The interaction of lexical frequency and phonetic variation in the perception of accented speech

Marie-Catherine de Marneffe (mcdm@stanford.edu)
Department of Linguistics, Bldg. 460
Stanford University Stanford, CA 94305-2150

John Tomlinson, Jr. (Tomlinsonj2@cardiff.ac.uk)
School of Psychology, Cardiff University
Cardiff, CF10 3AT

Marisa Tice (middyp@stanford.edu)
Department of Linguistics, Bldg. 460, Stanford University
Stanford, CA 94305-2150

Meghan Sumner (sumner@stanford.edu)
Department of Linguistics, Bldg. 460, Stanford University
Stanford, CA 94305-2150

Abstract

How listeners understand spoken words despite massive variation in the speech signal is a central issue for linguistic theory. A recent focus on lexical frequency and specificity has proved fruitful in accounting for this phenomenon. Speech perception, though, is a multi-faceted process and likely incorporates a number of mechanisms to map a variable signal to meaning. We examine a well-established language use factor — lexical frequency — and how this factor is integrated with phonetic variability during the perception of accented speech. We show that an integrated perspective highlights a low-level perceptual mechanism that accounts for the perception of accented speech absent native contrasts, while shedding light on the use of interactive language factors in the perception of spoken words.

Keywords: speech perception, cross-accent perception, lexical frequency, phonetic variation.

Introduction

A single word is produced differently *each time* it is uttered by one speaker. A single speaker naturally produces a wide array of sound tokens that differ greatly in any number of acoustic values — amplitude, F_0, duration, formant transitions, and so on. Each of these acoustically distinct tokens must be understood sometimes as a single sound (within-category), and other times as different sounds (across-category). A central issue is how listeners, oftentimes with no prior experience with a speaker, learn to navigate through this variation to perceive two variants of a token as instances of the same word or two different words. This is particularly challenging considering that minimal differences between words are oftentimes meaningful.

Theories of speech perception that are sensitive to variation, both phonetic and phonological, have typically focused on language use factors to explain how listeners accomplish the task of mapping a variable signal onto meaning (Goldinger, 1996; Johnson, 1997; Newman, Clouse, & Burnham, 2001; Pierrehumbert, 2002). One factor that plays an important role in this mapping task is frequency, or how often a linguistic unit is produced (or experienced by a listener). We know that frequently produced units come with perceptual benefits, e.g., in recognition time, (Dahan, Magnuson, & Tanenhaus, 2001; Forster, 1976; Fox, 1984; Grosjean, 1980; McClelland & Rumelhart, 1981). For example, Fox (1984) found that listeners made more *b* responses to words in a *bad–dad* continuum than to syllables in a *ba–da* continuum when asked to identify the initial sound in the word. He attributed this to a categorization bias toward frequent lexical items, showing that we can get Ganong-like (lexical bias) effects with lexical frequency over and above categorical word/non-word effects. Similar effects have been shown for units smaller than a word (e.g., frequent phonological variants, Connine, 2004; Deelman & Connine, 2001) and units larger than a word (e.g., chunks of commonly co-occurring words, Arnon & Snider, 2010).

Lexical frequency is also a factor shown to influence both the production and perception of spoken words. For example, Jurafsky, Bell, Gregory, and Raymond (2001) showed that frequent words are shorter in duration than infrequent words. Examining the perception of lexically-specific phonological variants, LoCasto and Connine (2002) found that frequent words that typically occur with a reduced vowel in the first syllable (e.g., *police*) are recognized faster and more accurately with a reduced vowel than with a full initial vowel. On the other hand an infrequent word, like *obese*, that is most often produced with a full vowel, is recognized faster and more accurately with a full vowel. Their study thus shows that listeners are sensitive to phonological variant frequency differences across words of a similar phonological shape.

