Title

Phonetic category activation drives dimension-based adaptive tuning in speech perception

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 40(0)

Authors

Wu, Yunan Charles
Holt, Lori L

Publication Date

2018

Phonetic category activation drives dimension-based adaptive tuning in speech perception

Yunan Charles Wu (charleswu@cmu.edu)
Department of Psychology, Baker Hall, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Lori L. Holt (loriholt@cmu.edu)
Department of Psychology, Baker Hall, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

Multiple acoustic dimensions contribute to speech categorization. Yet highly diagnostic dimensions contribute greater 'perceptual weight' in influencing speech categorization than less diagnostic dimensions. Recent research demonstrates that perturbations in short-term input regularities lead to rapid dynamic re-weighting of auditory dimensions. Here, we test the hypothesis that phonetic-category-level activation via a highly diagnostic acoustic dimension is critical in driving this rapid tuning of how input maps to phonetic categories. To do so, we manipulate the inherent relative effectiveness, the perceptual weight, of two acoustic dimensions in signaling English vowel categorization using noise-vocoded versus clear speech. We observe that rapid tuning across statistical regularities is affected by dimensions' effectiveness in signaling the vowel categories. These findings indicate that category activation via a highly diagnostic dimension drives adaptive tuning in speech perception, consistent with error-driven supervised learning.

Keywords: Auditory dimension-based statistical learning; Speech perception; Noise vocoding; Perceptual tuning; Vowel categorization; Error-driven learning

Introduction

Everyday speech communication seems effortless, but it presents listeners a series of perceptual challenges in mapping from the acoustic signal to linguistically-relevant categories. One challenge is acoustic variability. The acoustics underlying speech signals are intrinsically variable such that there is no simple mapping from the many possible acoustic realizations to a particular phoneme or word. Even within a single talker's speech, a given word may be realized with variable acoustics, depending on factors such as articulation rate (Miller, Grosjean, & Lomanto, 1984) and coarticulation (Ohman, 1966). This is exacerbated across groups of talkers that vary in gender, age, foreign accent, or dialect (Hillenbrand, Getty, & Clark, 1995; Lee, Potamianos & Narayana, 1999). As an example, local Pittsburghers pronounce their home football team the 'Steelers' with a short /I/ (as in *hill*) rather than a long /i/ (as in *heel*), departing from the mapping of acoustics to vowel categories typical of Standard American English.

Recent research has demonstrated that the perceptual challenge introduced by acoustic variability is met, at least in part, by short-term learning, recalibration, or perceptual tuning of speech categorization according to regularities experienced across short-term speech input. For example, the perception of acoustically-ambiguous speech can be shifted with presentation of disambiguating supportive information, such as lexical information (Kraljic & Samuel, 2006). In the Steelers example above, knowledge that *Steelers* is a word whereas *Stillers* is not can lead more /I/-like speech sounds to be categorized as /i/, even when lexical context is no longer available.

Researchers have described the disambiguating lexical information as a potential 'teacher signal' that may drive adaptive tuning of speech categorization (Norris McQueen, & Cutler, 2003), consistent with supervised error-driven learning (Idemaru & Holt, 2011; Guediche et al. 2013). By this view, the disambiguating information helps to resolve the mapping of the ambiguous speech acoustics to a phonetic category. In doing so, it generates expectations or predictions of input typical of the category. If the ambiguous acoustic input is a poor match with predictions, it may generate a 'mismatch' or 'error' that drives accommodation of the mismatch and leads to subsequent shifts in speech categorization that are apparent even when the disambiguating information is no longer available (see Guediche et al., 2013 for a review).

In the present work, we examine this possibility more closely in the context of adaptive tuning driven by statistical regularities in the input, so-called dimension-based statistical learning (Idemaru & Holt, 2011; 2014; Liu & Holt, 2015). As an example, in Liu and Holt (2015), participants were instructed to respond whether they heard *set* or *sat*, while listening to the stimuli varying across two acoustic dimensions, spectral quality (SQ, related to the pattern of formant frequencies) and vowel duration (DU). In one block of trials, the majority of sounds ('exposure' trials) were sampled from the typical English acoustic space whereby vowels with formant frequencies (SQ) associated with *sat* tend to have longer DU than those associated with *set*. American English listeners tend to give SQ greater 'perceptual weight' than DU (Liu & Holt, 2015). So, for these exposure trials, the most heavily perceptually-weighted acoustic dimension, SQ, unambiguously signals vowel category membership. Moreover, the secondary, DU, dimension is correlated with SQ in a manner consistent with long-term experience with English. In this case, listeners rely upon DU when SQ is perceptually ambiguous on a

