INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Do listeners learn better from natural speech?

Michael McAuliffe[1], *Molly Babel*[2], *Charlotte Vaughn*[3]

[1]McGill University, Canada
[2]University of British Columbia, Canada
[3]University of Oregon, USA

michael.mcauliffe@mail.mcgill.ca, molly.babel@ubc.ca, cvaughn@uoregon.edu

Abstract

Perceptual learning of novel pronunciations is a seemingly robust and efficient process for adapting to unfamiliar speech patterns. In this study we compare perceptual learning of /s/ words where a medially occurring /s/ is substituted with /ʃ/, rendering, for example, *castle* as /kæʃl/ instead of /kæsl/. Exposure to the novel pronunciations is presented in the guise of a lexical decision task. Perceptual learning is assessed in a categorization task where listeners are presented with minimal pair continua (e.g., *sock-shock*). Given recent suggestions that perceptual learning may be more robust with natural as opposed to synthesized speech, we compare perceptual learning in groups that either receive natural /s/-to-/ʃ/ words or resynthesized /s/-to-/ʃ/ words. Despite low word endorsement rates in the lexical decision task, both groups of listeners show robust generalization in perceptual learning to the novel minimal pair continua, thereby indicating that at least with high quality resynthesis, perceptual learning in natural and synthesized speech is roughly equivalent.

Index Terms: perceptual learning, adaptation, synthesis, lexical decision

1. Introduction

Spoken language is a highly variable signal. How listeners manage fast and accurate recognition in spite of this – a phenomenon known as perceptual constancy [1] or recognition equivalence [2] – is a critical question in the speech sciences. The variability in speech stems from several sources. While some variability, like that from variation in talker anatomy and physiology [3, 4, 5], appears to be at least superficially easy for listeners to manage (i.e., accurate recognition is easily achieved), other sources of variability like noise, language disorders, and unfamiliar accent present more of a challenge. These latter factors all contribute to an increased difficulty in recognition processes, and thus are often lumped together as adverse listening conditions (see [6] for a review). Listeners, however, readily adapt to these adverse listening conditions and one of the processes involved in adaptation is thought to be perceptual learning. Simply, increased exposure to speech in adverse conditions increases its intelligibility. Researchers have explored perceptual learning in speech from a range of perspectives and have demonstrated the robustness of the behaviour [7, 8, 9, 10].

A classic example of perceptual learning in speech is as follows: listeners hear the word *monsoon* pronounced not with a canonical /s/, but a fricative that is ambiguous between /s/ and /ʃ/: mon?soon. In hearing this novel pronunciation in the context of a known lexical item, listeners learn to expand their /s/ category to include more /ʃ/-like sounds [7, 8].

In perceptual learning studies, some sort of variation or ambiguity is present in the speech signal that a listener needs to learn. The literature has exploited a range of means of introducing such variation. Naturally produced non-native accents [11, 12], manipulated non-native accents [13], synthesized non-existent accents [14], and synthesized sounds within naturally produced native accents [7] have all been used successfully to demonstrate and explore the limits of adaptation and perceptual learning.

Recently, [15] found robust generalization in the perceptual learning of a vowel shift from speech that was naturally produced with category mismatched vowels that were presented as being involved in an unattested back vowel chain shift. That is, the speaker's vowel space was shifted such that any back vowel was pronounced like the adjacent vowel one step higher in F1 (i.e., lower in the vowel space). In a series of experiments, Weatherholtz demonstrated that listeners exposed to part of this shift in a leave-one-out paradigm generalized the entire shift, including vowel pronunciations they had not previously heard. The robustness of this generalization contrasts with the more vowel-specific findings of [14]. Weatherholtz suggested that the patterns of vowel specific learning with a synthesized non-existent accent in [14] could have been due to the quality of the concatenative text-to-speech synthesizer used in that work, which caused listeners to adopt a more vowel-specific listening strategy that may have impaired the generalization of a vowel shift. Listeners do perform differently on a range of tasks when presented with synthetic speech [16, 17, 18]. [19] demonstrated listeners struggle subjectively and objectively in comprehending spectrally degraded synthetic speech compared to unmasked natural speech, in addition to experiencing physiological stress responses when exposed to synthetic speech. Moreover, the physiological responses and increased processing demands recruited to process the synthetic speech were different from the responses to masked natural speech.

