mostly focus on its claims about the innateness of speech (see Chapter 2, Section 2.3.2 for further discussion of Motor Theory and innateness). Regardless of whether speech is innate, however, the evidence is unclear for whether perception and production share a single representation, or whether the two have separate representations that occasionally inform each other.

This question can be addressed by studying how learning with high variability affects each. Multiple talker training is perceptual learning: participants do not produce the new words during training. Thus, if the representation of a word for its perception is not closely connected with its representation for production, then the benefit of multiple talkers would not be expected to extend into production. Most studies have focused exclusively on the benefit of high variability training...
in the perception of sound categories (Lively et al., 1993; Kingston, 2003; Holt & Lotus, 2006; Rost & McMurray, 2009, 2010). On the other hand, Richtsmeier et al. (2009) found that children produced words more accurately if they had learned them from multiple talkers, but did not address their perceptual accuracy. Very few studies have addressed both perception and production simultaneously. Of those studies that have addressed both, learners were found to better perceive and produce foreign speech sounds (within familiar minimal pairs), when they learned from multiple talkers (Bradlow et al., 1997, 1999). Again, however, these studies only examined L2 speakers, who are not proficient with the language.

Testing both proficient and less proficient speakers on both perception and production would further tease apart the two. If the multiple talker benefit does indeed taper off with increasing proficiency, it is possible that this could happen at separate times for perception and production. This would indicate that the two do not share a representation. Alternatively, the two may parallel each other, suggesting a very close relationship between them.

1.5 Summary and explanation of terminology

It is hypothesized that the helpfulness of high variability training depends on the language proficiency of the learner. In particular, variability is often assumed to be helpful because it draws attention to the more reliable cues for speech category recognition. This reasoning suggests that only less proficient learners benefit from variability across the board, in both perception and production of new words. Looked at another way, variability is also considered to help with forming
a generalizable, possibly more abstract representation. This view is supported if only less proficient learners benefit from variability. A more proficient learner has generalizable representations of the phonemes of the word already, and so can use these representations to form a generalizable word representation.

The facilitative role of talker variability may also be different in perception vs. production. If variability is helpful for all learners’ productions of new words, but is helpful only for inexperienced learners’ ability to perceive new words, then this suggests that people use a different representation for perceiving vs producing a word. This runs counter to the hypothesis that people use a single representation for words, as has sometimes been argued.

Table 1.1 summarizes what is currently known about the interaction between phonological experience, task, and benefit from multiple talkers in learning word forms, and the gaps that will be filled by the proposed experiments. Filling in these gaps will address the questions that have been introduced in this chapter.

<table>
<thead>
<tr>
<th></th>
<th>L1 infants</th>
<th>L1 children</th>
<th>less proficient adults (like L2 speaker)</th>
<th>proficient adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>segment perception</td>
<td>-</td>
<td>-</td>
<td>yes</td>
<td>NA</td>
</tr>
<tr>
<td>segment production</td>
<td>-</td>
<td>-</td>
<td>yes</td>
<td>NA</td>
</tr>
<tr>
<td>novel word perception</td>
<td>yes</td>
<td>Experiment 1a</td>
<td>Experiment 3a</td>
<td>Experiment 2a and 3a</td>
</tr>
<tr>
<td>novel word production</td>
<td>NA</td>
<td>yes</td>
<td>Experiment 3b</td>
<td>Experiment 2b and 3b</td>
</tr>
</tbody>
</table>

Table 1.1: Does variability help?
1.5.1 Terminology

Terms may be used in different ways, even within the same discipline. This dissertation will use a variety commonly used terms, that the reader may think of in a different way than they are being used in this dissertation. Therefore, some terms, and a few abbreviations, will be given a brief explanation here, for the reader’s reference.

**language proficiency** This will refer to the broad knowledge that a speaker of a language acquires about that language from experience. A native speaker has complete language proficiency, while a newborn infant or a beginning second language learner has essentially no language proficiency. Older infants, children, and second language learners with some experience are somewhere in between.

**phoneme** Phoneme is generally used to refer to an abstract unit of sound, which is strictly categorical and does not retain any phonetic detail. However, the term is not used consistently. Here, the term will be used in a more general sense, to mean a sub lexical speech sound category, which a word may be parsed into. For further discussion of conceptions of the phoneme, see Chapter 2, Section 2.2.1.

**speech sound category** Speech sound categories refer here to word forms (representations) and to phonemes. Both of these will be considered categories, and the recognition of each as a categorization process (Holt & Lotto, 2010).
Abbreviations

**L1 speaker**  First language speaker, a native speaker

**L2 speaker**  Second language speaker
CHAPTER 2

An Overview of Previous Research on Learning from Variability
2.1 Introduction

In this chapter, I review previous work related to how learners form robust speech categories, including segments and word forms.

2.2 The importance of phonology

Languages differ in their sound systems. For one, languages have different segment categories. Hindi has a contrast between retroflex and dental plosives, whereas English does not. English has labial fricatives, whereas Hindi and Mandarin do not. Mandarin has retroflex fricatives, whereas English and Hindi do not. Additionally, the context in which particular segments may occur differs between languages. Thus, while both Mandarin and English have a speech sound that is represented as /s/, in English /s/ may occur before /t/ word-initially, and it may occur word-finally, but words in Mandarin do not allow /s/ in either of these contexts. Additionally, languages differ in stress, tone, and intonation.

Further, speakers of a language are sensitive to these patterns in their language. A wide range of studies shows that speakers are sensitive to phonotactics Ohala (1999); Bailey & Hahn (2001); Hay et al. (2003); Edwards et al. (2004); Richtsmeier et al. (2010); Daland et al. (2011), and to subtle variation in pronunciation Port & O’Dell (1985); Warner et al. (2004); McMurray & Aslin (2005); Warner et al. (2009).

Because languages differ in both their sound categories and the patterns of these categories, the phonology of a particular language must be learned. Minimally, in order to recognize and ultimately understand speech, language users must learn which acoustic dimensions are most predictive of a given sound category. Further,
they must be able to adjust to new contexts, such as new speakers, different dialects, or noisy environments. That is, language users must have a robust, flexible representation of a sound category.

The definition of mature language ability is having arrived at this sort of representation for the native language’s speech categories. Language users may adjust with new experience, but any adjustments will be smaller than for infants, young children, and second language learners. Learners, on the other hand, do not yet have robust representations, nor complete knowledge of the sound system. They must learn the phonology.

2.2.1 The role of phonemes and phonology

While the existence of linguistic sound systems is uncontroversial, there has been some debate about the particulars. The existence of abstract sound units - phonemes - has been especially controversial. Adding to the confusion is that the term *phoneme* is not used consistently. Sometimes the term is assumed to refer strictly to the classic definition, an abstract unit which is strictly categorical and does not retain any phonetic detail. However, it may also also be used to refer to a less well-defined unit of sub-syllabic sound, which may or may not include phonetic detail. Viewpoints range from assuming that phonemes are indeed entirely abstract (McClelland & Elman, 1986), to arguing against their existence at all Port (2007).

If phonemes do not exist, then an alternative viewpoint is that word forms are recognized whole, as categories in themselves. However, there are several reasons to believe that this is not the case. That words *can* be consciously segmented
into smaller subunits is clear from alphabets and rhymes in poetry. Further, children progress in their ability to consciously segment words into smaller units, first separating rhymes, then syllables, and finally individual speech segments (Anthony et al., 2002).

But all this evidence for sub-lexical structure does not absolutely imply that speakers segment words into phonemes during speech processing. Nor is it evidence that words are in fact stored in memory as collections of smaller speech categories, or that these smaller speech categories exist as abstract units, as categories could also be explained as constellations of exemplars (see Section 2.3.2).

However, Johnson (1997a) gives further reason to believe that phoneme-like units exist. Johnson points out adults can use the transitional probabilities present between syllables to segment speech into words (Saffran et al., 1996); thus the process of segmenting the speech stream into words necessitates sub-word processing. That said, Johnson’s view on phonemes is more nuanced than either of the extremes, of phonemes being entirely abstract, or not existing at all. In his model, the task of recognizing words is accomplished without explicitly segmenting the speech stream into abstract subunits such as phonemes. Johnson (1997a) does not claim that segments do not exist, but rather than they are emergent phenomena, that do not exist independently as abstract representations.

In contrast to Exemplar models, the TRACE model of speech perception makes use of entirely abstract phonemes (McClelland & Elman, 1986). In this model, words are recognized phoneme-by-phoneme. From a sequence of phonemes, coded in features, the TRACE model uses features to recognize phonemes, which activate
all words containing that sequence of phonemes. As more of the phonemes from the word are recognized, the number of words corresponding to that sequence is reduced, until finally the entire sequence of phonemes has been recognized, and the word corresponding to that sequence is then recognized. This model requires that speech be segmentable into abstract, letter-like units: the classic definition of a phoneme.

Port (2007) argues strongly against the idea of phonemes, and particularly against word representations being composed of phonemes. Here, a phoneme is defined as an abstract speech sound unit, which is invariant across speakers, and is invariant in how it is represented in the minds of speakers. While Port’s definition of a phoneme is an extreme version that would be difficult to argue in favor of, he implies an important question: how are words stored? What do sub-word sound categories look like? There is reason to believe that in fact, words are stored whole, with phonetics that are specific to a particular word.

Pierrehumbert (2003) reviews many studies that point to words having specific phonetic detail, which either does not spread or does not spread immediately to other words. In particular, she notes that more frequent words undergo a sound change sooner than less frequent words, while very infrequent words may not undergo the change at all. She sketches out how word-specific phonology can be accommodated in Exemplar Theory (see Section 2.3.2). In her model of Exemplar Theory, clouds of exemplars are associated with phonological units - something like phonemes - that are contained within words. Thus, these phonological units are emergent, rather than existing on their own. A sound change that applies to a broader class of words
occurs when an environment common to a set of words is targeted. That is, if, for example, schwa-deletion occurs in unstressed syllables, then the set of words with [ə] is targeted. This does explain how words can have word-specific phonetics, while also accommodating phonological generalizations.

Sumner et al. (2013) notes the contradiction between findings that both the more frequent pronunciation of an allophone and its canonical form (which is often a less frequent variant) facilitate recognition of a word containing that allophone. They hypothesize that the apparent discrepancy can be explained by the acoustics of the rest of the word. That is, if the acoustics of the rest of the word matches what a listener would expect for the canonical variant, then the canonical variant facilitates word recognition, while the same is true for the frequent variant. In a set of experiments, they find that listeners categorizing phonemes are more influenced by phonetic frames than they are by the lexical or non-lexical status of words or pseudo-words. Stimuli were [b] and [p] initial words, where one word was a real word and its pair was a pseudo-word: bottom-pottom, and these stimuli varied continuously on VOT. Participants were asked to categorize what they heard as one or the other. VOT is influenced by speaking style, such that it is longer for a [p] in a canonical style and shorter in a casual style. Participants were more influenced by what Sumner et al. refer to as phonetic frames than they were by word-status. Similar phonetic patterns occur across different word forms, such that within a word, either canonical or casual pronunciation variants will occur, but they will not be mixed within a word - this is what Sumner et al. mean by phonetic frame. In this case, the overall phonetic frame of the word biased listeners’ responses, while whether
their response was a word or pseudo-word did not seem to influence them. These results are not explainable by word-specific acoustics, because the pseudo-words used do not have a lexical representation, and thus can have no word-specific acoustics associated with them. Sumner et al. thus argue that storing exemplars at the word level cannot explain their results - a sub-lexical level of representation is required.

Adaptive Resonance Theory (ART), adapted to speech perception by Goldinger & Azuma (2003), provides a compromise between these differing viewpoints. Like Johnson (1997a), they also consider units of speech perception to be emergent properties. In ART, bottom-up and top-down processing occur at the same time, and provide feedback to each other. However, units of speech perception are not confined to classic phonemes:

“Processing in ART begins when featural input activates items (feature clusters) in working memory. Items, in turn, activate list chunks in memory. These are products of prior learning (perhaps prototypes) that may correspond to any combination of features. Possible chunks therefore include phonemes, syllables, and words.” (Goldinger & Azuma, 2003, p. 307)

In other words, speech processing units could be just about anything, although larger units are favored in perception, all else being equal. These units come into awareness only as they are brought to attention. Thus, speech processing is flexible, adaptive, and self-optimizing.

In summary, word forms could either be the smallest unit of representation, or they could be a category (the word) that is made up of smaller categories (phonemes). The way that word forms are represented is relevant to how word forms
are learned. Word forms could be learned whole, without parsing them into smaller units of sound. They could be learned by recognizing sub-lexical, phoneme-like units, and then constructing the representation out of these categories. Or, they could be learned in a flexible way, such that the learner can learn words using either method, and chooses the one optimal to the situation. Words themselves may be represented flexibly, as noted by Goldinger & Azuma (2003). Indeed, this may even be the definition of a generalizable word representation: a word representation that may be used to perceive and produce words from variable voices and in variable situations may require multiple levels of representation.

2.2.2 First language learners and second language phonological learning

Infants are born sensitive to a wide range of speech sounds. By the time they are one year old, they have largely narrowed that sensitivity to native speech sounds (Werker & Tees, 1984a). They continue to refine their knowledge of their native sound system for many years, becoming more and more sensitive to smaller and smaller subunits of words (Anthony et al., 2003, 2002), and refining their knowledge of suprasegmental information such as stress as late as 12 years of age (Vogel & Raimy, 2002).

How we go from pre-linguistic infants to native speakers is only partly answered, but is almost certainly a combination of overall developmental changes, inborn biases, and experience. What changes occur in order for this to happen is also an important question. Adult native speakers of a language are better at learning new word forms, for example - what is it that allows native speakers to do this?
Second language phonological learning is in some ways similar to first language phonological learning. However, second language learners typically vary more widely in their phonological abilities, even with the same amount of exposure to a language (Piske et al., 2001; Kondaurova & Francis, 2010). In contrast, first language learners do vary, but the variation is more subtle (Schertz, 2014). A contributing factor is likely first language interference (Best, 1995; Flege, 1997). Whatever drives the difference, it is clear that experience with the second language does not entirely predict proficiency or attainment of a native-like phonology.

Much of the study of second language learners has focused on how the first language interferes with learning the sound system of the second language. For example, the difficulty Japanese learners have with English /r/-/l/ has been particularly well-studied. Native English-speakers rely primarily on the third formant to distinguish this contrast, which is not used for any contrast in Japanese. Accordingly, Japanese listeners use the second formant to make the distinction, a cue with which they are familiar, but which is less reliable (Iverson et al., 2003). Best (1995) proposes that second language learners are better with sounds that differ by a feature that exists in their language, but struggle with sounds that differ by a feature that is not used contrastively in their language. For example, English speakers are better at distinguishing voiced and voiceless lateral fricatives, /\&/ and /\&/, apparently because English has a voicing contrast. However, they struggle with plosive vs. implosive stops /b/ and /\&/, plosive vs. implosive being a distinction English does not make (Best, 1995).

Thus, while second language learners may get the same experience with a second
language as a first language infant learner, the second language learner must also
overcome previous learning. In a sense, when communicating in the second language,
she must unlearn some of what she has learned in her first language. Other
differences almost certainly exist between first and second language learning, but
this difference goes a long way to explaining why second language learners struggle
with some parts of the second language phonology more than others.

2.3 Finding consistency in an inconsistent world

It is commonly taught in introductory phonology classes, that English has a
phoneme, /t/, that has four allophones: t, tʰ, r, and ?; In other words, one might
be tempted to say that English has four kinds of /t/, which all count as the same
/t/. This is however a simplified view, as English has, in fact, infinitely many t’s.
No two are ever pronounced in exactly the same way twice. And yet, most English
speakers think of English as having only one /t/.