These behavioral patterns are the result of years of experience and exposure with a native language (English). As listeners, we must also map variable speech with less familiar speech patterns, and these speech patterns oftentimes use contrasts that are novel to listeners. To circumvent this issue, a number of studies have examined the perception of regional

and non-native accented speech after large training sessions, providing context for and experience with a novel contrast (Barcroft & Sommers, 2005; Bradlow & Bent, 2008; Lively, Logan, & Pisoni, 1993; Logan, Lively, & Pisoni, 1991; Clark & Garrett, 2004). A robust result stemming from this research is that highly variable training stimuli yield improved learning (where high variability is achieved via multiple speakers).

Factors such as lexical frequency, phonological and phonetic variability, and cross-accent speech perception are oftentimes examined independent of each other. In our everyday linguistic interactions, though, it is not unlikely that these factors are rapidly integrated during speech perception. We believe that further insight into the basic mechanisms underlying the perception of speech can be gained by examining the established effects of lexical frequency in combination with other language use factors — in this case, phonetic variability. In this paper, we examine the perception of non-native accented speech. We focus on the perception of accented speech that is missing a particular native contrast (final voicing contrast), to better understand the roles frequency and phonetic variability have collaboratively in speech perception. We show that listeners rely on both in order to understand spoken words, and this utility surfaces by examining accented speech with novel, unfamiliar contrasts.

Specifically, we show that as lexical frequency increases, listeners increasingly rely on phonetic variability to map sound to meaning. And, when understanding accented speech, an invariant, but representative, token of a word produced in French-accented English is more costly than three variable productions of that word. We consider two accounts of this result — a cost of phonetic invariance and a benefit of phonetic variance.

Final devoicing in French-accented English

Contrast is language-dependent, and phonetic cues signaling contrast may be lost in the speech of a non-native speaker. This loss of contrast may require an adaptation on the part of the listener: the use of a different cue to signal contrast, perhaps ultimately resulting in new mapping strategies. In this paper, we are interested in the loss of the voicing contrast in English words like *bet – bed* when produced by native speakers of French. Both of these productions sound like *bet* to English listeners (Hazan & Boulakia, 1993). Our interest in this loss of contrast is to understand how listeners move from understanding a two instantiations of a word as instances of one word to understanding two instantiations as different words. We believe greater understanding will result from examining the recognition of the variable and invariable acoustic values in words collectively with the influence of lexical frequency. First, we collected productions of voiceless-voiced minimal pairs from native French speakers of English to have a better understanding of both the loss of contrast and the natural phonetic variability that is inherent to voiced final and voiceless final words. The results of this production experiment were then used to create the stimuli for a second experiment, where we examine the joint effect of variability in word production and lexical frequency on the recognition of accented spoken words.

Production experiment

Participants. Eleven native speakers of Belgian French participated in the study (7 female, 4 male). Participants were between the ages of 25 and 61. Exposure to English varied among the participants. In addition, as a native baseline, eleven speakers of American English were recorded with the same stimuli.

Stimuli. Productions were elicited from a reading passage and a word list. Embedded in the passage were fourteen critical pairs differing in the voicing of the word-final stop (e.g., *tack – tag*). Critical items were all monosyllabic. Monosyllabic words provide more control over differences in stress, while offering more minimal pairs for the perceptual experiment than bisyllabic pairs.

Procedure. Participants were asked to read the passage and reading list four times at a comfortable speaking rate. All recordings were made in a quiet room.

Acoustic measurements. Words ending in final stops were extracted from the narrative and the reading list. Words were annotated for vowel onset, vowel offset, and release onset, and were measured with the Praat speech editor (Boersma & Weenink, 2004).

Results. The average vowel duration, final consonant duration, as well as the computed V/C ratio of the rhyme, show gross difference between the native baseline and the French-accented speakers of English. The average duration of the vowels and consonants for voiced-final and voiceless-final words are shown in Figure 1. For the native English speakers, the V/C ratio of voiced-final words is 3.69, and 2.05 for the voiceless-final words. The native French speakers had shorter vowels overall, and nearly identical V/C ratios for the voiced- and voiceless-final words (1.08 vs. 1.06).