As summarized above, the literature on perceptual learning has exploited both natural and synthetic ambiguity to elicit learning, so we know that listeners can and do perceptually learn from natural and synthetically ambiguous spoken language. However, to our knowledge no study has directly compared adaptation and generalization processes in synthesized and naturally-produced speech, and the differences between the Weatherholtz and Maye and colleagues' studies motivate this investigation. Moreover, understanding the differences in how listeners respond to and learn natural and synthetic speech has implications for human-computer interaction and the learning of language for children and adults via online platforms (e.g., Duolingo).

http://dx.doi.org/10.21437/Interspeech.2016-610

To this end we use a lexical decision exposure phase to test whether synthesis affects the perceptual learning of items where the /s/ has been replaced by /ʃ/ – e.g., *castle* is pronounced as /kæʃl/ not /kæsl/. In a between-subjects design, listeners were either exposed to 20 category mismatched /s/-to-/ʃ/ items in their naturally produced forms or after having been run through a synthesizer. To maximize the likelihood of perceiving the mismatched items as words, all items with /s/-to-/ʃ/ had the sound in medial word positions [20, 21].

2. Experiment

2.1. Methods

2.1.1. Materials

Materials were constructed for a lexical decision task and a categorization task. For the lexical decision task, a total of 200 multisyllabic stimuli were presented to listeners, equally divided between word and nonword stimuli. Critical exposure stimuli were 20 words that had /s/ in the onset of the final syllable (henceforth, /s/ or /s/-to-/ʃ/ words); these words are presented in Table 1. The remaining 80 word stimuli and all nonword stimuli contained no sibilant fricatives (/s z ʃ ʒ t͡ʃ d͡ʒ/). The items for the categorization task were six middle steps out of a continuum originally synthesized to have 11 steps for four minimal pairs (*sack-shack, sigh-shy, sin-shin, sock-shock*). For two of these pairs the /s/ word is more frequent, while the /ʃ/ words are more frequent for the other two items [22], thus eliminating any systematic response bias.

Table 1: *Critical items used for perceptual learning in the exposure phase with their standard pronunciations and pronunciations used in the current experiment.*

Word	Standard	Experiment Pronunciation
carousel	kæ.ɹə.sɛl	kæ.ɹə.ʃɛl
castle	kæ.sl	kæ.ʃl
concert	kɑn.sɹt	kɑn.ʃɹt
croissant	kɹə.sɑnt	kɹə.ʃɑnt
currency	kɹ.ɛn.si	kɹ.ɛn.ʃi
cursor	kɹ.sɹ	kɹ.ʃɹ
curtsy	kɹt.si	kɹt.ʃi
dancer	dæn.sɹ	dæn.ʃɹ
dinosaur	daɪ.nо͡ʊ.soɹ	daɪ.nо͡ʊ.ʃoɹ
faucet	fɑ.sɹt	fɑ.ʃɹt
fossil	fɑ.sl	fɑ.ʃl
galaxy	gæ.lɪk.si	gæ.lɪk.ʃi
medicine	mɛ.də.sɪn	mɛ.də.ʃɪn
missile	mɪ.sl	mɪ.ʃl
monsoon	mɑn.sun	mɑn.ʃun
pencil	pɛn.sl	pɛn.ʃl
pharmacy	fɑɹ.mə.si	fɑɹ.mə.ʃi
tassel	tæ.sl	tæ.ʃl
taxi	tæk.si	tæk.ʃi
whistle	wɪ.sl	wɪ.ʃl

The stimuli were recorded by a male native speaker of North American English. The /s/ words were produced twice, once with a normal /s/ and then immediately after with /s/ replaced by /ʃ/. The speaker was instructed to produce both tokens with as similar prosody, speech rate, and style as possible. In the lexical decision exposure phase, all listeners were presented with the /ʃ/ version of the /s/ word, rather than the normally produced /s/ word.

The between-participant manipulation in this experiment was whether listeners were presented with the resynthesized or unchanged natural productions in the exposure phase. Resynthesis was done using STRAIGHT [23], which has been widely used in recent perceptual learning studies (e.g., [24, 20] and others). The stimuli were originally recorded for [20], which used synthesized ambiguous steps of /s/-/ʃ/ continua for perceptual learning exposure. In contrast to [20], the materials used in the current experiment are either natural or resynthesized versions of the natural productions, with no tokens from intermediate continuum steps present in the exposure phase. All tokens in the categorization phase, however, were synthesized items along from the middle of continua, and did not include the original canonical endpoints. Categorization steps were determined in a pre-test [20].