More broadly speaking, no speech sound or word is ever pronounced the same way
twice. Not only does speech vary across dialects and between individual speakers,
it varies even within an individual speaker.

Given this, recognizing that different pronunciations of a speech sound should
be categorized as the same sound is not a trivial task. While this is obscured by the
ease with which adults are able to recognize speech across contexts and speakers,
the difficulty is highlighted by the generally low accuracy of machine recognition
(Bernstein & Franco, 1996).
2.3.1 The search for invariance

Much of speech perception research has been devoted to the question of how language users are able to recognize a sound as a constant - a single category - despite the amount of variation in its realization. In the literature, this has become known as the search for invariance. While no speech sound is ever pronounced the same way twice, it is theoretically possible that some part of the acoustic signal that is /t/ is invariant. In other words, it is possible that there exist invariant acoustic cues to the identity of a speech sound. Early speech researchers hypothesized that this is what language users are picking up on, and using to categorize all variant pronunciations of /t/ as the same category. The rest of the acoustic signal could then be ignored; it would be considered irrelevant to the categorization of the sound. Potter & Steinberg (1950, p. 807) outlined the goal of early research in speech perception, saying “The problem is to determine those physical properties that are invariant in the several [different] utterances that enable the ear to identify them as a given vowel.”

In examining the vowels of 76 speakers, Potter & Steinberg (1950) were looking for acoustic measurements that could be used to specify vowels. In other words, they desired to find the acoustic description of each vowel, that would differentiate it from other other vowels. Instead, they observed that women, men, and children have different ranges of fundamental frequency, and of the first three formants. Further, they noted that the vowel spaces of the vowels, as produced by these three groups, overlapped considerably. In other words, while F1 and F2 are undoubtedly important in differentiating vowels from one another, they cannot uniquely identify a particular vowel.
Peterson & Barney (1951) used the data from Potter & Steinberg (1950) in a vowel identification task. They found that vowels were correctly identified as the intended vowel 94% of the time, despite the considerable overlap in the vowel space. Human listeners were able to do what the researchers’ model for vowel identification could not: they could identify the vowels much better than a model based solely on F1 and F2. It did not seem to be the case that vowels, or perhaps any other speech sounds, could be identified based on a simple two-dimensional graph of F1 and F2.

Thus far, all models of vowel identification had been based on intrinsic models - that is, that the acoustic signal of the vowel itself was sufficient for identifying it. Joos (1948) suggested a more complex model, whose predictions were then tested by Ladefoged & Broadbent (1957). In their classic experiment on vowel perception in context, they used a speech synthesizer to manipulate the range of formant frequencies in a carrier sentence, “Please say what this word is”. The following word was either “bit”, “bet”, “bat”, or “but”. They found that the same acoustic word was identified differently depending on the formant range in the carrier sentence. They concluded that the identity of a vowel could be affected by extrinsic information, such as the formant ranges in the vowels that preceded it. In particular, they concluded that a listener could identify a vowel based on their previous knowledge of the speaker’s vowel space. This began the idea of “speaker normalization”. But still, this did not entirely explain how listeners identified vowels: listeners are able to identify a vowel produced in isolation, even if you have never heard the speaker say anything before, and even if you know nothing about the speaker.
Syrdal & Gopal (1986) postulated that the difference between F0, F1, F2, and F3, calculated in Bark, might be important for vowel identification. They found that much of the variability previously noted was reduced, and this information was able to identify some, but not all, American English vowels. Unlike the Ladefoged & Broadbent (1957) approach, this model was an intrinsic model, which would explain how speakers are able to identify vowels in isolation.

The continuing failure to find true acoustic invariance led to an alternative approach. Liberman et al. (1962) suggested that the invariance was not to be found in the acoustics, but in the motor commands to the articulators. In other words, while the acoustic signal that is produced might not be consistent across speakers, or even within a speaker, the motor commands given to the tongue, vocal tract, etc. are always the same, for a given speech category. Speech perception happened through the mapping of variable acoustics to the motor commands, predicting that speech production is a prerequisite to speech perception. This came to be called Motor Theory.

A number of challenges quickly called Motor Theory into question. In addition to a critical review by Lane (1965), the death knell for Motor Theory as it was first conceived came from a study by Eimas et al. (1971). In this study, one-month-old and four-month-old babies were tested on their perception of a /b/-/p/ continuum. Both groups showed categorical perception of the continuum, just like adult listeners, despite never having produced these sounds. If perception required knowledge of production, then these results would not be possible.
2.3.2 Speech categories: invariant and innate or variant and learned?

Regardless of whether speech perception researchers are able to find the invariance in the variant acoustic signal, listeners undoubtedly do. Again as noted by Peterson & Barney (1951), despite overlapping the overlapping acoustics of vowels, human listeners are able to identify vowels in isolation with 94% accuracy.

After Eimas et al. (1971) disproved the original Motor Theory, a new conception of Motor Theory proposed that knowledge of invariant articulatory gestures was innate (Liberman & Mattingly, 1985). While this does solve the problem of very young infants being able to distinguish speech categories, it does not explain how speech remains perceptable, even when the traditional cues, which presumably would be the innate ones, are removed. Remez et al. (1981) showed that listeners were still able to perceive speech, even when everything but the dynamic spectral pattern was removed from the signal. In other words, if there were some invariant cue that was present in speech, then researchers had, at the least, not been looking at the right ones.

An alternative viewpoint is that speech does not in fact contain true invariance. Lindblom (1996) called the search for invariance into question:

Do we have reasons to believe that, under the surface of all the variable phonetic shapes, there is a “core” of constancy that accounts for the perceptual success of each individual realization? No, not at all. Any set of intelligible pronunciations are, by definition, articulatorily, acoustically, and auditorily equivalent with respect to the goal of perceiving the given lexical item correctly, but that does not logically
entail assuming articulatory, acoustic, or auditory invariance in the phonetic behavior. The common point of these examples is that an invariant (nonsignal) end is reached by variable (signal) means.

He calls his hypothesis for how listeners identify sound categories the “signal-plus-knowledge” hypothesis. By this he means that the speech signal alone is not enough to perceive speech, an idea going back to Ladefoged & Broadbent (1957), who first proposed an extrinsic model of speech perception. Lindblom, however, was the first to propose that the signal might not contain any invariance, and that invariance was in fact unnecessary for perceiving speech.

If the signal itself does not contain invariance, then the question becomes how listeners do use the signal. In this view, listeners may make use of a variety of acoustic cues, and may change how much weight they give each of these cues’ importance depending on the circumstances. Cue weight is flexible, and may change both depending on the circumstances, and over time. In other words, cue weighting may be learned.

Thus, to recognize multiple instances of a sound as belonging to the same sound category, it is necessary to know which parts of the acoustic signal of a sound - the acoustic cues - are most important for distinguishing between categories, and also how to weight those cues under different circumstances. This is true for both individual speech sounds and for word forms.

In support of this view is the fact that adult listeners continue to be aware of the parts of the acoustic signal which do not seem to contribute to speech category identification. Goldinger (1996, 1998) found that adult listeners are sensitive to
individual voice characteristics, and further that they store fine phonetic detail in memory. McMurray & Aslin (2005) found that 8-month-old infants likewise are sensitive to fine-grained detail. These studies showed that listeners are more accurate at remembering a list of words they heard previously when the words are spoken in the same voice at test as during training. In other words, people are sensitive to between-talker phonetic variation. This overturned previous thinking, that in order to perceive a category, such as a word, the perceiver needed to discard information that is not relevant to categorization, and to arrive at an invariant structure. Categorization must be able to occur while preserving variance. Thus, listeners must balance their sensitivity to different acoustic cues, and weight more highly the more reliable cues for distinguishing speech sounds. However, continuing sensitivity to less reliable cues, even in listeners who are very experienced with which cues are the more reliable ones, suggests that listeners may change their cue weighting - for example, in a second language where the more reliable cues may be different.

A range of theoretical approaches rose from this idea of no invariance. Many models do not make strong claims about whether invariant representations of speech categories exist. McMurray & Jongman (2011) refers to these as “cue integration” models. Statistical learning models fall in this category. Statistical learning models propose that learners use the distribution of perceptual cues to group tokens together into categories. Maye et al. (2002) showed that infants are able to keep track of statistical distributions of VOT not only to differentiate [d] from [t], but also to conflate them as the same category. Whether the infants perceived [d] and [t] as two
categories or one category depended on whether they had been exposed to tokens of each in a bimodal (two categories) or unimodal (single category) distribution. Based on these results, McMurray et al. (2009) proposed a computational model, in which the learner forms categories based on Maximum Likelihood Estimation by stochastic gradient descent. The basic idea of a Maximum Likelihood Estimation model is to maximize the likelihood of the parameters given the data. The initial model, which used statistical learning alone, did not converge on a reasonable solution. McMurray et al. (2009) found that the model only converged when competition - meaning that only one distribution (category) was changed at a time - was added. They conclude that statistical learning with competition is sufficient for learning speech categories.

Common to all these cue integration models is the use of a great variety of cues for categorization. Thus, while no explicit claim is necessarily made that incoming speech sounds are reinterpreted - normalized - to an invariant, abstract form, such a step is unlikely to be required in these models. Instead, as proposed by Nearey (1997) the vast numbers of cues available can separate categories from one another.

Exemplar Theory makes a stronger claim about the nature of representation: it claims that invariant representations are unnecessary (Goldinger, 1998; Johnson, 1997b; Pisoni, 1997; Pierrehumbert, 2003). Pisoni (1997) first argued that, regardless of whether invariant cues exist, they are unnecessary, as speech recognition may occur without normalization. Goldinger (1998), Johnson (1997b), and Pierrehumbert (2003) take an even stronger stand, claiming that invariant cues are not only unnecessary, but are unavailable.

Johnson (1997a) proposes a model for recognizing words, and processing them
into subunits. The model simulates experimental results in word recognition, such as competition and recognition of words within non-words. In this model, the unit of exemplars stored in memory is the word, and recognition takes place by comparing the incoming word to those stored in memory, and categorizing it as the one that it is most similar to. Johnson (2006) extends this model to explain preferences for stereotypical speech, given sociolinguistic expectations.

Pierrehumbert (2003) suggests how sound categories might be acquired in an exemplar model. Pierrehumbert views the first units of speech perception as “positional segment variants”, by which she means segments that occur in a particular context. These sub-word categories are learned bottom-up, from statistical distributions. However, as words are acquired, they provide lexical feedback to refine these categories. Further, it is assumed that words, in addition to position segment variants, are stored as exemplars within categories in their own right.

Exemplar Theory thus proposes a model for how categories may exist without reference to invariant cues. While it had been assumed that acoustic cues that were irrelevant to the category of a speech sound were “cleaned up” prior to categorization, Exemplar Theory assumes that all acoustic cues remain in memory, as an exemplar of that category. It is a theory which crucially relies on fine phonetic detail remaining in memory. Rather than storing a single, defining representation of a category, all exemplars of that category are stored, in an exemplar cloud, and all exemplars bear the same category name. Because these “traces” are all stored in memory with great detail including context, then the listener does not need to
compensate for context, provided that she has had experience with that context.

What is common to all models that lack cue invariance, is that speech categories are not inborn. Indeed, this would have to be the case. If there is no invariance present in the signal, then there is not a finite knowledge of speech categories that one could be born with. The very idea of Exemplar Theory is that the cloud of exemplars is constantly in flux - although with greater experience, and thus more tokens of a type, the category may look as though it is completely stable.

From the vast literature in speech perception, what can be taken away is that:

1. The speech signal is highly variable, and may lack any sort of invariance altogether

2. Infants and adults are sensitive to this variation, and retain it in memory whether or not it is retained during categorization

3. Adult native speakers recognize speech categories with a high degree of accuracy

Despite the highly variable signal and their continued sensitivity to individual voice characteristics, adult native speakers do recognize variable pronunciations as belonging to the same speech category. While it is under debate how much knowledge of speech categories is inborn, because languages differ from one another in their sound systems, some learning is necessary. One proposal for how they learn to do this is discussed in section 2.4.

Table 2.1 summarizes the different approaches to how speech is consistently recognized, despite the amount of variance it contains.
2.4 Variation promotes generalization

Research suggests variation promotes category generalization in learning, in diverse domains including motor development (Thelen et al., 1993), problem solving (Oplfer & Siegler, 2007), and learning grammatical features in natural and artificial languages (Gomez, 2002; Hudson Kam & Newport, 2005). Additionally, learning from variable input may promote higher order generalizations. Perry et al. (2010) examined 18-month-old children’s ability to learn new object categories, with and without high variability of objects. They found that the children trained with highly variable instances of new object categories generalized to novel instances of the category. More broadly, the children also developed higher-order biases for learning new object categories in general, being more likely to use certain dimensions for object categorization than others.
Additionally, learning with variation has been found to be helpful for learning generalizable sound categories. The reasoning behind this is the same as for other category learning. Because the real world is highly variable, it is not possible to learn the range of variation within a category from a single token. Variable tokens not only demonstrate more of the range of variation possible for predictive cues; they also draw attention to the which cues are predictive vs. which cues may be regarded as noise. More specifically, if acoustic cue A varies in a proscribed way, and acoustic cue B varies randomly, then a learner may decide that acoustic cue A is more reliable for deciding if a speech sound is, for example, a /p/ or a /b/. This allows the learner to then more reliably recognize a category, especially when presented with examples of that category that she has never heard before.

2.4.1 Perception

The benefits of learning with variability were first noted in perception. In most studies, participants learn from multiple talkers, who vary in their pronunciation, and then are tested with a new talker, who they have not heard before.

In one of the earliest studies on the benefit of high variability training, Lively et al. (1993) found that monolingual Japanese speakers were able to discriminate English [i] from [l] in new speakers when trained on the distinction with multiple talkers. The distinction between [i] and [l] is difficult for Japanese speakers, because [i] and [l] are allophonic - that is, which one will occur is predictable from surrounding context - in Japanese. Participants improved their ability to distinguish [i] and [l] in words and context they were familiar with. Additionally, participants
also improved their distinction ability in contexts they had not heard, and in nonce words. Hardison (2003) likewise found that Korean and Japanese listeners improved their ability to distinguish English [i] and [l] when trained with multiple talkers.

Wang et al. (1999) showed that multiple talker training also improves learning supra-segmental foreign speech sounds. For American English speakers learning Mandarin, distinguishing Mandarin tones is difficult. They were trained with real Mandarin words, pronounced by multiple talkers. Similar to the studies with [i] and [l], they were asked to identify which of two tones a word was, and given feedback. Unlike previous studies, they were learning four categories rather than two: there are four tones in Mandarin. Wang et al. found that participants significantly improved their identification of the tones by an average 21%, and generalized this improvement to novel Mandarin words and to Mandarin words spoken by a novel talker. Further, participants were re-tested 6 months after training, and had retained the 21% improvement. Multiple talker training thus is helpful for supra-segmental as well as segmental speech categories, and remains helpful even when learning more than two categories.

Using the same multiple talker training paradigm, Ylinen et al. (2010) trained Finnish learners of English to change their cue weighting. Finnish has long and short vowels, and the Finnish participants were expected to rely more on durational cues than on spectral cues when identifying English vowels [i] and [i]. Native English speakers rely primarily on spectral cues, or on a combination of spectral and durational cues for distinguishing these vowels. Before training, if durational cues for the vowels were removed, Finnish speakers indeed identified the vowels at chance.
Following multiple talker training, they were able to identify [i] and [ɪ] even when durational cues had been removed, suggesting their cue weighting had changed. This change was reflected also in electrophysiological data from ERP and MMN. This shift in cue weighting explains what is happening during multiple talkers training: learners are adjusting their cue weights, presumably to a more native-like weighting.