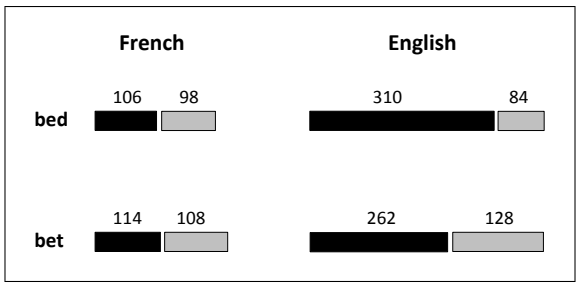

Figure 1: Average vowel (black) and consonant (gray) duration for voiced-final and voiceless-final words for native English speakers and French-accented speakers.

Perception experiment

We examined the effect of within-speaker variability in V/C ratios and the interaction of this variability with lexical frequency in the perception of accented speech. The speech reflects naturally occurring variation in French-accented English, but is very different from the sound patterns of English. The patterns that surfaced from our production study suggest that native English listeners should be biased toward perceiving a voiced-final word (e.g., *tag*) as voiceless (e.g., *tack*). We are specifically interested in whether listeners are more likely to shift away from this bias and correctly identify these tokens as voiced-final. More critical, though, is the potential effect variability (or the lack of variability) on well-established effects of lexical frequency.

We used a click-on paradigm in which participants saw two pictures (e.g., a picture of a *tag* and a *tack*), heard a word, and were asked to click on the picture corresponding to the auditory stimulus.

Participants. Forty-eight native English speakers participated in this experiment. All subjects were undergraduate students at Stanford University. None had any explicit exposure to French or Spanish (in their education, friends, and family).

Stimuli. As critical items for the auditory stimuli, we used twenty monosyllabic words ending in voiced consonants. Each word is a minimal pair to its voiceless counterpart (e.g., *tag – tack*). We used words that are easily identifiable in pictures. Important to our investigation of lexical frequency and variation, we included a range of frequency values for both the produced words and their minimal pair counterparts (voiced-final: 0.16 – 415 msec, voiceless-final: 0.17 – 302 msec). In addition to the 20 critical pairs, we included 60 filler pairs. Forty of the pairs differed word-initially by one sound (*lush – rush*), and twenty pairs differed word-finally by one sound (*moss – moth*), to balance the location of the disambiguating sound. In total, we had 80 auditory items. We recorded all the items produced by a male, native speaker of Belgian French with an obvious accent in English. Target pictures representing each word were used (160 pictures total). The pictures were normalized for color, size of the object in the picture, and mode of representation (i.e., photography, drawing).

Critical manipulations. The critical words were manipulated to form two experimental conditions: one in which we varied the V/C ratio of the words (Variance) and one in which we left the ratio constant (NoVariance). Our manipulations were based on the production patterns of the native French speakers of English in our production study. For each speaker, we computed the average duration for each vowel-consonant pair (e.g., the V/C ratio for /ed/, /eg/, etc.). For some vowels and consonants, the range of averages per speaker was quite large. Taking the average of these averages did not result in a duration close to one that was naturally uttered. Thus to avoid the unintentional use of idiosyncratic patterns, we categorized the vowel and consonant duration averages from each speaker into bins (0.03-0.04, 0.04-0.05, . . . , 0.26-0.27), and took the averages of the values falling in the median bin. Based on these values, we created our stimuli. For each average V/C pair, we created three variants: Lo, Med, High. To do this, we modified the natural tokens recorded by our speaker. The Med stimuli were generated using the median durations of vowel and consonant. For the Lo and Hi variants, we altered the consonant duration using PSOLA in Praat. We lengthened or shorted the consonant durations by half of the standard deviation of the speaker averages resulting in a range of V/C ratios from 0.831 – 2.627 for Lo variants, 0.707 – 1.802 for Med variants, and 0.631 – 1.372 for Hi variants. In the Variance condition, subjects heard variable ratios: the Lo, Med and High versions of each word. The NoVariance condition contained only the Med variant.