2.1.2. Participants

A total of 50 native speakers of English (27 = female, 22 = male, 1 = gender non-binary, average age = 19.6 years) completed the task for course credit. Twenty-five participants were assigned to each condition. None of the participants reported any uncorrected speech, language, or hearing disorders.

2.1.3. Procedures

Participants completed an exposure task and a categorization task in E-Prime [25]. Exposure was a lexical decision task. Participants heard auditory stimuli and were instructed to respond with either "word" if they thought what they heard was a word or "nonword" if they did not think it was a word. The buttons corresponding to "word" and "nonword" were counterbalanced across participants. Trial order followed the recommendations of [10]. The auditory stimulus was presented 500 ms after the presentation of the word/nonword response options. Participants had 3000 ms from the onset of the auditory stimulus to respond.

In the categorization task, participants were presented with an auditory stimulus and asked to categorize it as one of two words (e.g., *sin* and *shin*), which were presented as options on the response screen. The six most ambiguous steps of the minimal pair continua were used with seven repetitions each. Thus, there was a total of 168 categorization trials.

2.2. Analysis and Results

We first analyze the accuracy and response times in the lexical decision exposure phase before turning to the evidence for perceptual learning from the categorization task.

2.2.1. Exposure phase

Performance in the lexical decision task was high: overall 97% of filler words were correctly identified as words and 90% of nonwords were correctly identified as nonwords. Trials with nonword stimuli and trials with response times below 200 ms and greater than 2500 ms were removed, following previous work. A logistic mixed effects model with accuracy as the dependent variable and Trial, Item (Filler, /s/ words), and Condition (natural, synthesized) as fixed effects. The random effects structure was as maximal as permitted by the data. The model intercept was significant [$\beta = 1.50, SE = 0.22, z = 6.93, p < 0.001$]. There were significant effects of Trial [$\beta = 0.27, SE = 0.09, z = 3.05, p < 0.01$] and Item [$\beta = 5.97, SE = 0.44, z = 13.64, p < 0.001$]. An interaction

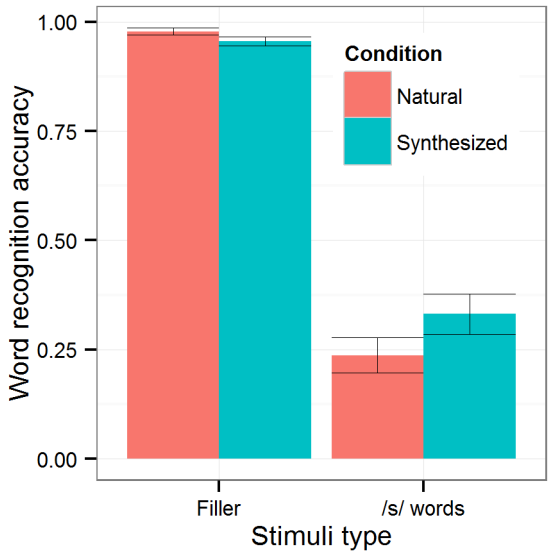

Figure 1: *Word recognition accuracy for filler words and /s/-to-/ʃ/ words in the lexical decision task.*

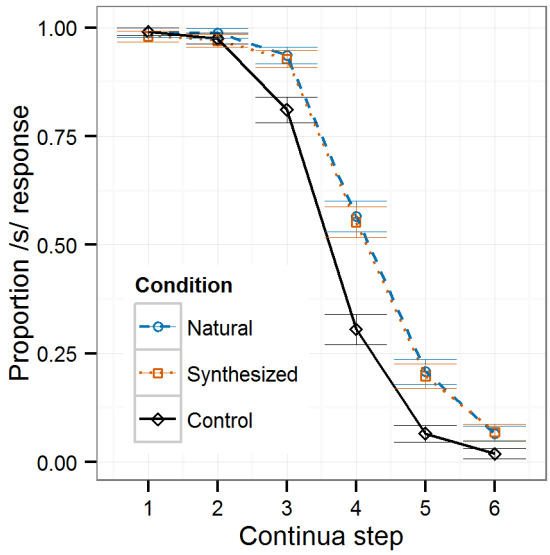

Figure 2: *Proportion of /s/ responses across the continuum for listeners exposed to naturally produced /s/-to-/ʃ/ substitutions, listeners who heard the same items through the STRAIGHT synthesizer, and a control group for comparison. Control group participants are from [20].*

between Item and Condition also surfaced $[\beta = 1.88, SE = 0.62, z = 3.0, p < 0.01]$. These results are shown in Figure 1. Listeners who heard the items that had been passed through the synthesizer were slightly less accurate on filler items and were more likely to endorse the mismatched /s/-to-/ʃ/ items as words.