Kingston (2003) examined American English-speaking listeners ability to identify German vowels differing in tenseness and height: [ i, ɪ, y, ʏ, u, ʊ, e, ɛ, ø, œ, o, ɔ ]. Unlike the case of Korean and Japanese speakers learning English [ɪ] and [l], a similar vowel system exists in English, with contrasts in tenseness and height. However, phonetically, there are differences: in American English, tenseness is accompanied by dipthongization, which is not the case in German. Additionally, many American English speakers do not distinguish between [o] and [ɔ], and American English does not have front or mid rounded vowels [ɣ], [ ø ], [œ]. Kingston compared three kinds of high variability training: talker variability, context-of-the-vowel variability, and both. In contrast to other studies which report increased learning with more variability, listeners actually improved their ability to distinguish by tenseness more when context did not vary during training. Kingston concludes that there is a “Goldilocks effect” with high variability training: either too little or too much is harmful. It may be that because the listeners were not learners of German, and had not been much if at all exposed to German before, the number of new things to attend to was overwhelming, and hindered learning.

Variability training may also benefit learners who are exposed to “natural” variability in their environment. Iverson et al. (2012) trained two groups of
native-French-speaking, English-learners with multiple talkers, on English vowels. English vowels can be difficult for French speakers, because they have a tense-lax distinction that does not exist in French (i.e., /ɪ/-/ɨ/, /ɛ/-/ɛ/). Both groups had learned English in high school, but one group had lived in England for 6 months, where they were presumably daily exposed to English from a variety of speakers. Additionally, in order to reduce the chances of ceiling effects for the England-living group, only participants who scored below 85% correct on identifying English vowels were used. Thus, while they had had more experience of English, they were not highly proficient in English. The authors noted that the group living in England might not benefit from additional high variability training, since they were already getting such training in their environment. However, they hypothesized that the training might still have value, because environmental exposure would be less focused than the training received in the lab. Specifically, listeners may learn more poorly in a natural environment because non-phonetic information, such as sentence context, can be relied upon to disambiguate word meaning. In contrast, in a laboratory setting, participants are forced to rely exclusively on phonetic information. Participants were tested on a) vowel identification, in which they selected which vowel they had heard, b) discrimination, in which they heard three vowels and chose which one was different, and c) production. Both experienced and inexperienced participants improved their vowel identification after training, and benefited to a similar degree. Both groups also slightly improved in category discrimination. Finally, both groups slightly improved in their production accuracy. Iverson et al. conclude that phonetic training provides a benefit that natural
variability does not necessarily provide. However, given that they pre-selected participants who performed at less than 85% correct, it may be that natural variability is beneficial for some people, while others need the focused phonetic training in the laboratory. Those people who benefit from natural variability would have performed too well to be included in the experiment. Thus, there is a crucial distinction between experience and proficiency: two learners may have had similar experience in a language, but one may, for whatever reason, have attained greater proficiency.

Learning word forms also seems to be aided by phonetic variability. Infants are known to have difficulty learning similar-sounding words (see Section 2.5.1). Rost & McMurray (2009) proposed that infants have immature phonetic categories. While immature phonetic categories are enough for simple tasks like discrimination, they are not mature enough for learning novel minimal pair words. They further note that in previous studies, learning words from a single talker might make maintaining a partially developed phonetic category more difficult. In contrast, learning from multiple talker might make maintaining an immature category easier. They trained 14-month-old infants on English-like words *buk* and *puk*, either with 18 different talkers, or a single talker. Supporting their hypothesis, the infants trained with multiple talkers learned the minimal pair, while the infants trained with a single talker did not.

While most studies have focused on variability between talkers, the affect of a talker can also affect recognition of words. Singh et al. (2004) found that 7.5-month-olds do not recognize words they were familiarized with when the affect
(positive vs. neutral) changed between familiarization and test. By 10.5 months, however, infants did recognize familiarized words that had changed in affect. Singh (2008) found that 7.5-month-olds were able to recognize familiarized words that had changed in affect, if they were familiarized with the words with _variable_ affect, thus suggesting a benefit for variable affect.

Further, Rost & McMurray (2010) found that phonetic variation was only helpful when the variation was on less reliable dimensions, or, as they put it, “irrelevant dimensions”. As in their previous study, infants were trained on new minimal pairs _buk_ and _puk_, which differ from each other only in the voicing of the initial stop. Once again, in one condition infants learned these words when trained with multiple talkers, but with a caveat: the Voice Onset Time (VOT) was manipulated to remain constant across all instances of _buk_ and to remain constant across all instances of _puk_. In another condition, infants learned the words from a single talker, but with the caveat that VOT was manipulated to vary. Thus, in one condition, there was variability except in VOT, the primary cue for voicing in English, and in the other condition, there was variability only in VOT. Infants learned the words in the former but not the latter case. Rost & McMurray conclude that variability is only helpful if the variability is in irrelevant dimensions.

A previous study supports this conclusion, with evidence from non-speech category learning. Holt & Lotto (2006) explicitly manipulated two acoustic cues for a sound category, where only one acoustic cue could be reliably used to distinguish between the two sound categories. Listeners were trained with feedback to identify non-speech sounds as belonging to one of two categories. The sounds varied in
two acoustic dimensions, where one dimension had higher variance than the other. Holt & Lotto found that when one dimension had higher variance during training, listeners relied less on that dimension for categorization. This result suggests that higher variation of irrelevant cues helps learners to know which parts of the acoustic signal reliably distinguish between sound categories.

Apfelbaum & McMurray (2011) later modeled the results of Rost & McMurray (2010), using a computational associative model of word form-object pairing. Their model predicts early stages of word learning, where the learner associates the individual acoustic forms of a word with the word itself (or object). For example, if the learner sees a ball and hears it named, she associates that production of the word with the object. They propose that early words are not under-represented, but over-represented, unless learned from multiple talkers. The new words are too specific to a particular speaker, as the new word will be associated only with the details particular to that speaker. In contrast, if the word is learned from multiple talkers, then the word will be associated with more variable pronunciations, and the learner can notice which cues are varying predictably, and thus should be used for identifying a new instance of that word. While they do not model the process, they further propose that this sort of model can bootstrap phonological development. Over time, the learner will notice that, for example, VOT is reliable for distinguishing similar-sounding words, across the lexicon, and not only for particular word pairs. As phonology improves, the learner should be able to learn words from a single talker.
2.4.2 Production

Numerous studies have found a link between perception and production. For example, Yamada et al. (1994) found that if Japanese L2 speakers of English were better at perceiving the contrast between English r-l, they were also more accurate at producing the difference. It is thus expected that the benefits of high variability training would extend to production.

There has been less work on the benefits of high variability training for production, either of new speech sounds or new words. However, results so far suggest that the benefits of high variability training do indeed extend into production.

Richtsmeier et al. (2009) found that the benefits of phonetic variation extend to production of new words, and to young children. In their study, they trained 4-year-olds on CVCCVC (consonant-vowel-consonant-consonant-vowel-consonant) words either learning the word from ten different talkers, each saying the word once, or repeated ten times with the same talker. They found that 4-year-old children produce new words more accurately and more quickly when they learn the words from ten talkers. They conclude that speech production depends, at least in part, on perceptual learning, as the children had had no articulatory training. Production improvements, with the right kind of variability, can extend beyond a single word, to the improvement of a sub-lexical sequence across words. Richtsmeier et al. (2010) explored children’s accuracy of frequent and infrequent phonotactic sequences. They found that when children were exposed to a sequence in multiple words and by multiple talkers, they improved their production of those sequences in novel words,
showing generalized learning of phonotactics.

Second language learners can also benefit in production from high variability training. Again training Japanese learners of English on the English /r/-/l/ distinction, Bradlow et al. (1997) tested participants on their production as well as their perception of the contrast. Participants pronounced a set of words containing /r/ and /l/ in various positions, before and after multiple talker training. Production then was measured in two ways. American English speakers were asked to rate the pre- and post-training productions. Secondly, different American English speakers were asked to identify words pronounced by the Japanese participants, as one of two minimal pairs differing by /r/ and /l/. The post-training productions were overall rated better than pre-training productions, and the post-training productions were more accurately identified. Although the Japanese participants had only had perceptual training, not production training, they nonetheless improved their production accuracy through high variability training. A second study replicated these results, and found that the benefit was maintained three months later, in both perception and production (Bradlow et al., 1999). Hardison (2003) also found that Korean learners of English benefited from multiple talker training in production as well as in perception, and Wang et al. (2003) found that English learners of Mandarin improved in tone production, having previously shown that they improved in perception (Wang et al., 1999).

However, despite the results indicating high variability perceptual training benefits production as well as perception, Iverson et al. (2012) argues that the apparent correlation is circumstantial. He notes that while Bradlow et al. (1997)
found that high variability training benefits production as well as perception, they did not find a correlation in the degree of improvement in each. That is, a participant might benefit more in one than the other. More strikingly, Hattori & Iverson (2008) found that giving Japanese speakers production training in American English /r/ and /l/ did improve production, but did not improve perception. Because both perception and production rely on experience, Iverson et al. argues, the two may become correlated, but the processes may yet be distinct. Thus, while high variability training may improve both, it is not clear that benefit in one implies benefit in the other.

2.5 Word form learning

2.5.1 Infants and young children

Infants and young children have difficulty encoding word representations in an adult-like way. Much of the debate around infants’ early word representations has centered on whether or not their representations are fully specified. If they are not fully specified, then infants would be expected to accept a wider range of pronunciations as examples of a particular word. What an adult would consider to be a mispronunciation, or a different word altogether, an infant might accept as an example of a word that she knows. It has also been suggested that this occurs because an infant has a very small vocabulary; she does not need to be able to differentiate between very many words, and potentially no minimal pairs - words differing by a single segment. Thus, a second debate centers around whether an increase in minimal pairs is what drives infants to store word representations with
Findings from early studies on infants’ word-form learning suggest that infants do not store words in great detail. Garnica (1973) found that American children between 17 and 22 months have difficulty distinguishing minimal pairs. In the course of her study, she found that only one in 5 children managed to distinguish above chance words differing in the voicing of the initial consonant.

Halle & Boysson-Bardie (1996) explored the degree to which infants accept incorrect pronunciations when recognizing a word. They found that French-acquiring 11-month infants were not able to distinguish minimal pairs that differed only in voicing. In their experiments, infants accepted many nonce-word minimal pairs as real words: for example, [kato] for [gato]. The 11-month-olds accepted mispronunciations of the voicing and even of the manner of articulation of the initial consonant. However, the infants did not accept a nonce-word minimal pair for its real word counterpart when the first consonant was entirely suppressed: [ato] for [kato]. They conclude that when infants are recognizing words, they are less sensitive to phonetic detail. That is, their representation of a word for purposes of recognition is more broad than an adult’s representation, in the range of variation that is included in the category. Infants will thus accept a greater range of pronunciation as belonging to a particular word category.

Stager & Werker (1997) also found that infants do not distinguish minimal pairs differing in place of articulation. They familiarized infants with object-sound pairs: Object A with bih and Object B with dih. At test, the infants failed to notice when the object sound pairs were switched: Object A with dih and Object B with bih.
However, they found that infants distinguished the same pair of nonce words when looking at a checkerboard pattern, although they had failed to distinguish them when looking at objects. Stager & Werker interpret this to mean that infants are able to distinguish the sound categories, but that the increased attentional demands of learning an object-sound mapping reduce their ability to distinguish the minimal pair. The infants were, however, able to distinguish highly dissimilar object-sound pairs, *neem* and *lif*. This suggests that it is the combination of phonetic similarity and the increased demands of object-sound pair association that causes the failure to distinguish minimal pairs.

What is striking about the results of these studies, is that younger infants do distinguish these same phonetic differences, when they are not recognizing them as words. Eimas et al. (1971) found that infants as young as one month, and continuing at 4 months, distinguish [b] from [p], a voicing contrast. In a classic study, Werker & Tees (1984a) tested infants’ ability to distinguish non-native contrasts: the Salish contrast /k/ and /q/ and the Hindi contrast /t/ and /t/. They found that 6 to 8-month-olds do distinguish these non-native contrasts, while by 8 to 10 months fewer were able to, and by 10 to 12 months the infants matched the poor performance of adults and young children. In contrast, infants who were being raised in Hindi-speaking environments retained their discrimination of Hindi /t/ and /t/ and infants being raised in Salish-speaking environments retained their discrimination of Salish /k/ and /q/. These results would suggest a relatively mature phonology; thus it is surprising that 14-22-month infants would not distinguish native sound contrasts in words.
Swingley & Aslin (2000) partially refuted these past studies, showing that 18-22 month children are better able to recognize words that are correctly pronounced. Further, they did not find any evidence that children with larger vocabulary size were better at distinguishing correct from incorrect pronunciations. This goes against the hypothesis that word forms are only represented in more detail as children acquire more minimal pairs in their lexicon. Swingley & Aslin suggest that in the past studies, the poor recognition was due to the infants and young children not being sufficiently familiar with the words. Thus, they argue, it is not that words are underspecified at a particular age, but that newly learned or less familiar words are not as detailed in their representation as familiar, known words.

The processing demands of the task may cause the failure to distinguish novel minimal pairs. Fennell (2012) argues that infants and young children succeed with familiar words because the task demand is lower: they are not simultaneously needing to associate a new object with a new word form. She notes the possibility that in Swingley & Aslin (2000), young children’s representations of highly familiar words are due not to knowing which segments are contrastive, but to the formation of more detailed exemplars. Thus, the greater detail stored with known words is specific to that word, and might not generalize to new words. In other words, if this were true, infants would not be driven to store detailed word representations until they were highly familiar. To test whether infants can store novel word forms in greater detail, Fennell familiarized half of her 14-month infant participants with objects, several weeks prior to learning to pair the object with a word form. When tested on their ability to distinguish minimal pairs /din/, /dm/ and /gin/,
In children, only the infants who had been pre-familiarized with the objects were able to distinguish them. Thus, Fennell concludes that infants are able to store novel words in detail, but only if the task demands are not too high. If the task demand is high, they are liable to have a more fuzzy representation.

2.5.2 Native speakers and second language learners

In principle, adults and older children could also have difficulty with encoding new word forms in detail, and there is some evidence that they do. Aitchison & Chiat (1981) found that older as well as younger children have difficulties with recalling newly learned word forms, and frequently made phonological/phonetic errors when attempting to recall a newly learned word form. Smith et al. (1991) found the same for adults, although they argue that adults make qualitatively different errors than children.

Second language learners also seem to have more difficulty remembering word forms that contain difficult sounds. Diaz et al. (2012) found that more Dutch late learners of English reached native levels of performance when categorizing a difficult English contrast - /æ/-/ɛ/ - than did when the task was lexical identification or lexical decision. This mirrors the result with infants, who are able to distinguish a speech sound contrast in isolation, but have more difficulty with distinguishing minimal pair words differing by that same contrast.