Design. The experiment contained three blocks. Each block contained the 80 words, but the presentation order was randomized. In the Variance condition, each variant of a word (Lo, Med, Hi) was randomly assigned to one block. In the NoVariance condition, a Med variant of the words was presented in each block. Participants heard each lexical item three times by the end of the experiment, but in one condition (NoVariance) they heard only the median V/C ratios and in the other condition (Variance), they heard three different V/C ratios for each lexical item.

Procedure. The procedure consisted of a picture familiarization phase followed by the main task. Participants first took a familiarization phase for all 160 target pictures. For each word, they saw a picture and the word the picture represented written next to the picture. Subjects went through this phase at their own pace, clicking on a key to see the next item. This familiarization procedure was used to give the participants an opportunity to see the pictures before the task began. No auditory stimuli were presented during the familiarization phase.

Following familiarization, participants began the main task. Listeners were presented with two target pictures followed by an auditory stimulus 100 msec after the onset of the pictures. They were asked to click on the picture representing the word they heard. Before each trial, the mouse was moved to a neutral prompt in the bottom center of the screen (Freeman & Ambady, 2010). There was 1 second of silence before the next trial began. If no response was made after 3 seconds, the next trial began automatically. Participants responded to all three blocks of 80 pairs. The pictures were 225 x 225 pixels and appeared at either side of the screen. Each picture was 1/5 of the way up the screen and 1/5 the length of the horizontal axis. To assure that participants did not develop a strategy for selecting targets across blocks, the picture presentation order (either left of right) was counterbalanced for

both fillers and targets across the three blocks.

Results. In the regression analyses reported, we excluded two items (*cod* and *cad*) as listeners were unable to recognize any variant as voiced. Examining the reaction times, we removed every data point above 2.5 standard deviations from the overall mean and every data point below 500 ms. Based on these criteria, 10% of the data were excluded from all analyses. To examine the interaction of lexical frequency and phonetic variability, we ran a mixed-effect logistic regression predicting voiced responses on all the subjects (with subject and item as random effects). The condition (Variance or NoVariance), the ratio of lexical frequencies of the words in the pair (FreqRatio), the lexical frequency of the word clicked on by the participant (FreqChoice), and the lexical frequency of the word not chosen (FreqOther) were included as fixed effects. The fixed effects of the model, as well as the interaction between the variance condition and the frequency ratio, are presented in Table 1.

We found significant main effects of the frequency ratio of the words in the pair (voiced word frequency/voiceless word frequency) as well as the frequency of the word clicked on by the subject. Importantly, the interaction between the frequency ratio and the variability condition just missed significance at $p = .057$. One critical factor not to be overlooked in this experiment is the fact that in the variable condition, the Lo items (with a short consonant), by nature have a V/C ratio that is closer to the native English ratio. This alone could bias the participants toward the voiced-final words in the variable condition. To rule out this bias, we ran two regressions, one with only Lo and Med items, and one with only Med and Hi items. The interaction effect goes away with the Lo/Med analysis ($p = 0.118$) suggesting that the variant closest to the natural bias is not driving this effect. Excluding these items, the analysis of Med/Hi items results in a significant interaction ($p < .001$). Figure 2 compares the two conditions (Variance and NoVariance) and shows that variability leads to more percent voiced responses as the frequency ratio between voiced and voiceless words becomes bigger.