Response times for word items from the lexical decision task were analyzed in a mixed effects model with centered and log-scaled response times and Trial, Item, and Condition as fixed effects. Like the accuracy model, the random effects structure was as maximal as permitted by the design. The model intercept was significant $[\beta = 0.28, SE = 0.07, t = 3.84, p < 0.001]$. There were significant effects of Trial $[\beta = -0.07, SE = 0.03, t = -2.63, p < 0.01]$ and Item $[\beta = -0.9, SE = 0.09, t = -9.73, p < 0.001]$. Response times increased through the course of the task and listeners were faster on filler items compared to /s/-to-/ʃ/ items. Overall, response times were not significantly slower in the synthesized condition (M = 985 ms, SD = 282 ms) than the natural condition (M = 981 ms, SD = 298 ms).

2.2.2. *Categorization*

A logistic mixed-effects model was used to model subjects' categorization responses (1 = /s/, 0 = /ʃ/). Step, which was centered, and Condition (natural, synthesized) were fixed effects, and the random effects structure was as maximal as the data allowed. The intercept of the model was significant $[\beta = 1.19, SE = 0.28, z = 4.22, p < 0.001]$, indicating that participants overall showed a perceptual learning effect (i.e., participants categorized more of the continua as /s/ than as /ʃ/). There was an effect of Step $[\beta = -2.1, SE = 0.12, z = -17.37, p < 0.001]$; as shown in Figure 2, listeners were more likely to categorize lower (more /s/-like) steps as /s/ words.

The two experimental conditions were also compared directly to a control group who only completed the categorization task in a model where the control group was the reference level. The model intercept was not significant. However, the model revealed an effect of Step $[\beta = -2.51, SE = 0.17, z = -14.7, p < 0.001]$, in addition to Condition effects such that both the natural $[\beta = 1.13, SE = 0.29, z = 3.9, p < 0.001]$ and synthesized $[\beta = 0.91, SE = 0.26, z = 3.51, p < 0.001]$ conditions showed learning compared to control. The interaction between Step and Condition (Synthesized) was significant $[\beta = 0.51, SE = 0.19, z = 2.79, p < 0.01]$, while the interaction between Step and Condition (Natural) was beyond significant $[\beta = 0.30, SE = 0.19, z = 1.59, p = 0.11]$. Listeners in the experimental conditions were more likely to categorize the test continua as /s/-words compared to the control group and this pattern strengthened for the more ambiguous middle steps and the /ʃ/-end of the continua. These results can be seen in Figure 2.

2.2.3. *Performance across tasks*

The kind of perceptual learning under study in this experiment is often called lexically-guided perceptual learning, as it does not occur with nonwords [7]. Therefore, we expect that listeners who categorize more of the /s/-to-/ʃ/ words as words (and not nonword) in the lexical decision exposure task will show more perceptual learning in the post-test. The relationship between /s/-to-/ʃ/ item word endorsement rates and perceptual learning in the post-test is shown in Figure 3 and separated by condition. There is no apparent relationship between these measures.

3. Discussion

This study compared perceptual learning in groups of listeners who heard /s/-words (e.g., *castle*) pronounced with substituted /ʃ/ sounds (e.g., /kæʃl/) which has been either resynthesized or naturally produced. In the lexical decision exposure phase, those who heard the resynthesized tokens were slightly more