Pajak et al. (2012) examined whether Mandarin and Korean speakers would attend to phonetic detail when learning minimal pair words. These new words differed by contrasts that did not exist in their native language: length of various
consonants, and alveo-palatal vs. retroflex place of articulation. While Korean has a consonantal length distinction in nasals, it does not have that distinction in other consonant natural classes. Likewise, while Mandarin has an alveo-palatal/retroflex distinction, it does not have that distinction in all consonant natural classes. Because of this, Pajak et al. predicted that Korean learners should do better with minimal pairs with a length distinction, while Mandarin speakers should do better with minimal pairs with an alveo-palatal/retroflex distinction. However, while Korean learners were better at perceptual distinction of length, and Mandarin speakers were better at perceptual distinction of alveo-palatal vs. retroflex, language background did not affect how well they learned minimal pair words. When the groups were split into top-scoring and bottom-scoring, however, the top-scoring group did show the predicted pattern: Korean speakers were better with length minimal pairs, and Mandarin speakers were better with alveo-palatal/retroflex minimal pairs. However, there was no such pattern in the low-scoring group. In other words, some learners use the knowledge that they have about contrasting acoustic dimensions while others do not. Regardless, performing well in this task requires generalization of an acoustic dimension contrast to a whole class of speech sounds, even though that acoustic dimension contrast does not exist for a whole class of speech sounds in the native language. In other words, the task was aided by generalizing, for a Korean speaker, the knowledge that nasals contrast in length, to the idea that stops and fricatives could also contrast in length, even though Korean stops and fricatives do not contrast in length.

The ability to distinguish the sound contrast that differentiates a minimal
pair word does seem to be a prerequisite to distinguishing a minimal pair. Sebastian-Galles & Baus (2005) tested 80 Spanish-Catalan bilinguals on their categorization of an isolated speech sound, lexical decision, and word identification. They found that 90% of participants performed best on the lower level task of speech sound categorization, as compared with lexical decision and word identification; that is, to perform better at lexical tasks, one must first improve performance at sub-lexical tasks.

Why adult second language learners should have more difficulty with words containing difficult sounds, even when they can distinguish those same sounds, is less clear than the case for infants. Infants are more limited in their processing abilities than adults, and have smaller working memories. It makes sense that the higher processing demands of learning a new object-sound pair would limit what they retain in memory. Keeping in mind, however, that even native speaking adults have trouble remembering a recently learned word, it seems likely that a newly learned word is not always immediately represented in a robust and detailed way, regardless of the learner’s level of proficiency or processing limitations. What also seems likely is that greater knowledge of sub-lexical categories does assist in learning new word forms. If a learner already has knowledge of sub-lexical categories, and can use that knowledge, it is one less thing to have to learn in detail at the time of learning a new word.

Pajak et al. (2014) address the question of why learning similar-sounding words are more difficult to learn. They found that as newly learned words containing non-native contrasts become more and more similar, they become more difficult
to distinguish for learners. This is in line with studies finding that even for words that follow native phonology, more similar-sounding newly-learned words are more confusable (Creel et al., 2006; Creel & Dahan, 2010; White et al., 2013). The difficulty of learning new, similar-sounding words, may then be a combination between memory limitations and low confidence in distinguishing similar phonological contrasts, even when those contrasts are distinguishable in isolation.

2.5.3 Words as sound categories

Word forms, syllables, and speech segments can all be thought of as sound categories. In learning to categorize any of these, the learner runs into the same problem: the variability in how they are produced. What is different about word forms, is that they are not “primitives of language” (Carroll, 2013). While they are themselves categories, they may be viewed in turn as being made up of smaller categories - syllables and segments. Thus, learning word forms presents a slightly different problem from learning segments or syllables.

What further differentiates word forms from other speech categories, like syllables and segments, is that they are theoretically infinite in number. While the number of segment types in a language is constrained, and the possible number of syllable types is constrained due to a language’s phonotactics, the number of word forms in a language has no limit. Thus, while an adult native speaker may be assumed to know all the segments and syllables in her language, she will never know all the word forms. She continues to learn word forms throughout her life.
The studies presented here all focus on learning word forms, for these two reasons. Unlike learning categories which are limited in number, even highly proficient speakers continue to learn new words, allowing the comparison of proficient with less proficient learners. Further, because they are built up of smaller categories, they may be learned either by building the word out of the smaller categories - syllables and segments - or by learning the word as a whole. The first strategy is the more economic strategy, but relies on a good deal of experience with the smaller categories - proficiency in the language. The second strategy does not require as much, if any, experience with the smaller categories, but is less economical. It is hypothesized that it requires more variable input, to make up for the varied experience that a more proficient learner would have.

2.6 Is variation always useful?

As discussed in Section 2.4, variation can be helpful for learning new categories, and being able to generalize to new voices. As discussed in Section 2.3.1, Exemplar Theory and cue integration approaches view variation as something inherent not only in speech, but in speech perception. Listeners retain the variability of speech sounds in memory, and may make use of a diversity of cues for recognizing speech sounds.

However, in some ways it is surprising to think that variability could not but increase the difficulty of speech perception. After all, many phonetic studies have been performed looking for how to get rid of variability (i.e., the search for invariance). And indeed, while variability can be useful for learning, it often adds
difficulty to a task.

2.6.1 Learning to cope with variation

Jusczyk et al. (1992) found that 2-month-old infants have not completely mastered recognition of words across variable production. While they are able to discriminate minimal pairs *bug* and *dug* across 12 different talkers, they failed to discriminate if there was a two minute delay between exposure to tokens from one word and testing with tokens from the other word. Infants thus can recognize words across talkers, but the difficulty of doing so has a cost.

Houston (2000) also found that infants have some difficulty generalizing to new voices when recognizing words in fluent speech. 7.5-month-old infants who were familiarized with words from a single female talker were able to recognize the words in fluent speech, when spoken by a different female talker. The same was true for infants familiarized to a single male talker and tested on a different male talker. However, if infants were familiarized with the words from a single female talker and had to generalize to a male talker, they did not recognize the words. 10.5-month-old infants, on the other hand, were able to generalize the words across sexes. Singh et al. (2004) found a similar effect when the variability was affect: 7.5-month-old infants do not recognize familiarized words when the affect changes from positive to neutral and vice versa, but 10.5-month-olds do.

Even young children seem to have trouble accommodating different speakers. Ryalls & Pisoni (1997) tested 3-, 4-, and 5-year-olds ability to recognize words, when words were spoken by multiple talkers vs. a single talker. When words were presented
in quiet, all ages were less accurate at recognizing words spoken by multiple talkers, but only if the multiple talker condition were presented first. When words were presented in noise, all ages were less accurate with words spoken by multiple talkers, regardless of presentation order. They did find that processing multiple talkers became easier with age.

The more different a talker, the older infants are before they can adjust to the difference. Mulak et al. (2013) compared Australian 15-month-old and 19-month-old infants’ ability to identify words pronounced by a Jamaican speaker. They found that only the older infants identified Jamaican pronunciations of a word, whereas both the younger and older infants identified the Australian pronunciation. Further, they found a correlation between vocabulary size and the younger infants’ ability to identify words pronounced by a Jamaican speaker. What is additionally interesting about this result, is that the older infants did not have any more exposure to Jamaican English than the younger infants. However, the older infants did presumably have more mature phonological categories, and may thus have been less confused by the change in dialect. This suggests that more experience allows for broader generalization outside of experience. In line with this idea, Perfors & Dunbar (2010) found that training on a novel phonetic contrast made both learning words containing that contrast, and discriminating a related contrast, easier. It may be that as a sound category becomes more firmly entrenched, it becomes easier to make broader generalizations about that category and related categories. But establishing mature sound categories, which are recognized across speakers and contexts, is not automatic; it takes time and experience.
2.6.2 Continuing to cope: adults also have difficulties

Adjusting for different talkers is not just difficult for infants and young children. Ryalls & Pisoni (1997) found that adults as well as children were slower to repeat words from a list when pronounced by multiple talkers vs. a single talker. Further, Strand (2000) found that listeners process words faster when pronounced by a stereotypical voice, suggesting that adjusting for talkers is not without cost.

Listening to different talkers causes changes in speech processing. That is, listening to one voice is not the same as listening to another. Wong et al. (2004) examined the cortical mechanisms behind listening to multiple talkers in a single block vs. listening to only one talker per block. They found that listening to multiple talkers in a block activated middle/superior temporal and superior parietal regions to a greater degree than listening to one talker per block.

Talker variability also affects vowel perception (Summerfield & Haggard, 1973; Strange et al., 1976; Verbrugge et al., 1976; Assmann et al., 1982) and word identification (Mullennix et al., 1989). Mullennix et al. had adult listeners identify words, where words were presented either by a single talker, or where the talker varied by trial. Listeners were slower to identify words when the talker varied by trial. Mullennix et al. conclude that processing speech by multiple talkers requires more working memory resources than processing speech by a single talker.

Talker variability also disrupts recall. Martin et al. (1989) tested listeners ability to recall words from a list. Lists were read either by a single talker, by ten talkers of a single gender, or by ten talkers with five of each gender. Listeners then had to recall the list in the exact order it had been read. Items that were early in the list
were recalled better in the single talker condition than in either of the multiple talker conditions. These results suggest that memory encoding processes are disrupted by talker variability.

Variability in talkers may be an especial problem for second language speakers. Bent et al. (2010) tested American English speakers and Korean speakers perception of American English vowels in noise. The vowels were produced by 10 different talkers, and presented at three signal-to-noise-ratios. Unsurprisingly, American English listeners were overall more accurate at identifying the vowels. Two vowel pairs - /i/-/ɪ/ and /ɑ/-/ʌ/ - had especially variable accuracy for Korean listeners but high accuracy for American listeners. Acoustic analysis of the production of these vowels showed that Korean listeners were influenced by the acoustic variability. This variability was, however, within normal range for American English. This suggests that at least for some foreign speech sounds, variability across talkers is more difficult for less proficient speakers.

However, not all studies find that variability affects non-native speakers more than native speakers. Lee et al. (2012) compared native and non-native speakers’ perception of Mandarin fricatives. Stimuli were presented at five signal-to-noise ratios, and either blocked by talker or multiple talkers in the same block. All listeners did worse with increasing noise, but non-native speakers’ performance decreased even more. In contrast, multiple talkers caused all listeners’ performance to fall equally. This may be because the non-native speakers had been exposed to multiple talkers, but not to noise, in the second language.

Collectively, these studies point to a processing cost for adjusting for variation
between talkers. Another study suggests that this difficulty is due to the multiple demands made on attention, and thus a failure in selective attention. Mullennix & Pisoni (1990) used a two-choice speeded classification task to assess how variability affects selective attention. To succeed at this task, the listener must ignore one dimension while attending to the other. Mullennix & Pisoni used talker voice and VOT as the dimensions, and exposed listeners to three conditions for each dimension. In the control condition, neither dimension varies. In the orthogonal condition, one dimension is held constant while the other varies randomly. In the correlated condition, the value of one dimension is always paired with the other, such that there is redundancy for cueing the correct response. If the orthogonal condition slows processing while the correlated condition speeds processing, then this suggests that listeners always use all dimensions, and cannot ignore a dimension even when it is irrelevant. This is what Mullennix & Pisoni found. When classifying words when VOT is constant but talker varies, listeners are slower than when both are constant. When classifying the talker, and talker is constant but VOT varies, listeners are also slower than when both are constant. Listeners cannot completely ignore part of the acoustic signal, although they presumably can attend to it less. While this again shows that variability increases processing demands, it also shows why adults can learn new phonetic contrasts, and why they are aided by variability. If adults did not continue to attend to irrelevant dimensions, then they could not change their attention from native-language cues to second language cues, when learning a foreign language. Further, high variability of the irrelevant dimensions is itself a cue to its irrelevance. However, once it is known which dimensions should be attended
to for accurate categorization, then variability, as here, has a cost.

Variability training, however, does not seem to be universally helpful. Wade et al. (2007) had native English speakers listen to multiple Spanish-accented speakers, and identify words. Recognition accuracy was very low - 60%. Recognition only improved for speakers used during training; improvement did not generalize to new speakers. Wade et al. additionally measured the confusability of vowels, which seemed to be the main problem for the English listeners. Comparing the Spanish-accented production to native productions, the Spanish-accented productions were much more variable. Finally, a final experiment specifically looked at vowel recognition, examining the effects of native vs. non-native accent, and effects of variability. Recognition accuracy was largely driven by variability and confusability - more variable productions of a vowel resulted in lower recognition accuracy. Additionally, they found that high variability training only increased sensitivity to the most easily recognized vowel: natively-pronounced [i]. For the least easily recognized vowel, the opposite was observed: non-natively pronounced [i], those trained with less variability performed better. Thus, the effects of variability training are not completely straight-forward, and are not always beneficial.

Although adults have learned to accommodate multiple talkers, these studies all point to there being a processing cost for this accommodation. Thus, while variability can help with learning to generalize across talkers, changes between speakers generally makes perception more difficult, and may also make remembering more difficult.
2.7 Situating the present work in context

To be a native speaker of a language means that you are fully proficient in the phonology of your native language. The nature of this sound system may be based in innate human auditory biases, such that certain patterns and similar sound categories recur across language. However, given that there is no universal phonology, phonology must be learned. This ability to learn a phonology exists not only in infants and children, but remains in adults, as attested to by their ability to learn a second language.

A wealth of studies shows that variability is helpful to both infants and second language learning adults. These attest to the similarity between infant and adult phonological learning, despite the undoubted differences between them. Further, phonological learning can be helpful even for native speakers in their own language. They may need to adjust to new listening conditions, such as a noisy restaurant or construction site, or even to a new talker. In these situations, it may be that the acoustic dimension most used in other conditions is no longer the most reliable cue. In a world that is full of variation, flexibility and the ability to learn and adjust is very helpful. Learning from variable input may induce a more robust category representation by making that category representation more flexible. Learning a category from variable input may also make a learner more willing to generalize beyond their experiences (Perry et al., 2010).

Thus, learning from variable input has repeatedly been shown to benefit generalization of a category. However, while variability is undoubtedly helpful when learning something entirely new, it may not be if the learner has another strategy
available to her. Rost & McMurray (2009) proposed that the difficulty in learning similar-sounding words is due to immature phonological categories - i.e., phonemes. As mentioned in Section 2.2.1, previous experiments cast doubt on the existence of abstract phonemes (Goldinger, 1996, 1998). However, it seems equally unlikely that no sub-lexical categories exist at all. Given that sound patterns exist in all languages, it would be strange if learners did not notice them, become familiar with them, and develop phonological categories. Further, it would be strange if they did not put this familiarity with phonological categories to use. One way they may use them is when learning new categories like word forms. Rather than starting from scratch, as would be necessary for a less proficient learner, a proficient learner may put together known phonological categories, to form a robust, generalizable word representation. Proficient learners should not then benefit from variable input. In contrast, less proficient learners are not familiar with the phonological categories making up the new word; they should benefit from variable input. The experiments here are designed to test this prediction.
CHAPTER 3

Experiment 1: English-Acquiring Children
3.1 Introduction

Language learning begins in infancy, but it is by no means complete by childhood. Young children’s knowledge of their native language’s sound continues to mature, as late as 12 years of age (Hazan & Barrett, 2000). To have adult-like competence in a language’s sound-system, a child must learn to tell the difference between acoustic changes that change the meaning of a word, such as voice onset time, and acoustic changes that do not change the meaning of the word, such as speaker identity.

Fourteen-month infants are able to do this with newly learned words when they learn from multiple talkers, but not from a single talker (Rost & McMurray, 2009, 2010). Three-year-olds more accurately produce new words when they learn from multiple talkers (Richtsmeier et al., 2009). Experiment 1 was designed to test whether children continue to benefit from learning from multiple talkers in perception as well as production, and also if the effect disappears as children get older and more experienced in their language.

3.2 Experiment 1a: Perception

3.2.1 Participants

Participants were native learners of English, between the ages of 2.5-6, with no record of hearing loss or speech disorders. In total, 35 children participated. 9 participants were excluded from analysis due to fussiness. An additional participant was excluded for failing to correctly answer the practice questions, indicating they did not understand the task. This left 25 subjects.
3.2.2 Stimuli

Children learned the following four nonsense words: *dapu*, *gube*, *pegu*, and *tegi*. These words were recorded from 10 native English speakers. One of these speakers was chosen to be the voice in the single talker condition, while all 10 speakers were used in the multiple talker condition. To choose the voice for the single talker condition, the mean pitch was found for each speaker, and then compared to the mean pitch overall, for all speakers. The speaker with the mean pitch closest to the overall mean pitch was chosen as the voice for the single talker condition. Each speaker’s pitch intensity was normalized in Praat, so that it was a mean of 70 dB.