small proportion of 'test' trials. In a subsequent block, Liu and Holt (2015) introduced an artificial 'accent' on the same voice. The trials in this block sampled the acoustic space with a reversed correlation between the SQ and DU dimensions that was not typical in English (i.e., vowels with formant frequencies associated with *sat* had shorter durations). As in the first block, listeners' strong reliance on SQ was sufficient to lead to successful vowel categorization for exposure trials. Yet, experiencing the reversed correlation between the acoustic dimensions in short-term input resulted in rapid perceptual tuning such that listeners down-weighted reliance upon DU in vowel categorization. But, when the next block of trials reinstated the typical English SQxDU correlation, listeners rapidly returned to relying upon DU in vowel categorization. In all, this pattern of speech categorization indicates that listeners track the regularities among acoustic dimensions. Moreover, the effectiveness of acoustic dimensions in signaling speech categories is dynamically adjusted according to short-term experience with regularities across dimensions.

Similar to lexically-guided perceptual tuning, this down-weighting of acoustic dimensions' contribution to phonetic categorization may arise from a mismatch between the expected, and actual, acoustic input (Guediche et al., 2013). Yet, there are some differences across paradigms. In the case of dimension-based statistical learning, the acoustic input is not inherently ambiguous. The exposure trials always unambiguously signal a category via the dominant acoustic dimension (SQ in the case above), regardless of the changing correlation between acoustic dimensions across blocks, even upon introduction of the 'accent'. Only the relationship of SQ to the secondary dimension DU shifts across blocks. In this way, dimension-based statistical learning provides an excellent test-bed for addressing the generality of supervised learning approaches to adaptive tuning in speech categorization.

In the present studies, we test a specific hypothesis about the information available to drive adaptive tuning in dimension-based statistical learning. Idemaru and Holt (2011) proposed that the primary, heavily perceptually-weighted acoustic dimension (SQ in the case of Liu and Holt, 2015) may serve as a teacher signal to drive supervised learning in dimension-based statistical learning. Across all blocks, the primary dimension always unambiguously signals a phonetic category. A supervised learning account would predict that category activation via the dominant, unambiguous dimension generates predictions about patterns of input typically associated with the category, including secondary acoustic dimensions. Upon introduction of the accent, the relationship of the secondary dimension, DU, falls out of alignment with predictions potentially generating a mismatch that drives learning. This model leads to a specific prediction: the effectiveness of the dominant dimension in unambiguously signaling category membership should affect the degree of perceptual tuning across acoustic dimensions.

Here, we test this prediction by manipulating the stimuli used by Liu and Holt (2015). We use noise vocoding to reduce spectral resolution, and therefore the effectiveness of the dominant dimension, SQ, in signaling vowel category while preserving DU effectiveness. We expect the relative SQ/DU perceptual weights to shift relative to clear speech. Correspondingly, we predict that DU's greater effectiveness in noise vocoded speech will lead it to serve as a 'teacher signal' that will result in down-weighting SQ -- the opposite pattern observed by Liu and Holt (2015) for clear speech.

To test these predictions, we first characterized the baseline perceptual weights for clear and noise-vocoded speech tokens of *set* and *sat* varying across a two-dimensional acoustic space defined by SQ and DU. With this as a foundation, we predict qualitatively different patterns of tuning will be apparent across noise-vocoded and clear speech as a function of which dimension most effectively activates category representation. We predict that activation of categories by a dominant acoustic dimension is the driving contributor to the perceptual tuning observed in dimension-based statistical learning.

Experiment 1

Methods

Participants Twenty-five Carnegie Mellon University students (18-27 yrs) participated in the study. They received course credit or pay. All reported normal hearing and English as the language used at home since age two.

Stimuli The stimuli were based on those of Liu and Holt (2015). The stimulus space was defined across a SQ (spectral quality) and a DU (duration) dimension with 7 steps along each dimension creating 49 unique stimuli in a two-dimensional acoustic space. This acoustic space served as the basis for two stimulus sets: Clear and Noise-vocoded.