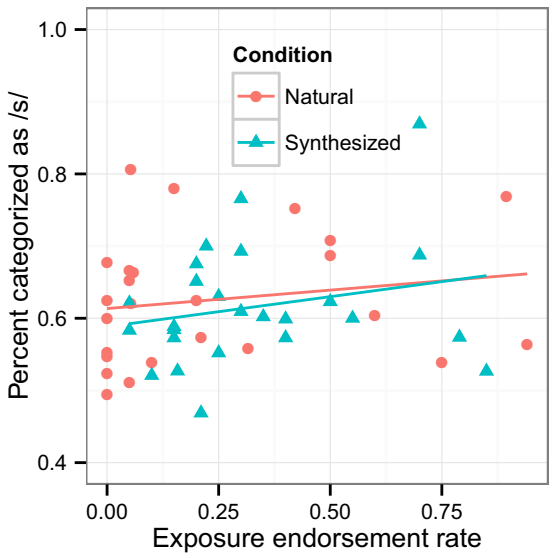

Figure 3: *Relationship between perceptual learning in the categorization task, as measured by proportion of /s/-to-/ʃ/ items categorized as words, and the proportion of items identified as /s/ words in the minimal pair continua in the categorization task. Each dot represents one participant's responses.*

likely to identify the /s/-to-/ʃ/ words as words than those who heard the natural productions. When tested with novel stimuli in a categorization task, both groups of listeners showed evidence of perceptual learning, categorizing more of the /s/ to /ʃ/ continuum as /s/ words, which indicates an increase in the size of the /s/ perceptual category compared to baseline performance of a control group. All together, however, the exposure to the resynthesized speech had minimal effect on perceptual learning as compared with naturally produced speech.

Weatherholtz [15] had hypothesized that he found greater generalization in perceptual learning than [14] because of Maye and colleagues' use of resynthesized speech. It may be the case that lower quality synthesis negatively impacts perceptual learning, whereas the higher quality resynthesis we used in the current study does not. Further work addressing the role of synthesis quality is necessary to disambiguate the current findings. Our future work tests the role of synthesis quality on the robustness of perceptual learning. Additionally, a systematic analysis of the acoustic properties of the synthesized and naturally produced productions of these tokens will be informative about the cues present to listeners in each condition. These results demonstrate, however, that synthesis in and of itself does not attenuate perceptual learning.

Most inquiries into the perceptual learning of consonants use ambiguous pronunciations that fall in between two phonological categories of a language. In this study, we fully substituted /ʃ/ for /s/, similar to what has been done in recent studies of perceptual learning of English vowels [15, 14]. Unlike these studies, however, we exposed listeners to the items in the context of a lexical decision task, as opposed to the continuous narratives used in the mismatched vowel studies. Focusing on perceptual learning of shifting VOT distributions for /p/ and /b/, [13] included a condition that was modelled after naturally produced French-accented English where listeners heard /p/ VOT distributions in the range of 0 to +10ms, which overlaps with the typical pronunciation of /p/ in most varieties of English, and /b/ VOT distributions of -70 to -60 ms, which are strongly pre-voiced tokens. Listeners were exposed to these pronunciations in the context of an accent rating task, where the voice used in the experiment spoken French-accented English, and then tested in a p/b categorization task. Contrary to the mismatched vowel findings and the current results, [13] showed no evidence of perceptual learning from these shifted VOT distributions. A systematic examination of potential differences in perceptual learning performance under a variety of exposure tasks (e.g. lexical decision versus narrative), as well as differences in learning of different categories of segments (e.g. vowels versus consonants), should be taken up in future work.

The word endorsement rates for the /s/-to-/ʃ/ words in this study were low – 29% for the natural productions and 35% for the synthesized productions. This contrasts with the higher endorsement rates of the same words with ambiguous fricatives in other work – 71% [20]. While we cannot directly compare across these studies, it may be the case that the presence of only unambiguous pronunciations in the present study encouraged listeners to adopt a more stringent criterion in the lexical decision task as to what should be considered a word. Regardless of such differences in endorsement rates, however, the magnitude of generalization of perceptual learning across these studies is similar. Earlier work had indicated that listeners do not perceptually learn novel pronunciations from nonword stimuli (eg., [7]), but for listeners in this study, perceptual learning was robust in the absence of strong lexical endorsement. One likely explanation for this is that while listeners did not overtly endorse these items as correct pronunciations of known words, the overall phonetic and phonological similarity to the known words activated these lexical representations, facilitating the updating of the distribution of the /s/ category. In other words, this could be an issue with our interpretation of performance in lexical decision tasks. Lexical decision tasks may tap more meta-linguistic judgments about perceived "correctness" of a pronunciation of an item, as opposed to simply measuring lexical activation. This possibility has interesting implications about the level at which perceptual learning operates – i.e., do post-perceptual biases or strategies affect lexically-guided perceptual learning?