Each nonsense word corresponded to a drawing of a pretend animal. These drawings were taken from Ohala (1999).

3.2.3 Procedure

Participants learned four nonsense words of English, and were then tested on their perception and production of the words. The perception test was a mispronunciation detection task, in which children were asked to identify a word as correct or incorrectly pronounced.

In order to engage the participants, a puppet was used. A participant was introduced to the puppet, and the puppet would in this first interaction mispronounce a word. For example, the puppet might say, “Hi! That’s a great *bink* shirt you have!” The experimenter then corrected the puppet’s pronunciation, and explained to the child that the puppet sometimes said words wrong. In order for the child to fully understand, the puppet was asked to say what color a variety
of objects in the room were, correctly pronouncing the word half the time, and incorrectly pronouncing it the other half of the time. Each time, the experimenter asked the child if the puppet had said the word correctly or incorrectly. The child was encouraged to correct the puppet him or herself.

After the puppet had said at least four words (two correctly, two incorrectly), the experimenter moved to a computer, which was set up on a desk at child height, with a child-sized chair in front of it. A picture of the first practice trial, a dog, was already on the screen. The experimenter invited the child to sit in the chair. If the child didn’t want to, the parent was invited to sit nearby, which generally had the desired effect of getting the child to sit in the chair. The experimenter asked the puppet to name the animal that was on the screen. After the puppet had (correctly) named the animal, the experimenter took out a fake candy, and a rag. She explained to the child that when the puppet says a word correctly, we give her a candy, and when she said it incorrectly, we give her a rag. Then the experimenter asked which should be given to the puppet. The puppet reacted to the candy by making a yummy sound, and saying 'thank you!', and to the rag by spitting it out and saying 'yuck yuck, ick!' More practice trials followed, with a dog, a cat, and a bear, where any incorrect pronunciations were alterations of voicing of the initial consonant. For example, *gat* for *cat*. This continued until the child seemed to understand the game, and got several correct in a row.

Once the child seemed to understand the game, the child was asked to help the puppet to learn the names of some new animals. These animals were presented on the computer screen using Powerpoint, with pre-recorded voices pronouncing the
word corresponding to the animal. Each animal was seen ten times, in all cases. There were two conditions. In the **multiple talker** condition, each time the animal was presented and its name was said, the name was said by a different voice (ten voices in total). In the **single talker** condition, each time the animal was presented, its name was said by the same voice (also ten times in total). To keep the children engaged, each animal moved across the screen in a PowerPoint animation. If the child’s attention wandered, the experimenter would tell the puppet to pay attention. This usually caused the child to look at the screen again.

Following the training phase, participants were given a mispronunciation detection test. For each animal, its picture was shown twice in total, in random order. One of the times each animal was seen, the puppet correctly pronounced its name, and the other time the puppet incorrectly pronounced its name. The child had not previously heard the puppet’s voice pronouncing the new word, so the task required generalizing a word’s pronunciation to a new voice.

In order to avoid biasing the child to a particular response, the experimenter put the puppet in front of her face while the puppet pronounced the word. The child would then give either the rag or the candy to indicate if the puppet had said the name correctly. The entire training and test phase was video and audio recorded.

Participants were discarded if there were noise distractions during the training phase, or if they didn’t get all the filler trials correct, in the test phase of the experiment.

Mispronunciation detection accuracy was scored for each response: correct or incorrect.
3.2.4 Hypotheses

If children have developed sufficiently adult-like phonologies, then they are predicted to not benefit from multiple talkers. They should be able to rely on their knowledge of the allowable variability in speech sounds, to develop robust representations of the new words right away. In this case, children should either perform better in the single talker condition, or there should be no difference between conditions.

It may be that this age group spans the change in children’s phonological maturity, such that the older children have sufficiently adult-like phonologies while younger children do not. Participants’ ages ranged from 2.5 to 6 years. If the older children have more adult-like phonologies while the younger ones do not, then there should be an interaction between age and whether or not children benefit from multiple talkers.

If none of the age groups have acquired sufficiently adult-like phonologies, then all ages should benefit from multiple talkers, like the infants in Rost & McMurray (2009, 2010).

3.2.5 Results

Figure 3.1 shows the results for the mispronunciation detection task. In the single talker condition, children had a mean score of 58.75% correct, in the multiple talker condition children had a mean score of 53.33% correct.
Figure 3.1: Mispronunciation detection accuracy was not higher in the multiple talker condition

Mixed effects models were used to determine if children were more accurate when trained in one condition vs. another, and to determine if there was an interaction between age and training condition.

In all cases, for random effects, intercepts were used for participants and items,
and a by-item random slope was used for the effect of condition. P-values were obtained by likelihood ratio tests of the model with the fixed effect in question against a model without.

A mixed effects model comparison does not support training condition as predictive of accuracy ($\chi^2(1)=0.41$, $p=0.52$). Subjects were not more accurate when they learned from multiple talkers. While in both cases participants perform significantly above chance (one-tailed t-test, $p=0.0307$), neither group performed very well. No interaction was found with age ($\chi^2(1)=1.02$, $p=0.31$), nor was age alone predictive of perceptual accuracy ($\chi^2(1)=0.0228$, $p=0.88$).

A t-test did show that children performed significantly above chance level, $t(24)=1.96$, $p=0.03$. They showed a slight yes-bias, saying that a word was correctly pronounced 64.5% of the time, when the true correct pronunciation rate was 50%.

Given the slight yes-bias, a second analysis was carried out using the sensitivity index $d'$ as the dependent variable. Unlike raw response data, $d'$ gives an unbiased sensitivity score. It is calculated as the z-transform of the hit rate minus the false alarm rate:

$$d' = z(\text{hitrate}) - z(\text{falsealarmrate}) \quad (3.1)$$

The mean $d'$ in the single talker condition was 0.18, and 0.67 in the multiple talker condition, both very low scores. A simple linear regression with $d'$ as the dependent variable still did not support training condition as predictive ($F(1, 23) = 0.89$, $p=0.35$). Thus, even with bias removed, children do not perform better in either condition, nor do they seem to have learned very well in either condition.
A scatter plot of children’s performance is shown in Figure 3.2.

![Scatter plot of children's performance](image)

Figure 3.2: Mispronunciation detection accuracy was not higher in the multiple talker condition.

Performance is generally around chance. Further, it is visually clear that children do not perform better when they learn from multiple talkers.
3.2.6 Discussion

Children did not better identify correct vs. incorrect pronunciations when trained with multiple talkers, and this was true across ages. Further, children did not improve their mispronunciation detection as they got older; older children performed as badly as the younger ones.

This is surprising, given the amount of language experience children have had by 5 or 6 years of age. It is possible that the task itself is at the root of the children’s poor performance. The task here differs from the task in Rost & McMurray (2009) in several ways. For one, the words being learned are not minimal pairs; the minimal pair foil is only heard during the test, and has no referent. For example, for the word *dapu*, children only hear *tapu* during the test, and are asked if it is the correct pronunciation. Additionally, children are asked to say whether the pronunciation is correct or not, rather than being asked to choose between two pronunciations. Choosing the better of two pronunciations may be easier than deciding if a pronunciation is correct or not. It is also reasonable to think that children would be more accepting of mispronunciations in newly learned words, which is born out by a slight yes-bias.

3.3 Experiment 1b: Production

3.3.1 Participants

Participants were the same as in Experiment 1a.
3.3.2 Stimuli

Stimuli were the same as in Experiment 1a.

3.3.3 Procedure

Experiment 1b directly followed Experiment 1a described above. Training was the same as described in Section 3.2.3.

Following the mispronunciation detection task described, the puppet said she was thirsty. The experimenter asked the child if they would come with the puppet to get a drink of water. This provided a short break between the perception and production tasks.

After the break, the child was asked if she would help the puppet say some words. Again, the child was asked to sit in front of the computer, on which appeared a series of pictures. The child was audio- and video-recorded while pronouncing the words corresponding to the pictures on the screen. The first three words were English words that the child was familiar with: *glasses, ghost, and eggs*. This made the child more comfortable with the task. Following this, the four newly learned words appeared consecutively. For each newly learned word, the experimenter said once “This is a X, can you say that?”, where X stands for the word corresponding to the picture. A pre-recorded prompt was not used because in pilot testing the younger children were unwilling to repeat from one. The child’s responses were audio recorded. If a child did not say the word, or if the experimenter had to repeat the word before the child said it, then that trial was discarded.

Production accuracy was measured by tallying points for production errors. For
example, if a child produced [tabu] when the correct pronunciation was [tapu], the word would be given a score of 1. If the child produced [dabu] when the correct pronunciation was [tapu], then the word would be given a score of 2. In practice, few mispronounced words contained more than one error. Additionally, for each inserted syllable, 2 points were added to the score, following Richtsmeier et al. (2009). For example, if a child produced [hetapu] when the correct word was [tapu], then the word would have a score of 2. If the word was produced with no errors, then it had a score of 0. Vowel errors were not included.

3.3.4 Hypotheses

3-year-old children have previously been shown to benefit from multiple talker training (Richtsmeier et al., 2009); thus, it is expected that at least the children aged 3 and under will make fewer pronunciation errors when trained with multiple talkers.

As above, however, it may be this age group spans the change in children’s phonological maturity, such that the older children have sufficiently adult-like phonologies while younger children do not. Participants’ ages ranged from 2.5 to 6 years. If the older children have more adult-like phonologies while the younger ones do not, then there should be an interaction between age and whether or not children benefit from multiple talkers.
3.3.5 Results

Figure 3.1 shows the number of pronunciation errors made for each training condition.

![Graph showing pronunciation errors]

Figure 3.3: Fewer pronunciation errors were made when trained with multiple talkers.

Mixed effects models were used to determine if children pronounced the new words more accurately when trained in one condition vs. another, and to determine
if there was an interaction between age and training condition. As in the perception analysis, in all cases, for random effects, intercepts were used for participants and items, and a by-item random slope was used for the effect of condition. P-values were obtained by likelihood ratio tests of the model with the fixed effect in question against a model without.

A mixed model comparison showed that children pronounced words significantly more accurately when trained with multiple talkers ($\chi^2(1)=4.8186$, $p=0.03$). No interaction was found with age ($\chi^2(1)=0.97$, $p=0.33$), although age was predictive of pronunciation accuracy ($\chi^2(1)=17.00$, $p<0.001$), with older children more accurate than younger children.

3.3.6 Discussion

Children made fewer errors when they learned from multiple talkers, replicating Richtsmeier et al. (2009). Although no interaction was found with age, older children were more accurate than younger children, as would be expected. As age increases, it is likely ceiling performance would be reached, regardless of training condition. It is unclear whether learners would cease to benefit from multiple talker training before they reach this ceiling performance.

This experiment was largely a replication of Richtsmeier et al. (2009), although the stimuli were quite different. Despite the change in stimuli, children still produced new words more accurately when they learned from multiple talkers. This suggests that it is indeed a real effect. This is a rather remarkable result in general. In order to establish a robust representation for production, children must map from variable
adult acoustics to their own articulations, which is presumably what an adult would do in similar circumstances. In the case of a child, however, it is in fact impossible to match adult acoustics; children’s vocal tracts are not yet adult-like. Adults also cannot perfectly match another adult’s acoustics, because vocal tract shapes vary, but they at least can get closer than a child can.

3.4 General Discussion for Experiment 1a and 1b

Children were not found to benefit from multiple talker training in perception, but they did benefit in production. The production results replicate Richtsmeier et al. (2009). Even with a delay between learning the words and producing them, children produce new words more accurately when they learn the words from multiple talkers. This is surprising, since they do not get the same benefit in perception. Thus, multiple talker training is helpful for production for young children, but it is unclear if it is helpful in perception. There are several interpretations of these results.

It may be that the children have sufficiently adult-like phonologies, and they are using their phonological knowledge to learn word forms. In this case, their poor performance in the perceptual task was due to inefficient use of their knowledge. For the production task, if even fairly proficient speakers continue to benefit in production, they would continue to benefit into adulthood. It is possible that production is an inherently more difficult task, that requires a more detailed representation, and that this always benefits from exposure to multiple talkers. If this is the case, then even native-speaking adults should benefit from multiple talker training, if the test is production.
Alternatively, it may be that language production maturity, perhaps because it requires more detail, lags behind language perception maturity. In this case, multiple talker training would eventually cease to improve performance in production. However, performance in perception would cease to benefit first.

A final possibility is that children did not benefit from multiple talker training in the perception task, because of the task itself. The mispronunciation detection task differed from the tasks used in previous studies with multiple talker training, and may have been more difficult. It also differed from the production task, in the kind of memory it tapped into. While the production task did not involve explicit memory - the child simply had to repeat the word - the perception task did involve explicit memory. Further, children in both groups performed poorly. If there was any benefit from multiple talkers, it may have been obscured by the poor performance, which in turn may be due to young children’s difficulty with recall.

With the results from Experiments 1a and 1b, any of these explanations is plausible. A pilot experiment was conducted to look further into the possibility that it is the task that makes the difference.

3.5 Pilot experiment: a second test of perception

Previous studies of infants’ learning from multiple talkers used minimal pairs. Further, rather than the mispronunciation task used in Experiment 1a, they had to choose between two pictures, where the words for the two pictures were minimal pairs. This is an easier task. Whereas mispronunciation detection requires deciding if the word is pronounced closely enough to its mental representation, choosing
between pictures of minimal pairs requires choosing the better fit out of two. It may be that children’s performance will improve when they learn minimal pairs and choose between them. A pilot study was conducted to see if this would be the case. Because the question was whether children could have better performance, all children learned words from multiple talkers. It was assumed that if they were going to do well, it would be in the multiple talker condition, similar to the infants in Rost & McMurray (2009, 2010).

3.5.1 Participants

Participants were 5 three-year-old children.

3.5.2 Stimuli

Stimuli were a subset of the words and their minimal pair foils used in Experiment 1a and 1b: *begu, pegu, gube,* and *kube.*

3.5.3 Procedure

Participants learned four nonsense English minimal pairs, and were then tested on their perception. The perception test was a forced choice between a minimal pair.

In order to engage the participants, a puppet was used, as in Experiment 1a and 1b. Children were asked to help the puppet learn new words. After playing with the puppet for a few minutes, the experimenter moved to the computer, as described in Section 3.2.3. Practice was given, with children “helping” the puppet choose between a picture of a bear and a picture of a pear, when they heard the
experimenter say one of the two words. This continued until the child seemed to understand the game, and got several correct in a row.

Once the child seemed to understand the game, the child was asked to help the puppet to learn the names of some new animals. These animals were presented on the computer screen using Powerpoint, with pre-recorded voices pronouncing the word corresponding to the animal. Each animal was seen ten times, in all cases. All participants learned the words in the same way, identical to the multiple talker condition described in Section 3.2.3. That is, each time the animal was presented and its name was said, the name was said by a different voice (ten voices in total). To keep the children engaged, each animal moved across the screen in a Powerpoint animation. If the child’s attention wandered, the experimenter would tell the puppet to pay attention. This usually caused the child to look at the screen again.