Clear speech tokens were created from natural productions of *set* and *sat* by a female native-English speaker. SQ was manipulated across the steady-state portions of the vowels, spliced from their respective words. The first four formant trajectories were extracted in Praat (Boersma, 2001), and interpolated in equal steps between /ɛ/ and /æ/, then resynthesized to create a 7-step spectral series. Vowel steady-state duration varied from 175 milliseconds to 475 milliseconds. Each of these vowels was then concatenated with the same /s/ and /t/ segments to create the 49-stimulus grid varying from *set* to *sat*.

The noise-vocoded stimulus set was generated from these clear speech tokens, in a manner described previously (Hervais-Adelman et al., 2008; Shannon et al., 1995) using Praat. The frequency spectrum was divided into four logarithmically-spaced analysis bands between 50 and 5500 Hz and clear speech tokens were filtered by these analysis bands. The resulting envelops were applied to band-pass-filtered noise in the same frequency ranges, thereby reducing spectral resolution while preserving duration.

Procedure Seated in front of a computer monitor in a sound-attenuated booth, participants heard speech tokens diotically over headphones and responded whether the stimulus was *set* (Z key) or *sat* (M key) on a standard keyboard. There was a 1-second pause separating trials and no feedback. Visual prompts "SET" and "SAT" aligned with the relative position of the response keys.

All participants first categorized each of the 49 Noise-vocoded speech tokens 8 times. These trials were separated into four 98-trial blocks, between which participants took brief self-timed breaks. Immediately thereafter, they completed the same procedure for the clear stimuli. The order of noise-vocoded speech and clear speech blocks was not counterbalanced because exposure to clear speech can influence perception of noise-vocoded speech (Hervais-Adelman et al., 2008).

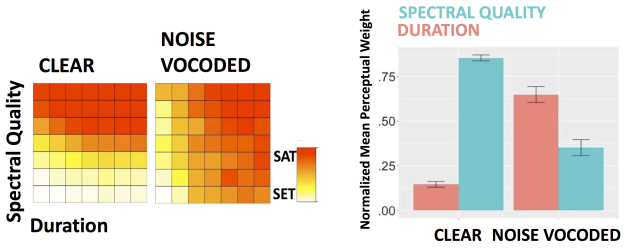

Figure 1. (A) Heat maps for vowel categorization across clear and noise-vocoded stimuli. **(B)** Average normalized perceptual weights across clear and noise-vocoded stimuli; error bars indicate standard error.

Results and Discussion

Figure 1a shows average percent *sat* responses, illustrating that listeners use both SQ and DU and that the weighting of each varies across clear and noise-vocoded speech. To quantify these data, dimension weights were calculated as the beta-coefficients in a regression model including SQ and DU as predictors of category responses for each participant (as in Liu & Holt, 2015). Coefficients were normalized to sum to one. Figure 1b plots these relative dimension weights, averaged across participants. These results replicate confirm American English participants rely more on SQ (M=.85) than DU (M=.15) in categorizing *set* and *sat* in clear speech (Liu & Holt, 2015). Crucially, noise vocoding shifted reliance away from SQ (M=.35) and toward DU (M=.65) in vowel categorization, $F(1, 24) = 56.16$, $p < 0.001$.

In summary, Experiment 1 establishes that manipulating signal quality through noise vocoding shifts the relative informativeness of SQ and DU in vowel categorization. In noise-vocoded speech, DU is the biggest contributor. Notably, the shift in perceptual dimension weighting in categorizing clear versus noise-vocoded speech was evident within-participants. In line with the proposal advanced by Holt and Lotto (2006), the perceptual weighting of the acoustic dimensions appears to have varied as a function of the dimensions' relative resolution in the auditory input. The fact that listeners quickly switched between these listening contexts in the course of the experiment is, itself, a form of rapid adaptive tuning of how input maps to speech categories. We return to this point in the General Discussion.

These results set the stage for Experiment 2. As described above, we expect dimension-based statistical learning of vowels will be qualitatively different across clear and noise-vocoded speech. Whereas in clear speech SQ is dominant and drives category activation, Experiment 1 indicates that DU will play a larger role in noise-vocoded speech categorization. As such, we predict down-weighting of the DU dimension for clear speech and down-weighting of SQ for noise-vocoded speech.