Table 1: Experimental fixed effects

Factor	Coefficient	SD	P-value
Variance	0.1098	0.2864	0.7014
Log(FreqRatio)	0.3111	0.0809	0.0001 ***
Log(FreqChoice)	0.2951	0.0548	0.0000 ***
Log(FreqOth)	-0.1358	0.556	0.0146 *
Variance x log(FreqRatio)	-0.1112	0.0585	0.0574

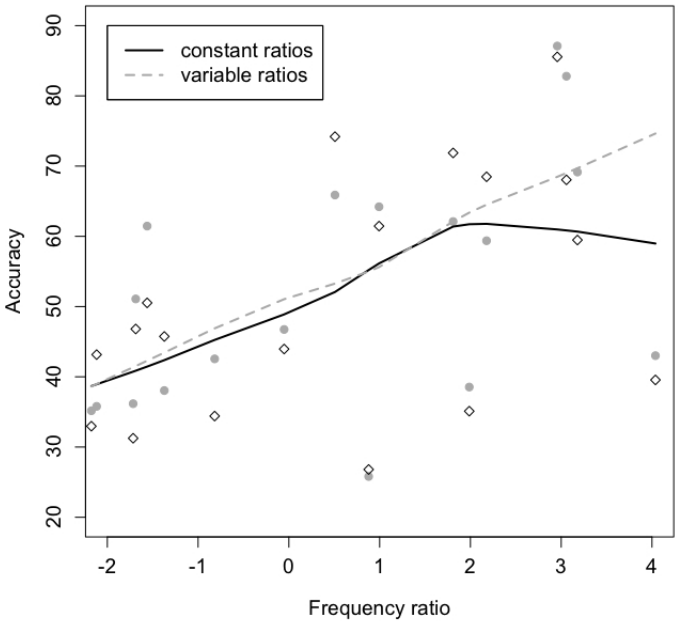

Figure 2: Percent voiced responses dependent on frequency ratios (voiced word/voiceless word).

General discussion

In this paper, we examined the interaction of lexical frequency and phonetic variation in the perception of accented speech. The result, excluding variability, is a Ganong-like pattern in which listeners choose a picture corresponding to a voiced-final word as the lexical frequency of the voiced-final words increases (Fox, 1984; Ganong, 1980). In other words, listeners are biased toward interpreting an ambiguous stimulus as the more frequent word of the pair. More interesting, though, is the split that variability introduces as the frequency of the voiced-final words increases. We consider two potential accounts of this effect, and explore implications for current theory.

Experience in production and perception

One question that arises from this study is whether frequent words are more variable phonetically than infrequent words. Unfortunately, we do not know the answer to this. We do know that frequent function words have more pronunciation variants (Keating, 1998), but whether the standard deviation of a cloud of phonetic variables is larger for frequent words than infrequent words, to our knowledge, has not been explored. This is an important consideration in linking language use factors in production to behaviors in perception. More variable words could, in an experience-based model of speech perception, predict greater perceptual flexibility in the perception of spoken words, as there are more potential candidates matching varied productions.

This idea may go a long way in explaining the increase in voiced responses when variability was introduced in the experiment, but unfortunately falls short in explaining the inability of listeners to identify items as the voiced-final mem-

ber of a pair when they were constantly produced the same way. In any consideration of distributions of values, the Med variant should always be the best. There is no reason a priori for a difference between the two conditions (especially since the interaction is driven by the Med and Hi variants) with this type of explanation. This leads us to ask not why there is a split, but whether the split is indicative of a benefit of variance or a cost of invariance.

Facilitation or inhibition?

Whether variance improves a listener's ability to identify the appropriate item, or invariance worsens this ability is hard to determine from this study. Having said that, we do believe that the answer is both. To support this interpretation, we consider first the NoVariance condition. Listeners are biased by lexical frequency on par with the Variance condition to a point. At the high frequency range, listeners plateau at approximately 60% voiced responses. One possible explanation is that a ceiling exists because on first presentation, listeners initially identified these tokens as voiceless final, and listeners are reluctant to shift from that classification upon subsequent presentations of the same item. If this were the case, why would this only occur to the frequent lexical items? Recalling work by Connine and colleagues, perhaps the reduction typical of frequent words is not present, and listeners shift toward the infrequent variant. This could be viewed as a cost of invariance.