Following the training phase, participants were tested on their learning. They saw two pictures on the screen, one on the right, and one on the left, where each picture pair corresponded to a minimal pair that had been in the training. The experiment said the name of one of the pictures, and the child was asked to help the puppet decide which picture corresponded to the name the experimenter had said. For example, there might be a picture of a *gube* on the right and a picture of a *kube* on the left, and the experiment might say *kube*. If the child has correctly learned the words for each animal, she will choose the *kube* on the left. For each animal, its picture was shown twice in total, in random order, appearing once on the left and once on the right. The child had not previously heard the experimenter’s voice pronouncing the new word, so the task required generalizing a word’s pronunciation.
to a new voice.

Accuracy was scored for each response: correct or incorrect.

3.5.4 Results and discussion

Results, by participant, are shown in Figure 3.4.

![Figure 3.4: Performance was about chance](image-url)
As can be seen in the graph, children still performed around chance, even when the task more closely mimicked the task in Rost & McMurray (2009, 2010). It may be that the method used for infants is less valid for children, or it may be something about the training used for the children. In any case, this study does not determine whether children can improve their perceptual memory of new words, when learning from multiple talkers. However, given the improved accuracy in production in Experiment 1b, which uses the same training method, the children must be learning something.

This study was discontinued, because it seemed very unlikely children would have better performance using this method.

3.6 General Discussion of Results

Children were more accurate when producing new words when they learned from multiple talkers. However, their performance in perception was close to chance, in both Experiment 1a, and a pilot experiment whose method more closely resembled the method used with infants (Rost & McMurray, 2009, 2010). It is puzzling why children do not benefit from multiple talkers, but even more puzzling why their performance is so poor. Children do successfully learn minimal pairs; the participants in these experiments were able to distinguish between practice items *pear* and *bear*.

It is not clear why children’s production is improved if they learn from multiple talkers, but not their perception. The pilot study described in Section 3.5 did not rule out any of the possible explanations described in Section 3.4.
To tease these explanations apart, two further experiments were conducted. Experiment 2 examines native English-speaking adults, and Experiment 3 examines second language learners of English, spanning a range of English proficiencies. Because the perception task may have caused the null result here, the task was changed to more closely match that used in Rost & McMurray (2009, 2010).

3.6.1 Suggestion for another experiment

None of the methods used in the above experiments showed a multiple talker learning benefit in perception. While not conducted here, another method could potentially do so. Swingley & Aslin (2000) found that young children accept mispronunciations of words, but that they are slower to recognize a word when it is mispronounced. Thus, they concluded that children can tell the difference between correct and incorrect pronunciations. This difference could be used to determine if children better learn correct pronunciations from multiple talkers.

Following training with the new words, children could be tested with correct and incorrect pronunciations, as in Experiment 1a, but with foils being known words instead of minimal pairs. For example, a child might hear begu while seeing a picture of a pegu and a picture of a cat. The pronunciation is wrong, but it is close. Presumably, the child would pick the picture of the pegu, even though the pronunciation is not an exact match. This could then be compared with a trial where the child hears pegu while seeing a picture of a pegu and a picture of a cat. In this case, the pronunciation is correct. In both cases, the child is expected to pick the picture of the pegu. However, if the child knows the correct pronunciation,
then she should be faster choosing the pegu when the pronunciation is correct than when it is incorrect. The hypothesis, again, is that children should better learn the correct pronunciation when learning from multiple talkers, like the infants in Rost & McMurray (2009, 2010). Thus, it is expected that children would be faster with correct pronunciations than with incorrect pronunciations when learning from multiple talkers, but not when learning from a single talker.

This experiment is left for future work.
CHAPTER 4

Experiment 2: Native English-Speaking Adults
4.1 Introduction

By the time a language user reaches adulthood, she is considered to be fully proficient in her native language. In particular, she is fully proficient in the sound system - the phonology - of her native language. She is very experienced with all her language’s speech sounds, and their combinations, across many different speakers and speaking situations. However, even in adulthood, a language user may run across words that she has never heard before, and must learn a new word representation, for both perceiving and producing them. To do this, the learner may use her already extensive knowledge of the phonology of her language. In other words, if the learner was not familiar with the nonsense word *rossip*, she is nonetheless familiar with /r/, /a/, /s/, /o/, and /p/, and further, she is familiar with how these speech sounds sound in combination, and as pronounced by different speakers. Thus, she should be able to form a robust representation of the word *rossip* without hearing a variety of speakers pronounce it.

Experiment 2 tested whether native English-speaking adults would benefit from phonetic variation, in the form of multiple talkers, in either perception or production. In the perception test, participants needed to discriminate between minimal pairs. In the production test, participants did a speeded production task. A speeded production task was used because it was expected that, unlike for children or second language learners, adult native speakers would produce English-like words with 100% accuracy. Additionally, a shorter lag time in producing a word should mean that the word’s representation is more robustly represented. In contrast, a longer lag time - i.e., hesitation - would indicate less certainty in the word’s representation.
The hypothesis was that adults would not benefit in perception, but might possibly in production, given the results with young children.

4.2 Experiment 2a: Perception

4.2.1 Participants

37 native English-speaking adults, with no reported record of hearing loss or language disorder, participated. 18 were in the single talker condition, and 19 in the multiple talker condition.

4.2.2 Stimuli

Participants learned the following four minimal pairs: reekus, leekus, slempet, flempet, splannet, sprannet, and rossip, lossip. These words were designed to be English-like words, while being difficult for Mandarin speakers to differentiate. This was important for comparing the results from the native English speakers in this experiment to the results in Experiment 3 with Mandarin speakers (see Chapter 5). All words had a score under 2 using the BLICK phonotactic probability calculator to calculate phonotactic probability (Hayes). For each word, the BLICK phonotactic probability calculator outputs "a numerical value, inversely reflecting the phonotactic ‘goodness’ (probability, well-formedness, word-likeness) of the word”. Thus, lower scores mean a more phonotactically probable word of English.

However, while these words are legal and phonotactically probable in English, all of these words are illegal in Mandarin, which disallows complex clusters, and obstruent codas. Additionally, these words differ from one another by contrasts
that do not exist in Mandarin: /r/ ~ /l/, /f/ ~ /s/. Thus, these words should be learnable by relying on the native phonology for English speakers, but not for Mandarin speakers.

These words were recorded from 10 native English speakers. One of these speakers was chosen to be the voice in the single talker condition, while all 10 speakers were used in the multiple talker condition. To choose the voice for the single talker condition, the mean pitch was found for each speaker, and then compared to the mean pitch overall, for all speakers. The speaker with the mean pitch closest to the overall mean pitch was chosen as the voice for the single talker condition. Each speaker’s pitch intensity was normalized in Praat, so that it was a mean of 70 dB.

As with the child experiment, each nonsense word corresponded to a drawing of a pretend animal. These drawings were also taken from Ohala (1999).

4.2.3 Procedure

Like the children in Experiment 1, participants learned words from either multiple talkers or a single talker, and were then tested on a new voice. Unlike Experiment 1, participants learned four minimal pairs, making eight words in total. This made the task more similar to Rost & McMurray (2009, 2010).

Participants received instructions both from the experimenter, and also in written form on a computer. They were told that they would be learning the names for some funny animals, and that following this, they would be tested on the names of the animals. Pictures of the animals appeared on the screen, and the name of the animal was heard from a sound file corresponding to the animal.
The experiment took place on a Macintosh computer in a sound modulated room. During the learning phase and the perception test, participants additionally wore headphones.

In the single talker condition, the name of the animal was pronounced always by the same token, for a given animal. In the multiple talker condition, the name of the animal was pronounced by 10 different tokens, where each token was a different speaker. Each trial lasted 1000 ms.

Following the learning phase, participants performed a perceptual test. Unlike with the children in Experiment 1, who were asked to decide if a pronunciation is the correct one or not, adult participants decided between two words, a simpler task. For each animal name, participants heard the name while being presented with a choice between two pictures, where the pictures corresponded to the animal being named and the animal whose name was the minimal pair. For example, if the name being pronounced were *reekus*, then the two pictures would be pictures of a *reekus* and a *leekus*, and the participant would have to choose between these pictures. To add time pressure, the participant had 2000 ms to respond. Before the main trials, the participant had eight practice trials, with known words cat˜dog and pear˜bear. For the real trials, each word was heard twice, with the correct picture on the right side of the screen in one case, and on the left in the other. Words were presented in random order.

Accuracy and reaction time were recorded.
4.2.4 Hypotheses

Past studies have found that at least less proficient learners - such as infants, young children, and L2 learners - benefit from multiple talker training, when generalizing to new voices. In Experiment 1, children benefited from multiple talker training on a production task, but not on a perception task. If they failed to benefit in the perceptual task due to the task being too difficult, then adult native speakers may benefit in perception. If, however, the benefit of multiple talker training tapers off with increasing proficiency, then adult native speakers should perform equally well whether they learn from a single talker or multiple talkers.

4.2.5 Results

Results are shown in Figure 4.1.

In the single talker condition, native English speakers had a mean score of 76.73% correct, in the multiple talker condition they had a mean score of 75.33% correct.

Mixed effects models were used to determine if native English speakers were more accurate when trained in one condition vs. another, and to determine if there was an interaction between age and training condition. In all cases, for random effects we used intercepts for participants and items, and a by-item random slope for the effect of condition. P-values were obtained by likelihood ratio tests of the model with the fixed effect in question against a model without.

A mixed effects model comparison does not support training condition as predictive of accuracy ($\chi^2(1)=0.10, p=0.75$). Native English speakers were not more accurate in either condition.
Figure 4.1: Accuracy was not higher in the multiple talker condition

Nor were participants faster in either condition, although this was less clear. Reaction time data is shown in Figure 4.2. Native English speakers had a mean reaction time of 1207 ms in the single talker condition, and 1165 ms in the multiple talker condition.
As is standard practice with reaction time data, incorrect responses were excluded, and the log of the reaction time was taken, to normalize the data. For each participant, reaction times that were more than 3 standard deviations from the participant’s mean were removed from analysis. Normally, data points more than 2 standard deviations from the mean are excluded, but because the stimuli
might inherently differ in how long they take to respond to, a wider range was kept in the analysis. A mixed model comparison does not support training condition as predictive of reaction time ($\chi^2(1)=1.70, p=0.19$). Native English speakers were not faster in either condition.

4.2.6 Discussion

Adult native speakers were not more accurate or faster on a perceptual task, when they had learned new words from multiple talkers. This contrast with previous multiple talker studies suggests that the multiple talker benefit tapers off with increasing proficiency. Adult native English speakers are highly proficient in English phonology. They may be using this knowledge to form robust word representations, and thus do not need the extra information provided by multiple talkers.

However, this null result could also be due to something about the experimental design. To distinguish between these two explanations, the same experiment run with less proficient English speakers needs to be compared with the results here. See Section 5.2.

4.3 Experiment 2b: Production

4.3.1 Participants

Participants were the same as in Experiment 2a, with three exclusions. They were excluded due to experimental error during the production test. This left 17 participants in the single talker condition and 17 in the multiple talker condition.
4.3.2 Stimuli

Stimuli were the same as in Experiment 2a.

4.3.3 Procedure

Experiment 2b directly followed Experiment 2a described above. Training was the same as described in Section 4.2.3

Following the perception test, participants were invited to take a short break. When they were ready, they then did a speeded production test with the words. A speeded production task was chosen, because it was expected that adult native speakers would be perfectly accurate in producing English-like words. Each word was pronounced while its picture was presented simultaneously, and the participant was instructed to say the word as quickly as possible. Thus, this was not a memory test, since participants were being prompted with the name of the animal. To add time pressure, participants were told they would have less than a second to say the name of the animal. Each trial, including the prompting of the word, lasted 1000 ms. The participants pronunciations were recorded using an AKG C3000 Multi Purpose Studio Vocal/Instrument Microphone. The time to production onset was measured from the onset of the prompt word to the onset of the participant’s pronunciation.

4.3.4 Results

Results are shown in Figure 4.3.
Figure 4.3: Time to production onset not faster in the multiple talker condition

Overall, there were three errors in word production; these trials were excluded. Production times that were more than 3 standard deviations from a participant’s mean were excluded. Finally, the log of the time to production onset was taken, to normalize the data.

In the single talker condition, native English speakers took a mean time of 818
ms to repeat the word. In the multiple talker condition they took a mean time of 780 ms.

A mixed model does not show training condition as predictive of the log of the time to production onset ($\chi^2(1)=0.64$, $p=0.42$). Participants were not faster to produce words in either condition.

4.3.5 Discussion

Adult native English speakers did not produce words faster when they learned them from multiple talkers. Greater lag time in production is generally considered to reflect increased processing time. Thus, this result suggests that there is no production processing benefit for learning from multiple talkers for adult native speakers. This could be because the adult native speakers are highly proficient in English, and thus can use their knowledge of English phonology to form robust representations of the new words. The extra information from multiple talkers is not needed, because they already have sufficient information.

Alternatively, as in Experiment 2a, this result could be due to something about the experimental design. Again, a comparison with less proficient speakers, who do not have as much knowledge of English phonology, is needed.

4.4 General Discussion

Unlike for the young children, who benefited from multiple talkers in production but not in perception, native English-speaking adults did not benefit in either. This suggests that, in becoming a proficient language user, production lags behind
perception. Thus, while the multiple talker benefit disappears with increasing proficiency, it tapers off first in perception. This is in line with much other research (Edwards, 1974; Menyuk & Anderson, 1969).

However, it is possible that there was something about the experiment design itself, that caused native English-speaking adults to not benefit from multiple talkers. Also, there may be something about the experiment design of Experiment 1, which caused the children to not benefit from multiple talkers in the perception task. A final experiment, with the same method but with second-language learning adults, was done as a comparison. This experiment is described in Chapter 5.
CHAPTER 5

Experiment 3: Second Language English Learners
5.1 Introduction

While infants and young children are known to benefit from learning from multiple talkers, Experiments 1 and 2 suggest that the effect tapers off as the learner gets older, and presumably becomes more proficient with the language’s phonology. Experiment 3 was done with second language (L2) learners, to see whether benefit from multiple talkers depends on proficiency.

Second language learners are known to have difficulty distinguishing similar-sounding words in a second language (Flege et al., 1997; Aoyama et al., 2004; Boersma, 2005; Cutler & Boersma, 2005; Bion et al., 2006), and also with learning words in a second language, especially when the words contain unfamiliar sounds (Kovacs & Racsmany, 2008; Escudero et al., 2008). Thus, the challenge for L2 learners, when learning words in the second language, comes from their lack of phonological proficiency in the second language.

5.2 Experiment 3a: Perception

5.2.1 Participants

Participants were L2 learners of English (native speakers of Taiwanese/Mandarin), recruited from students at National Chung Cheng University in Minxiong Township, Taiwan. They had a range of proficiency in English. All had had exposure to English in school, usually from age 8 or 9. Some were highly proficient, as many of them were recruited from the foreign languages department. Most came from southern Taiwan, however, where there is less exposure to in-person English speakers than
in the north, as there are very few foreigners. Others had not studied English as extensively after high school. Nonetheless, none could be called truly low proficiency, as National Chung Cheng University is a prestigious university, with mandatory English requirements. All participants included in analysis had grown up in Taiwan, and had not spent more than two months outside of Taiwan.

In total, 42 people participated in the experiment. Of these, data from two participants was not used: one had grown up in Vietnam, while the other was due to experimenter error. Table 5.1 shows how participants were divided into groups:

<table>
<thead>
<tr>
<th></th>
<th>multiple talker</th>
<th>single talker</th>
</tr>
</thead>
<tbody>
<tr>
<td>higher proficiency</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>lower proficiency</td>
<td>9</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.1: Participants in Experiment 3

Language proficiency was self-rated, based on a participant’s score on an English proficiency exam. See Section 5.2.4 for further discussion of how language proficiency was measured.

5.2.2 Stimuli

The stimuli were the same as used in Experiment 2. The description is repeated here for convenience.