Experiment 2

Methods

Participants Twenty-five Carnegie Mellon University students (18-21 yrs) participated in the study. They received course credit or pay. All reported normal hearing and English as the language used at home since age two.

Stimuli and Procedure The stimuli were sampled from the Experiment 1 stimuli. In a Canonical block, listeners heard 18 exposure stimuli (filled squares, Figure 2) and 4 test stimuli (filled diamonds, triangles, Figure 2) 8 times each in a random order. In this block, the sampling of exposure stimuli reflected long-term English norms: long DU was associated with /æ/-like SQ and short DU was associated with /ɛ/-like SQ. Two of the four test stimuli were distinguished by SQ (diamonds), with perceptually ambiguous DU; the other two test stimuli (triangles) varied in DU with SQ ambiguous. Exposure and test stimuli were intermixed within a block. Test stimuli measured perceptual reliance on SQ or DU.

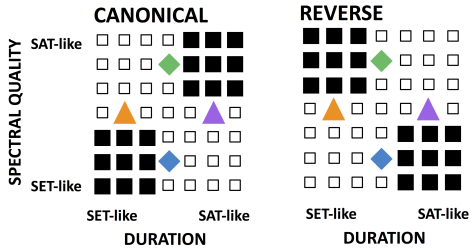

Figure 2. Experiment 2 stimulus distributions for Canonical and Reverse blocks.

The second block of the experiment involved introduction of an artificial 'accent' that reversed the SQxDU correlation typical of English. In this Reverse block, long DU was associated with /ɛ/-like SQ and short DU was associated with /æ/-like SQ. The test stimuli were identical to those of the Canonical block. Participants were not informed of any difference between blocks and simply responded whether they had heard *set* or *sat*.

The Canonical-Reverse block order was repeated for both noise-vocoded and clear speech, with the noise-vocoded condition preceding the clear condition. Overall,

participants experienced 4 blocks separated by self-time breaks. Participants were seated in the same room with the same equipment as in Experiment 1 and were given the same instructions, except that they were informed of the distinction between the clear sounds and 'degraded' sounds that might be very hard to identify.

The relative perceptual weighting differences across clear and noise-vocoded speech observed in Experiment 1 led us to expect different patterns of exposure-stimulus categorization for clear and noise-vocoded stimuli. As shown in Figure 3, reliance on SQ as a dominant dimension in clear speech leads to a different pattern of categorization in the Reverse block than for noise-vocoded speech categorization reliant upon DU. We predict that this difference across conditions will lead to different patterns of dimension-based statistical learning.

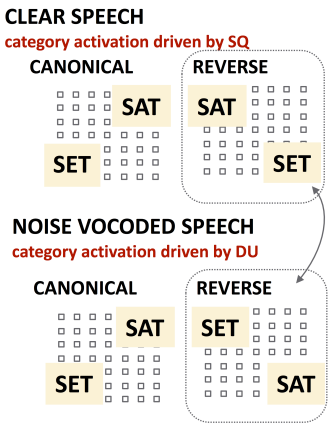

Figure 3. Expected category activation via exposure stimuli, as a function of the perceptual weights observed in Experiment 1.

Results and Discussion

Results are plotted in Figure 4.

Clear Speech We first consider listeners' reliance on the dominant SQ dimension in clear speech (diamonds, Figure 2). There was a significant main effect of SQ, $F(1, 24) = 1419$, $p < 0.0001$, but no interaction between Block and SQ, $F(1, 24) = 0.561$, $p = 0.461$. Replicating prior research, we observed dimension-based statistical learning across the DU dimension (triangles, Figure 2) as an interaction between Block and DU, $F(1, 24) = 56.69$, $p < 0.001$, with a main effect of DU, $F(1, 24) = 55.04$, $p < 0.0001$. These results suggest that as SQ continued to unambiguously signal category activation, listeners down-weighted DU upon introduction of the 'accent' in the Reverse block. The pattern of results is the same with analyses utilizing generalized linear mixed-effects models as a function of DU test stimuli, block, and participant as a random effect, with the response *set* coded as 0 and, *sat* coded as 1 (DUxBlock interaction, $\beta = -2.89$, $SE = 0.34$, $p < 0.0001$. To further investigate this interaction, we constructed models separately for each block examining the effects of different test stimuli on listeners' response. In the Canonical block, listeners differentially categorized DU test stimuli, $\beta = 3.19$, $SE = 0.33$, $p < 0.0001$. They did not in the Reverse block, β