4. Conclusions

In summary, the present study found that listeners demonstrated perceptual learning of fricatives regardless of whether they were exposed to resynthesized or naturally produced tokens. This was true even despite low endorsement rates of words in the lexical decision phase, which we posit may be related to meta-linguistic assessments of pronunciation correctness as opposed to a lack of lexical activation during exposure. These results raise many questions regarding the robustness of perceptual learning with respect to synthetic or naturalistic speech, which we will address in future work.

5. Acknowledgements

The authors would like to thank Matthew Bauer and Brandon Zuel for assisting with subject running.

6. References

[1] D. Shankweiler, W. Strange, and R. R. Verbrugge, "Speech and the problem of perceptual constancy," Hillsdale, NJ, pp. 315–345, 1977.

[2] M. Sumner and R. Kataoka, "Effects of phonetically-cued talker variation on semantic encoding." *The Journal of the Acoustical Society of America*, vol. 134, no. 6, Dec. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25669293

[3] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[4] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *Journal of the acoustical society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[5] B. McMurray and A. Jongman, "What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations," *Psychological Review*, vol. 118, pp. 219–246, 2011.

[6] S. L. Mattys, M. H. Davis, A. R. Bradlow, and S. K. Scott, "Speech recognition in adverse conditions: A review," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 953–978, 2012.

[7] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," vol. 47, no. 2, pp. 204–238, 2003.

[8] T. Kraljic and A. G. Samuel, "Generalization in perceptual learning for speech," *Psychonomic Bulletin & Review*, vol. 13, no. 2, pp. 262–268, Apr. 2006. [Online]. Available: http://link.springer.com/10.3758/BF03193841

[9] F. Eisner and J. M. McQueen, "The specificity of perceptual learning in speech processing." *Perception & psychophysics*, vol. 67, no. 2, pp. 224–238, 2005.

[10] E. Reinisch, A. Weber, and H. Mitterer, "Listeners retune phoneme categories across languages." *Journal of experimental psychology. Human perception and performance*, vol. 39, no. 1, pp. 75–86, Feb. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22545600

[11] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[12] M. J. Witteman, A. Weber, and J. M. McQueen, "Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation," *Attention, Perception, & Psychophysics*, vol. 75, no. 3, pp. 537–556, 2013.

[13] M. Sumner, "The role of variation in the perception of accented speech," *Cognition*, vol. 119, no. 1, pp. 131–136, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.cognition.2010.10.018

[14] J. Maye, R. N. Aslin, and M. K. Tanenhaus, "The weckud wetch of the wast: Lexical adaptation to a novel accent," *Cognitive Science*, vol. 32, no. 3, pp. 543–562, 2008.

[15] K. Weatherholtz, "Perceptual learning of systematic cross-category vowel variation," Ph.D. dissertation, The Ohio State University, 2015.

[16] J. A.-A. Smither, "Short term memory demands in processing synthetic speech by old and young adults," *Behavior and Information Technology*, vol. 12, pp. 330–335, 1993.

[17] S. Lattner, B. Maess, Y. Wang, M. Schauer, K. Alter, and A. D. Friederici, "Dissociation of human and computer voices in the brain: Evidence for a preattentive gestalt-like perception," *Human Brain Mapping*, vol. 20, no. 1, pp. 13–21, 2003.

[18] M. White, R. Rajkumar, K. Ito, and S. R. Speer, "Eye tracking for the online evaluation of prosody in speech synthesis: Not so fast!" in *INTERSPEECH*, 2009, pp. 2523–2526.

[19] A. L. Francis, M. K. MacPherson, B. Chandrasekaran, and A. M. Alvar, "Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort," *Frontiers in Psychology*, vol. 7, 2016.

[20] M. McAuliffe, "Attention and salience in lexically-guided perceptual learning," Ph.D. dissertation, University of British Columbia, 2015.

[21] M. Pitt and C. Szostak, "A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation," *Language and Cognitive Processes*, no. April 2013, pp. 37–41, 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01690965.2011.619370

[22] M. Brysbaert and B. New, "Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English." *Behavior research methods*, vol. 41, no. 4, pp. 977–990, 2009.

[23] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2008, pp. 3933–3936.

[24] E. Reinisch, D. R. Wozny, H. Mitterer, and L. L. Holt, "Phonetic category recalibration: What are the categories?" *Journal of Phonetics*, vol. 45, pp. 91–105, 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S009544701400045X

[25] I. Psychology Software Tools, "E-Prime," 2012. [Online]. Available: http://www.pstnet.com