Participants learned the following four minimal pairs: rekus, lekus, slempet, flempet, splannet, sprannet, and rossip, lossip. These words were designed to be English-like words, while being difficult for Mandarin speakers to differentiate. All words had a score under 2 using the BLICK phonotactic probability
calculator to calculate phonotactic probability. For each word, the BLICK phonotactic probability calculator outputs “a numerical value, inversely reflecting the phonotactic ‘goodness’ (probability, well-formedness, word-likeness) of the word”. Thus, lower scores mean a more phonotactically probable word of English. However, while these words are legal and phonotactically probable in English, all of these words are illegal in Mandarin, which disallows complex clusters, and obstruent codas. Additionally, these words differ from one another by contrasts that do not exist in Mandarin: /r/ ˜ /l/, /f/ ˜ /s/. Thus, these words should be learnable by relying on the native phonology for English speakers, but not for Mandarin speakers.

These words were recorded from 10 native English speakers. One of these speakers was chosen to be the voice in the single talker condition, while all 10 speakers were used in the multiple talker condition. To choose the voice for the single talker condition, the mean pitch was found for each speaker, and then compared to the mean pitch overall, for all speakers. The speaker with the mean pitch closest to the overall mean pitch was chosen as the voice for the single talker condition. Each speaker’s pitch intensity was normalized in Praat, so that it was a mean of 70 dB.

As with the child experiment, each nonsense word corresponded to a drawing of a pretend animal. These drawings were also taken from Ohala (1999).

5.2.3 Procedure

The methods for the L2 learners are largely identical to the methods for the native adult language speakers in Experiment 2.

The experiment took place on a Macintosh computer in a quiet room.
Throughout the experiment, participants additionally wore headphones. A dual output audio device aloud for audio output through both the headphones, and external speakers, which allowed for the microphone to record audio output at close range during the production task.

Participants were told that they would be learning eight imaginary words of English, and that there would be three parts to the experiment: learning, choosing between pictures, and recording. They were given extensive oral instruction at the start of the study, in case they did not understand the written instructions given during the experiment. While the experimenter explained what the tasks were, they were able to read a script, written in English with Chinese translation. Additionally, a Taiwanese interpreter, a linguistics graduate student at Chung Cheng University, was present throughout, and gave oral translation where needed. The study took place in a rural part of Taiwan, where foreigners are rare, so the interpreter also provided familiarity, as most students were not used to taking part in an experiment run by a foreigner.

During the experiment, participants were able to read the same written instructions that the participants in the native adult study read. As in the previous study, participants learned the same four new minimal pairs, making eight words in total. Pictures of the animals appeared on the screen, and the name of the animal was heard from a sound file corresponding to the animal. In the single talker condition, the name of the animal was pronounced always by the same token, for a given animal. In the multiple talker condition, the name of the animal was pronounced by 10 different tokens, where each token was a different speaker. Each
trial lasted 1000 ms.

For the perception task, as for the native adult participants in Experiment 2, participants had practice trials in order to familiarize them with the task. The same procedure as with the native adult learners was followed here for the L2 learners. For each animal name, participants heard the name while being presented with a choice between two pictures, where the pictures corresponded to the animal being named and the animal whose name was the minimal pair. For example, if the name being pronounced were *reekus*, then the two pictures would be pictures of a *reekus* and a *leekus*, and the participant would have to choose between these pictures. To add time pressure, the participant had 2000 ms to respond. Before the main trials, the participant had eight practice trials, with known words cat˜dog and pear˜bear. For the real trials, each word was heard twice, with the correct picture on the right side of the screen in one case, and on the left in the other. Words were presented in random order.

Accuracy and reaction time were recorded.

5.2.4 Determining English proficiency

All participants were asked to rate their English proficiency in speaking, comprehension, reading, and writing, on a scale of 1 to 4. In order not to intimidate participants prior to the experiment, this was done following the experiment. As mentioned, no subject was truly low proficiency, since all had studied English extensively throughout elementary, middle, and high school. Almost all subjects had also taken English proficiency exams, but since they had not necessarily taken
the same one(s), they were not asked for their scores directly.

In order to give a more uniform rating, participants were asked to use the standardized tests they had taken as a guideline for determining their score. Many seemed to simply translate their overall score on a test to a single number, which they used for all categories. Participants also frequently tended to underrate themselves. If the interpreter present felt that the self-assigned score was inaccurate, he would discuss with the participant further about the self-assigned score, and encourage them to put a higher score. This was done without looking at any of the data collected during the experiment.

Participants were divided into a high and a low proficiency group, using their self-rated scores. Those who had rated themselves with at least one 3 were put in the higher proficiency group, and the remainder were put in the lower proficiency group. Because participants were recruited at a competitive university in Taiwan, that requires a minimum level of English, none of the participants were truly low proficiency. Further, of those who participated, most were quite high proficiency. Thus, there are more people in the higher proficiency group than in the lower.

5.2.5 Hypotheses

Because the new words contain non-native sounds and non-native phonotactics, the L2 learners are expected to have difficulty with learning them, as compared with the native English speakers in Experiment 2. More proficient English speakers are expected to have less difficulty than less proficient speakers.

If benefit from multiple talkers tapers off with increasing proficiency, then there
should be an interaction between proficiency and learning condition, such that less proficient speakers benefit from multiple talker training while more proficient speakers do not. Past studies indicate that at least less proficient learners should benefit from multiple talkers (Lively et al., 1993; Bradlow et al., 1997, 1999; Wang et al., 1999; Kingston, 2003; Richtsmeier et al., 2009; Rost & McMurray, 2009, 2010; Ylinen et al., 2010). Thus, if there is something about the experimental design that caused the null result in Experiment 2, then no benefit from multiple talkers should be found, regardless of proficiency.

5.2.6 Results

Results are shown in Figure 5.1. A non-response was counted as an incorrect answer. A failure to respond indicates that the participants doesn’t know the answer.
Figure 5.1: Only less experienced learners benefited from multiple talkers in perceptual accuracy.

For the lower proficiency group, the mean accuracy was 40.00% correct in the single talker condition, and 57.75% correct in the multiple talker condition. For the higher proficiency group, the mean accuracy was 64.58% correct in the single talker condition and 60.80% correct in the multiple talker condition. Because
non-responses were counted as wrong, 50% correct is not chance level.

Mixed effects models were used to determine if the L2 learners were more accurate when trained in one condition vs. another, and to determine if there was an interaction between English proficiency and training condition. In all cases, for random effects I used intercepts for participants and items, and a by-item random slope for the effect of condition. P-values were obtained by likelihood ratio tests of the model with the fixed effect in question against a model without.

A mixed effects model, with an interaction between learning condition and language proficiency as fixed effects and participant and item as random effects, showed that a model including the interaction was significantly better than a model without the interaction ($\chi^2(1)=3.96$, $p=0.04$).

Planned comparisons were also carried out. The data was split into two groups by English proficiency: a higher proficiency group and a lower proficiency group. For the lower proficiency group, a mixed model, with learning condition as a fixed effect and participant and item as random effects, showed that lower proficiency participants were significantly more accurate when they learned from multiple talkers ($\chi^2(1)=6.12$, $p=0.01$). Interestingly, both the native English speakers and the higher proficiency L2 speakers did slightly better in the single talker condition. Figure 5.2 shows the L2 speakers compared with the native English speakers in Experiment 2.
Higher proficiency participants were not significantly more accurate when they learned from a single talker ($\chi^2(1)=0.27$, $p=0.61$), like native English speakers who were also not significantly more accurate when they learned from a single talker. That said, these parallel results do suggest the possibility of a small effect of multiple talkers on more proficient speakers - whether L1 or L2 speakers - such that they do
slightly worse when they learn from multiple talkers. This result may not have shown up here due to too small a sample size, but is in line with previous work showing that multiple talkers can impede speech processing (Summerfield & Haggard, 1973; Strange et al., 1976; Verbrugge et al., 1976; Assmann et al., 1982; Martin et al., 1989; Mullennix et al., 1989; Mullennix & Pisoni, 1990; Ryalls & Pisoni, 1997; Strand, 2000; Bent et al., 2010).

More proficient L2 learners, like native language speakers, do not benefit from variation, but less proficient L2 learners, like infants, do benefit from variation.

5.2.7 Discussion

These results support the hypothesis that only less experienced learners benefit from multiple talkers. At least in perception, the benefit from multiple talker training tapers off with increasing proficiency. As expected, more proficient learners performed better overall. However, less proficient learners trained with multiple talkers performed almost as well as more proficient learners.

More proficient learners may actually learn slightly less well when they learn from multiple talkers. Both adult native speakers in Experiment 2 and more proficient L2 speakers in this experiment performed slightly worse when they learned from multiple talkers. The effect was not significant, but this may be because it is a weak effect and the sample size is small. Previous studies have shown that changes in talker can make speech processing more difficult (Mullennix et al., 1989; Ryalls & Pisoni, 1997), and can hinder memory (Martin et al., 1989) (for more discussion, see Chapter 2, Section 2.6). If there is such an effect for the more proficient speakers,
then it would be consistent with these previous findings.

5.3 Experiment 3b: Production

5.3.1 Participants

Participants were the same as in Experiment 3a.

5.3.2 Stimuli

Stimuli were the same as in Experiment 3a.

5.3.3 Procedure

Experiment 3b directly followed Experiment 3a described above. Training was the same as described in Section 5.2.3.

Following the perception task, participants were invited to take a short break. When they were ready, their productions of each word were recorded. As with the native speakers, each word was pronounced while its picture was presented simultaneously, and the participant was instructed to say the word as quickly as possible. Thus, this was not a memory test, since participants were being prompted with the name of the animal. To add time pressure, participants were told they would have less than a second to say the name of the animal. Unlike the native English participants, the Taiwanese participants had practice trials for the production task. The first set of practice trials allowed participants 2500 ms in addition to the time the computer took to say the word (typically about 500 ms) to respond. The second set allowed 1500 ms, while the final set was identical
to the real trials, 500 ms. The practice trials ensured that Taiwanese participants native English participants would have the same amount of time to produce a word during the real trials, as it would otherwise be very challenging for the Taiwanese participants. The participants pronunciations were recorded using a AKG C3000 Multi Purpose Studio Vocal/Instrument Microphone.

Production accuracy was measured. All productions were transcribed by an undergraduate student in linguistics. The student had taken two phonetics courses, and had previously worked at Facebook, aligning speech. Following transcription, words were marked either as correctly pronounced or incorrectly pronounced. Changes in vowel quality did not change the correctness of the word. Otherwise, any error was marked as an incorrectly pronounced word. Unreleased final stop consonants were not counted as errors, but changes in manner or place of articulation of released consonants were counted as errors.

5.3.4 Hypotheses

Experiment 3a suggests that benefit from multiple talkers tapers off with increasing proficiency in perception. In Experiment 1 young children produced words more accurately when trained with multiple talkers, but did not derive any benefit in a perceptual test. In Experiment 2, which had methods identical to Experiment 3, adult native English speakers did not derive any benefit, in either perception or production, from multiple talkers. These results suggest that the benefit may taper off first in perception, and only later in production. Because none of the L2 speakers is completely proficient in English, both more and less proficient speakers
may benefit from multiple talkers in the production task. This would indicate that production lags behind perception, and suggests that the perceptual representation of a word is not the same as the representation used for production.

Alternatively, the null result in Experiment 1 may have been due to the difficulty of the perceptual task, which was more difficult than the one used in Experiments 2 and 3. In this case, there should be the same interaction between proficiency and learning condition as in the perceptual task of Experiment 3a.

5.3.5 Results

The less proficient L2 learners pronounced a mean of 10% of the words correctly in the single talker condition, and 31.43% in the multiple talker condition. The more proficient L2 learners pronounced a mean of 48.28% of the words correctly in the single talker condition, and 39.77% in the multiple talker condition. Results are shown in Figure 5.3.

Mixed effects models were used to determine if the L2 learners pronounced words correctly more often when trained in one condition vs. another, and to determine if there was an interaction between English proficiency and training condition. In all cases, for random effects I used intercepts for participants and items, and a by-item random slope for the effect of condition. P-values were obtained by likelihood ratio tests of the model with the fixed effect in question against a model without.

A mixed effects model, with an interaction between learning condition and language proficiency as fixed effects and participant and item as random effects, showed that a model including the interaction was significantly better than a model
Second language learners differed in production accuracy without the interaction ($\chi^2(1)=4.04$, $p=0.04$).

Planned comparisons were also carried out. The data was split into two groups by English proficiency: a higher proficiency group and a lower proficiency group. Visual inspection of the results (see Figure 5.3) suggest that lower proficiency speakers pronounce words more accurately when trained with multiple talkers. Although the
planned comparison of the lower proficiency group was not significant ($\chi^2(1)=2.88$, $p=0.09$), this is likely due to the lower power that resulted from splitting the data into smaller groups. Higher proficiency participants were not significantly more accurate in either condition ($\chi^2(1)=0.69$, $p=0.41$), but again performed slightly better in the single talker condition.

More proficient L2 learners, like native language speakers, do not benefit from variation, but less proficient L2 learners, like young children, do benefit from variation in production.

It is not possible to compare this experiment with the adult native English speakers in Experiment 2b, because the dependent variables in each measured different things. In this experiment, accuracy was measured, while in Experiment 2b the time to the onset of production was measured, since adult native speakers are expected to have perfect accuracy.

5.3.6 Discussion

The results for production accuracy mirror the results for the perception task in Experiment 3a. Less proficient L2 English speakers were more accurate when they learned from multiple talkers. More proficient L2 English speakers, on the other hand, were not more accurate when they learned from multiple talkers, and if anything were more accurate when they learned from a single talker. This is similar to the perception result in Experiment 3a.
5.4 General discussion

Less proficient L2 English speakers learned new English-like words better when they learned from multiple talkers, as has been found in previous multiple talker studies (Lively et al., 1993; Bradlow et al., 1997, 1999; Wang et al., 1999; Kingston, 2003; Richtsmeier et al., 2009; Rost & McMurray, 2009, 2010; Ylinen et al., 2010). This was true for both the perception and production of the new words.

This result complemented the result with native English speakers in Experiment 2. Only less proficient speakers benefit from multiple talker training, while more proficient speakers - whether first or second language speakers - can rely on their phonological knowledge when forming a new word’s representation.

Paralleling results from Experiment 2, both native English speakers and the more proficient L2 speakers performed slightly worse when they learned from multiple talkers, although the effect was not significant in either case. Additionally, the more proficient L2 speakers were slightly less accurate when they learned from multiple talkers, although again not significantly so. Thus, while none of these results was significant, they were consistent. It may be that more proficient speakers actually do slightly better learning from a single talker, but the effect is small and would require more power to reveal it. Such a result would fit in well with past studies showing that changes in talker can make speech processing more difficult (Martin et al., 1989; Mullennix et al., 1989; Ryalls & Pisoni, 1997).

Finally, the benefits of multiple talker training do not seem to taper off first in perception and then in production. Rather, they seem to parallel one another. This parallelism of perception and production is different from the results with children in
Experiment 1, and suggests a close relationship between the two language modalities. These inconstancies and the implications for the perception-production relationship will be discussed further in Chapter 6.
CHAPTER 6

Conclusion
6.1 Summary of main findings

The experiments presented here were designed to extend our understanding of how learning with variation affects the learning of a new category. Specifically, they were designed to test the prediction that the benefits of learning from multiple talkers tapers off with increasing language proficiency.

Past studies have found that variation is beneficial when learning a new speech category (Lively et al., 1993; Bradlow et al., 1997, 1999; Kingston, 2003; Richtsmeier et al., 2009; Rost & McMurray, 2009, 2010). Results from Experiment 1b and Experiments 3a and 3b corroborated this general finding that variation can be beneficial for learning. Given that speech researchers have long viewed variability to be a problem, and that many experiments find multiple talkers to likewise be a source of difficulty in speech processing, this alone is quite remarkable. In the right circumstances, variability can be helpful.