Why then, do listeners continue increasing voiced responses for variable items as lexical frequency increases? Sumner (2011) has recently shown that for accented speech sounds that map onto an unintended category (e.g., unaspirated [p] in Spanish-accented English is great match for English /b/), listeners are able to shift their categorization of that sound only when phonetic variability is infused into the stimuli, outperforming consistent exposure to invariant tokens. The listener behavior here exhibits a similar pattern. Why might variability facilitate identification of frequent voiced-final words? Taken together with the account of inability to shift, the addition of variable items might trigger a low-level matching mechanism in which the presence of a close-enough sound enables a listener to treat a variety of sounds as similar when they otherwise might not. While the interaction was not driven by the Lo tokens, this does not mean that the presence of these tokens had no effect on the perception of the Med/Hi items that might otherwise be categorized as the voiceless member of a pair. On the contrary, we suggest that the presence of these Lo items enabled listeners to interpret the other two variants as similar enough to be mapped to the same lexical item.

Again, though, we need to consider why this is the case with highly frequent words only. The trajectory we see when variability is introduced is what one would expect given the well-established Ganong effect. Listeners are constantly exposed to variable speech, and the presence of this variation does not alter the expected patterns based on lexical frequency. A low-level similarity metric may be used by listeners constantly during speech perception. This pattern is only detectable once we compare regular, variable speech to an invariant, but potentially exemplary token. In other words, the absence of this regular variation results in a reversal from the norm.

Conclusion

Our goal in this paper was to begin to understand the interaction of one major language use factor — lexical frequency —, with another language use factor — phonetic variability — in the perception of accented speech. We show that phonetic variation and lexical frequency interact in speech perception, highlighting both an inability of listeners to shift from one interpretation of a token based on production expectations driven by frequency and the constant reliance of listeners on phonetic variability in the perception of spoken words.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Numbers 0720054 made to Meghan Sumner. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. We appreciate comments and suggestions made by the members of the Stanford Phonetics Lab, and we thank Lily Guo for her help with data collection.

References

Arnon, I., & Snider, N. (2010). More than words: frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67-82.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*, 387-414.

Boersma, P., & Weenink, D. (2004). *Praat: doing phonetics by computer.* (http://www.praat.org.)

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707-729.

Clark, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*, 3647-3658.

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin Review*, *11*, 1084-1089.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 317-367.

Deelman, T., & Connine, C. (2001). Missing information in spoken word recognition: Nonreleased stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 656-663.

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to lan-*

guage mechanisms (p. 257-287). Amsterdam: North Holland.

Fox, R. A. (1984). Effect of lexical status on phonetics categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 526-540.

Freeman, J., & Ambady, N. (2010). Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226-241.

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

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 11661183.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, *45*, 189-195.

Hazan, V. L., & Boulakia, G. (1993). Perception and production of a voicing contrast by French-English bilinguals. *Language and Speech*, *36*, 17-38.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullenix (Eds.), *Talker variability in speech processing* (p. 145-166). San Diego: Academic Press.

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

Keating, P. A. (1998). Word-level phonetic variation in large speech corpora. *ZAS Papers in Linguistics*, *11*, 35-50.

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

LoCasto, P. C., & Connine, C. M. (2002). Rule-governed missing information in spoken word recognition: Schwa vowel deletion. *Perception Psychophysics*, *64*, 208-219.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. Part I: An account of basic findings. *Psychological Review*, *88*, 375-407.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*, 1181-1196.

Pierrehumbert, J. (2002). Word-specific phonetics. In *Laboratory phonology VII* (p. 101-139). Berlin: Mouton de Gruyter.

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *19*, 131-136.