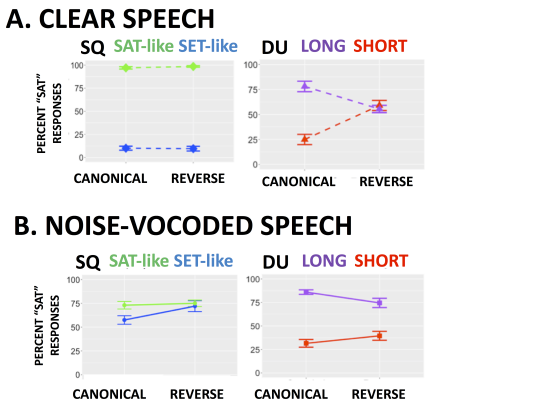

Figure 4. Results of Experiment 2, percent *sat* responses as a function of block and acoustic dimension.

$= -0.16$, $SE = 0.21$, $p = 0.46$.

Noise-Vocoded Speech As predicted, and contrary to the pattern of results for Clear Speech, we observed evidence of dimension-based statistical learning across the SQ dimension (triangles, Figure 2) for noise-vocoded speech. There was a significant interaction between Block and SQ, $F(1, 24) = 5.263$, $p = 0.031$ and also an overall main effect for SQ, $F(1, 24) = 4.377$, $p = 0.0472$. This suggests that in the context of noise-vocoded speech DU may have been sufficient enough to drive category activation. Unexpectedly, there was also evidence of dimension-based statistical learning across DU (Block x DU interaction, $F(1, 24) = 8.289$, $p = 0.01$), with a main effect of DU ($F(1, 24) = 89.58$, $p < 0.001$) as well. It is possible that this pattern of results emerges due to relatively more equal perceptual weightings in the noise-vocoded speech condition, or due to individual differences among participant who adopt different weighting schemes.

This general pattern of results emerges, as well, with generalized linear mixed-effects modeling. There was a marginally significant interaction between Block and SQ ($\beta = -0.60$, $SE = 0.33$, $p = 0.066$). Further investigation into the interaction showed that listeners differentiated the SQ test stimuli in the Canonical block, $\beta = -0.77$, $SE = 0.23$, $p < 0.001$, but not in the Reverse block $\beta = -0.17$, $SE = 0.24$, $p = 0.471$. There was a significant interaction between Block and DU, $\beta = -1.11$, $SE = 0.34$, $p = 0.001$, with significant effects in both the Canonical, $\beta = 2.66$, $SE = 0.27$, $p < 0.0001$, and Reverse, $\beta = 1.63$, $SE = 0.23$, $p < 0.001$, blocks. Interestingly, weighting of DU was also modulated in the paradigm even though it was the primary dimension. We return to this point in the General Discussion.

Overall, the results clearly indicate a difference in how dimension-based statistical learning plays out across clear and noise-vocoded speech. For each, there was a clear pattern of dimension down-weighting on the secondary

perceptual dimension (DU for clear speech, SQ for noise-vocoded speech). Additionally, and unexpectedly, we observed evidence of down-weighting of the *primary* dimension (DU) in the noise-vocoded speech.

General Discussion

By adulthood, the underlying representations for speech categorization reflect the long-term experience in one's native language. But to meet the challenge of acoustic variability in everyday situations, categorization remains flexible such that listeners can adapt to short-term regularities that depart from long-term norms. A critical challenge is understanding how the system flexibly adapts while maintaining long-term representations. There is mounting evidence that listeners use a variety of information sources to adaptively tune speech categorization (e.g., Norris et al. 2003; Bertelson, Vroomen & de Gelder, 2003; Idemaru & Holt, 2011).

Error-driven supervised learning has been explored as a potential contributor to the adaptive tuning (Guediche, et al., 2014; Norris, McQueen & Cutler, 2013; Vroomen, 2007). Under this framework, the mismatch between the expected and actual acoustic output could generate an error message that 'supervises' adaptive tuning of the mapping of speech to long-term representations. Idemaru & Holt (2011) hypothesized that speech categories activated by the primary acoustic dimension could generate expectations about the typical mapping of secondary dimensions. When these expectations are violated, it may generate an error signal that can be used to guide speech tuning via supervised learning.