The reasoning for the variability benefit is that variation highlights which cues are relevant for category membership, and which cues are irrelevant. However, all previous multiple talker studies tested a group of people who would not be expected to be proficient in the language used in the experiment. Rost & McMurray (2009, 2010) tested infants, Richtsmeier et al. (2009) tested young children, and Lively et al. (1993); Bradlow et al. (1997, 1999); Kingston (2003) tested L2 learners.

To test the prediction that the multiple talker benefit tapers off with increasing language proficiency, a series of experiments compared less proficient learners with more proficient learners in a word learning task. Unlike learning new phonemes, learning new word representations is something that continues even after
a language’s speech sounds are very familiar to the learner. Thus, an experienced learner could use her knowledge of the language’s speech sounds to quickly form a robust representation of a new word, even without exposure to multiple talkers. From this reasoning, the prediction follows that learning from multiple talkers should be beneficial for less proficient learners, but not for more proficient learners, who are already highly experienced with the sounds of the language.

This prediction was supported by the experiments presented here. English-acquiring children, who are more experienced than infants, but less experienced than adults, benefit to some degree from multiple talkers, showing a benefit in production but not perception (Chapter 3, Experiment 1a and 1b). Native English-speaking adults do not benefit from multiple talkers, in either perception or production (Chapter 4, Experiment 2a and 2b). Finally, second language learners of English benefit from multiple talkers if they are less proficient in English, but not if they are more proficient (Chapter 5, Experiment 3a and 3b). These findings highlight the general idea that the benefits of variability depend on the circumstances. In non-learning speech processing tasks, it is detrimental (Summerfield & Haggard, 1973; Strange et al., 1976; Verbrugge et al., 1976; Assmann et al., 1982; Martin et al., 1989; Mullennix et al., 1989; Mullennix & Pisoni, 1990; Jusczyk et al., 1992; Ryalls & Pisoni, 1997; Houston, 2000; Bent et al., 2010; Lee et al., 2012). In learning tasks, it is helpful only for less proficient learners.

Table 1.1 from Chapter 1 showed the gaps in knowledge, regarding which groups benefit from multiple talker learning. Table 6.1 shows those gaps filled in, with
bolded answers coming from experiments in this dissertation.

<table>
<thead>
<tr>
<th></th>
<th>L1 infants</th>
<th>L1 children</th>
<th>less proficient adults (like L2 speaker)</th>
<th>proficient adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>segment perception</td>
<td>yes</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>segment production</td>
<td>yes</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>word perception</td>
<td>yes</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td>(Experiment 1a)</td>
<td>(Experiment 3a)</td>
<td>(Experiment 2a and 3a)</td>
<td></td>
</tr>
<tr>
<td>word production</td>
<td>NA</td>
<td>yes</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Experiment 3b)</td>
<td>(Experiment 2b and 3b)</td>
</tr>
</tbody>
</table>

Table 6.1: Does variability help?

As expected from past studies, children and less proficient L2 speakers benefit from multiple talkers; although, children were found to benefit only in production and not perception, while the less proficient L2 speakers benefited in both (see Section 6.2 for discussion of this inconsistency). In contrast, adult native speakers and more proficient L2 speakers did not benefit from multiple talkers, and if anything performed better when they learned from a single talker (see Section 5.2.6 for details).

These findings thus suggest an important caveat to previous findings on variability. Learning with high variability is helpful in perception, but if there is a more efficient strategy available to the learner, then it no longer provides any benefit. In this case, a learner who is already very experienced with the variability that occurs between speakers can use this knowledge to form a robust word representation.
through exposure to a single speaker. Only a learner who is less proficient with the language will benefit from variability, because exposure to the variation within a category provides knowledge that a more proficient speaker would already have.

6.2 Perception and production: resolving an inconsistency

The results from Experiment 1, 2, and 3 generally give a consistent picture of when variability, in the form of multiple talkers, helps with learning a word representation. Less proficient speakers benefit from multiple talkers, while more proficient speakers do not. However, there is an inconsistency between Experiment 1 and 3.

Results from Experiment 1a and 1b suggested that the benefits of multiple talker training taper off first in perception, and then in production. Children produced words more accurately when they learned from multiple talkers, but showed no benefit in the perceptual task of mispronunciation detection. This result matches with studies that have found that perception develops before production (Menyuk & Anderson, 1969; Garnica, 1973; Edwards, 1974). From this perspective, the children in Experiment 1 were proficient in English perception, but their production abilities had not yet reached proficiency. Thus, in a perceptual task they would derive no benefit from multiple talkers, similar to the native English-speaking adults in Experiment 2 and the proficient L2 speakers in Experiment 3. And indeed, this is what happened, with no significant difference between children who had learned from multiple talkers and children who had learned from a single talker, when they were detecting mispronunciations. However, while young children may have adult-like perceptual representations, young children’s production abilities are not
mature. Children 2.5-6 years old do not pronounce words perfectly, even words they know. Thus, according to the general prediction that less proficient speakers benefit from learning from multiple talkers, the children are expected to benefit from multiple talkers in a production task, which is what happened in Experiment 1. The difference in multiple talker benefit between a perceptual task and a production task suggests that proficiency in each develops separately, and further that proficiency develops in perception before it develops in production. This view requires representations for perception and production to be different.

In contrast, in Experiment 3 less proficient L2 learners benefited from multiple talkers in perception, and in production, but more proficient L2 learners did not benefit in either. Likewise, in Experiment 2 native adult speakers did not benefit from multiple talkers in either perception or production. Whether or not learners benefited from multiple talkers in one domain implied the same behavior in the other domain. This is similar to previous findings: Bradlow et al. (1999) found that L2 learners improved both their perception and production of the English contrast /r/-/l/, and Yamada et al. (1994) found that if Japanese L2 speakers of English were better at perceiving the contrast between English r-l, they were also more accurate at producing the difference. Previous results, as well as the results in Experiments 2 and 3, thus follow the same pattern, and support a close link between perception and production. There is no evidence from any of these studies for production lagging behind perception.

One explanation for the difference between the children in Experiment 1 and the adult learners in Experiments 2 and 3 is that it is a result of the task. In Experiment
1, children were tested with mispronunciation detection, while in Experiment 2 and 3 the adults were tested with a forced choice task. It may be that the task of mispronunciation detection is more difficult than a forced choice task. If this is the case, the difficulty of the task may have obscured any benefit that the children got from multiple talker training. Thus, children are not actually proficient in either production or perception. They are not adult-like in either, but the perceptual task is so difficult for them that learning from multiple talkers is of no measurable benefit.

Another possible explanation is that the L2 learners are coming to the task with more background knowledge than L1 learners. For example, while an L2 learner might not be proficient in the L2, she is proficient in her native language, and thus knows something about how language works in general. In particular, she is more experienced with how pronunciation is related to acoustics, having had more experience with pronouncing words in her own language. A child, in contrast, is still learning how articulation works on a very general level. If this is the case, it could be that adults have closely linked perception and production representations, while children are still learning the relationship between the two. Children have separate representations for perception and production, while adults have a single representation, or at least tightly linked representations.

There are a few reasons, however, to believe more in the first explanation than the second. For one, if the children who participated in Experiment 1 had little knowledge of how acoustics and pronunciation are related, then it is hard to see how their production accuracy could have benefited from perceptual training. But for
another, children’s performance in the perceptual task was not very good, suggesting that the task was quite difficult for them. While the children performed significantly above chance, they did not perform much above chance, with a mean score of 55.5% correct. If the task had been easier, then children may also have benefited from multiple talker training, like the less proficient L2 learners.

6.3 Implications to theories of speech representation

A major controversy in phonetics exists over how speech sound categories are represented. Sub-lexical units like phonemes are regularly assumed in phonology, but are often disputed in psycholinguistics. Regardless of its psychological reality, the term *phoneme* is a useful way of talking about the patterns of sound in a language. But it is a different thing to postulate that phonemes exist as speech categories independent of words, or that words are regularly parsed into phonemes during language processing.

Port (2007, p. 144) argues against words being stored as phonemes, where phonemes are defined as “abstract symbol tokens that are invariant across context (so [d] is the same in the syllables we spell as [di, de, do], etc.) and the same across speakers.” (see Section 2.2.1) This definition is probably the most extreme possible interpretation of a phoneme. Experiments such as Goldinger (1996) and Goldinger (1998) strongly suggest that, even if such an abstract representation exists, variation is also represented. Thus, this sort of phoneme seems unlikely.

However, the results of these experiments support the existence of something like phonemes - that words may be composed of sub-lexical units of sound, and that
these are stored separately from words. If sub-lexical units exist, then it may be assumed that proficient learners have robust representations of these units, while less proficient learners do not. This would explain why more proficient learners do not benefit from variability, while less proficient learners do. Neither group has ever heard the new words before. Proficient learners may be building a representation of the new word that is robust to changes in talker, by combining sub-lexical units. Thus, they do not need the added experience provided by multiple talkers.

This idea of using familiar sub-lexical units to build new word representations also makes sense in context with results from other studies on the benefits of variability and phonetic training. Lively et al. (1993) found that high variability training with feedback on the English /r/-/l/ contrast not only improved participants abilities on the words they’d heard, but their improvement extended to words not used in training, and even to non-words. That is, the improvement of a particular segmental contrast extended to other instances of those segments, and not just to the words used in training. Similarly, Richtsmeier et al. (2010) found that, when children were exposed to phonotactic sequences in different words and spoken by multiple talkers, they improved their production of these sequences in novel words. Additionally, other studies have found that training on phonemes improved word learning. Perfors & Dunbar (2010) found that phonetic training made learning words easier for second language learners, and Thiessen (2007) found that infants learn similar-sounding words better when they had been pre-exposed to the phonemes distinguishing them, where the phonemes were presented in variable contexts in the word. Again, this suggests it is possible to take knowledge from one
kind of category - phonemes - and apply it to another - word forms.

There is an alternative explanation to the one outlined above: that proficient learners build new word representations out of sub-lexical units. The above explanation assumes parallel representations that are brought to awareness depending on the task. Nonetheless, it assumes that words are at least sometimes parsed into sub-lexical units, and hence the existence of sub-lexical units. However, words could be represented differently. Rather than viewing the lexicon as a sort of list, it may be a multi-dimensional space. In this view, learning words consists of filling in this space. A proficient learner knows more words than a less proficient learner, and thus has a very full lexical space, and thus more constraints on what a word can sound like. A more proficient learner has more information. This also predicts that more proficient learners would not benefit from the extra information provided by multiple talkers.

The experiments presented here are compatible with both explanations of word learning. Another experiment could distinguish between these explanations. If sub-lexical units exist, then proficient learners should be able to use them to build new word representations, even if those words violate the phonotactics of the language. Thus, proficient learners should still not benefit from multiple talkers when learning words that are either phonotactically highly improbable or impossible. However, if the lexicon is a constrained multi-dimensional space, then proficient learners should benefit from multiple talkers, if the words violate those constraints. This is left for future work.
6.3.1 If phonemes exist, when are they used? Perhaps for learning

Even if something like phonemes exist, it does not mean that they are used for all, most, or any speech processing tasks. On the one hand, there is no disputing that literate, adult speakers can talk about, for example, “a word starting with a B”, and hence are at least able to parse words into smaller units. This does not entail that speakers parse out sub-lexical units during speech-processing. Holt & Lotto (2010) note that the evidence for categorizing speech into units during everyday communication is not terribly strong. In particular, Broca’s aphasia patients have difficulty with syllable identification and categorical perception tasks (Blumstein, 1995), while at the same time being able to recognize words (Miceli et al., 1980) and understand conversation (Goodglass et al., 2001).

When then, if ever, do speakers parse speech into phonemes? It seems that, if language users were ever to parse words into smaller units, learning would be the time to do so. Using sub-lexical categories is one way of efficiently learning new, robust word representations. A proficient learner thus can quickly and easily learn new words. Learning to segment speech into words is another situation where sub-lexical units could come in handy. Saffran et al. (1996); Aslin et al. (1998) show that adults and infants can use transitional probabilities of syllables to segment running speech into words. Using transitional probabilities, of course, requires keeping track of sub-lexical units and how often they occur together.

If proficient learners are building words out of sub-lexical units, what do these units look like? As noted by Goldinger & Azuma (2003), many suggestions have been made for the building blocks of speech, including words, syllables, segments,
and phonemic features. The results from these experiments do not tease apart these various possibilities. Previous work in speech perception suggests that any of these levels of representation may be attended to. As suggested by Goldinger & Azuma’s Adaptive Resonance Theory, it may depend on the nature of the task. The results presented here do not support any particular level of sub-lexical representation, but are compatible with any, or all existing in parallel as in ART.

6.4 Final words

The work presented here is the first direct comparison of how multiple talker learning affects more vs. less proficient speakers’ learning of new word representations. Further, both perception and production were measured in all cases. Thus, because the multiple talker benefit tapered off with increasing proficiency, the development of perceptual and production proficiency could be compared. The results agreed with previous results which found a multiple talker benefit in both perception and production (Bradlow et al., 1999; Yamada et al., 1994). Less proficient speakers benefited in both perception and production, while more proficient speakers did not benefit in either: benefit or lack of benefit in one domain implied the same in the other.

These findings also have implications for the representation of words. Linguists have long noticed that language is hierarchical. Sentences can be parsed into phrases, phrases can be parsed into words, words can be parsed into morphemes or syllables, syllables can be parsed into phonemes, and phonemes into phonological features. A distinction can be made, however, between observing Language, and observing how
language users process it. Thus, while there is no doubt that language can be parsed in this way, whether language users regularly parse language into ever smaller units is a separate question.

The original concern of linguists was with the structure of language, and less with how people use it. Accordingly, the psychological reality of one level of the language hierarchy - phonemes - has only more recently come into question. Arguments for and against phonemes were discussed in Section 2.2.1, and various possibilities for how sound categories may be represented was discussed in Section 2.2.1 and Section 6.3. Generally, the picture we have of language as breaking down into ever smaller units has become more nuanced. Language users may not always parse linguistic categories into ever smaller units. Specifically, it seems unlikely that language users parse words into phonemes in all language processing tasks.

If language users were going to parse words into phonemes for any language processing task, learning would be a good time to do so. As discussed previously, one way that proficient speakers could quickly form a new robust word representation, is to use their familiarity with the phonemes of their language. More proficient speakers already know how chunks of speech in their language - phonemes - vary. Because a naive learners does not have robust sub-lexical representations, she cannot use them to build a word, and hence benefits from learning from multiple talkers. In this view of word-form learning, parsing words into smaller “chunks”, and using these chunks to form a word representation, is highly useful, and efficient. Accordingly, if phonemic representations exist, learning a new word representation would be the time to use them.
Finally, the results also support a basic assumption about why variability is helpful for learning. Previous studies with high variability training assume that variation helps with forming a generalizable, and possibly more abstract category representation. The finding that only less proficient learners benefit from multiple talker learning corroborates this assumption. Variation can only help with forming a generalizable representation if a generalizable representation has not already been formed. As mentioned above, proficient speakers have these kinds of representations for phonemes, familiar words, familiar phrases, and potentially many other categories. Therefore when they have to learn a new category, they need not do it from scratch. High variability training is redundant for them. Again, this sort of generalizable representation for various linguistic categories is lacking in a less proficient speaker. Thus, the high variability training is providing them with the information that they need to form a generalizable representation, information that proficient speakers already have.
REFERENCES


Hayes, B. (????). Blick: A phonotactic probability calculator.