The current study replicates previous findings on dimension-based statistical learning of vowels (Liu & Holt, 2015) thereby lending evidence that speech categorization is tuned via short-term regularities across acoustic dimensions experienced in the input (Idemaru & Holt, 2011). Additionally, the results are broadly consistent with an error-driven supervised learning account.

Experiment 1 confirmed Liu and Holt's (2015) prior results demonstrating that English listeners' primarily rely upon spectral quality in categorizing *set* versus *sat* in clear speech. Experiment 1 also makes the novel contribution that perceptual weights among the *same* listeners rapidly shift in the context of noise-vocoded speech. This aligns with Holt and Lotto's (2006) prediction that altering the acoustic signal can affect listeners' perceptual weighting. In the current study, using noise-vocoding reduced the spectral fine details of the vowels yet left DU intact. Under this manipulation, perceptual weights shifted from primary reliance on SQ to DU, which became the most informative dimension in vowel categorization in noise-vocoded speech. Note that this, in and of itself, constitutes a kind of rapid adaptive plasticity in how speech input maps to speech categories. It is notable that the same participants, in the same experimental session, rapidly shifted in the manner of mapping from acoustic input to speech categories as a function of signal quality (clear, noise-vocoded).

Experiment 2 demonstrates that this shift in dimensions' perceptual weights has a significant impact on dimension-based statistical learning. In the context of clear speech for which SQ is more effective at signaling category membership, listeners down-weighted reliance on the less-effective, secondary, dimension in response to the reversed dimension correlation in the 'accent.' Reliance on SQ remained stable, even upon introduction of the accent.

Most important to the goals of the present study, the same listeners showed a very different pattern of dimension-based statistical learning in the context of noise-vocoded speech for which DU is more effective at signaling vowel category. In this context, participants down-weighted reliance on SQ, the secondary dimension that was unaffected in the context of experiencing the artificial accent in clear speech. Curiously, there was also evidence of down-weighting for DU, the dominant dimension for noise-vocoded speech. In this regard, it may be important that SQ and DU were relatively more balanced in their contributions to noise-vocoded, compared to clear speech, vowel categorization. Although DU carried greater perceptual weight, its dominance over SQ was less (mean difference relative weight, 0.3) than the dominance of SQ over DU observed for clear speech (mean difference, 0.7). This brings up the possibility that dimension-based statistical learning may track with the reliability of category activation, or graded category activation according to a particular acoustic dimension. Relatedly, the smaller relative advantage for the dominant dimension in noise-vocoded speech may have led listeners to show less 'allegiance' to a particular dimension, perhaps switching over the course of the experiment. If participants (as individuals or as a group) use a mixed strategy of primary reliance on DU and SQ in vowel categorization, then re-weighting may be observed across both dimensions. Future studies could investigate this issue by examining a large cohort of listeners' relationship of baseline cue weights with the magnitude of down-weighting for each dimension.

In all, these studies provide preliminary support for the prediction that category activation via a dominant dimension can drive dimension-based statistical learning across acoustic dimensions. In this way, they are consistent with an error-driven supervised learning account. Other levels of analysis including computational-level Bayesian explanation may be able to account for these data (e.g., Kleinschmidt & Jaeger, 2015). However, a complication of any computational-level account is that it may be implemented mechanistically by multiple means. Here, we find preliminary support for a specific and neutrally-plausible mechanistic instantiation of dimension-based statistical learning in speech categorization. An appealing aspect of this account is that it potentially unites adaptive tuning of speech perception driven by lexical, visual, and acoustic information (Norris et al., 2003; Bertelson et al., 2003; Holt & Idemaru, 2011).

Acknowledgments

This work was funded by the National Institutes of Health (R01DC004674). We would like to thanks Christi Gomez and our undergraduate assistants for their help in data collection.

References

Bertelson, P., Vroomen, J., and de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14,* 592-597.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International 5:9/10*, 341-345.

Guediche, S., Blumstein, S. E., Fiez, J. A., & Holt, L. L. (2013). Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience, 7,* 126.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 460.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical Society of America*, *97*, 3099–3111.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*, 3059–3071.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 1939.

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1009-1021.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434-446.

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

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. The *Journal of the Acoustical Society of America*, *105*, 1455–1468.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 1783.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41, 215– 225.

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

Ohman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39, 151–168.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303.

Vroomen, J., van Linden, S., De Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptationin auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45, 572–577.