Listeners are still sensitive to the frequency distribution of cues in degraded speech, but only to the unimpaired ones

Yue Zhang, Stuart Rosen

Department of Speech, Hearing and Phonetic Sciences
University College London
yue.zhang.12@ucl.ac.uk, stuart@phon.ucl.ac.uk

Index Terms: degraded speech recognition, cochlear implant simulation, distributional learning

1. Introduction

Both infants and adults have shown to be able to harness statistical regularities to acquire new visual and auditory knowledge [1][2]. In learning new speech sounds, distributional training, which exposes listeners to a series of stimuli varying in a speech cue that has its frequency distribution following that of a new language, has been found to help learning non-native speech sound categories [3][4]. Studies have also shown that in integrating cues, listeners' reliance on a cue depends on its relative probability distribution for a word or speech category: that is how consistently a cue is correlated with a category [5][6]. The effect of distributional training on adaptation to the speech distorted by cochlear implant (CI) remains unclear. The spectral shift caused by incomplete electrode array insertion has a substantial impact on the spectral information essential to speech perception [7][8][9]. Distributional training on spectral cues might bridge the phonological mismatch between the distorted speech and listeners' representations by forming new phonemic categories. Manipulating the relative probability distribution of speech cues could help listeners to learn and rely on robust but distorted cues. These might improve speech intelligibility and ease listening effort [10][11][12][13][14]. The preliminary study reported here investigated the effect of distributional training on the adaptation to CI simulation (12 bands 4mm upward-shifted noise-vocoded), using tense and lax vowel /i/ and /I/, since they differ in both vowel duration and formant structure [15].

2. Methods and Results

Twenty native British English speakers were randomly assigned to two groups. Four word pairs (bit-beat, sit-seat, pit-peat, fit-feat) were recorded from a male speaker and modeled in the following way: ratios between vowels' F2 and F1 in each context were fitted with a custom distribution that was the sum of 2 Gaussian distributions (bimodal distribution); vowel durations were fitted with a linear regression as a function of F2/F1 ratios. All synthesising were based on these models to approximate in both formant structure and vowel naturalness to the original recording.

To make testing stimuli, for each context, 2 tokens were selected to model the typical dense and lax vowels (F2/F1 ratio 2*sd* away from each mean of the bimodal distribution). Using the 2 tokens, 150 continuous stimuli were synthesised, and among them 6 were selected to be of equal steps in F2/F1 ratio. 6 steps in duration were calculated based on the linear regressions and were paired orthogonally with the 6 steps in F2/F1 ratio, making 144 testing stimuli. Training materials for the Bimodal group were 60 tokens for each 2 contexts, with F2/F1 values of equal intervals in probability densities of the bimodal distribution; for the Uniform group stimuli were 60 tokens of equal intervals in probability densities of the uniform distributions fitted on the original recorded values (see [16] for details). Vowel durations for both groups were changed to the predictions from the linear regressions.

Participants were tested with 4 random lists of BKB sentences, and a cue weighting test on 4 contexts (categorising the word heard as containing tense or lax vowel). Trainings took place on 2 consecutive days, each for one hour. On each trial of the training session, participants chose from 4 words (2 unrelated foils and 1 containing tense/lax vowels in the same context). At the end of the training, the sentence and cue weighting tests were repeated.

For the cue weighting test, significant main effects were frequency steps and training. Significant interactions were: frequency-steps:training, training:word-type:group:duration-steps. The cue weighting pattern was analysed using the ratio between the frequency and duration step coefficients. The main effects of word type (whether words were trained or not) and training, and the interaction of word-type:training:group were significant. For sentence tests, training was significant.

3. Discussion

Significant frequency steps and training interaction indicated that after training participants became more sensitive to frequency cues. However, the lack of interaction with group suggested that exposing to different distribution patterns did not change listeners' reliance on frequency cues. However, Bimodal training made listeners better in using duration cues, but only when words were trained, where they relied less on frequency than duration cues after training. This suggested that for degraded speech, listeners were still sensitive to the statistical distribution of speech cues and adjusted their cue reliance accordingly, but only to cues that were not impaired. Also, this change in the cue weighting did not affect sentence recognition.

The lack of benefits from distributional training on frequency cues might be due to confounds in the experimental design of varying duration cues in both groups and applying passive training strategy [2][17]. They might distract listeners to more salient durational cues. Normal hearing adults could inherently benefit less from the training [18]. Further experiments will amplify spectral cues in training and investigate the training effect for CI users.

4. References

[1] J. Fiser and R. N. Aslin, "Statistical learning of new visual feature combinations by infants," *Proceedings of the National Academy of Sciences,* vol. 99, pp. 15822-15826, 2002.

[2] K. Wanrooij, P. Escudero, and M. E. Raijmakers, "What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning," *Journal of Phonetics*, vol. 41, pp. 307-319, 2013.

[3] E. M. Ingvalson, L. L. Holt, and J. L. McClelland, "Can native Japanese listeners learn to differentiate/r–l/on the basis of F3 onset frequency?," *Bilingualism: Language and Cognition,* vol. 15, pp. 255-274, 2012.

[4] P. Escudero and D. Williams, "Distributional learning has immediate and long-lasting effects," *Cognition,* vol. 133, pp. 408-413, 2014.

[5] M. Clayards, M. K. Tanenhaus, R. N. Aslin, and R. A. Jacobs, "Perception of speech reflects optimal use of probabilistic speech cues," *Cognition,* vol. 108, pp. 804-809, 2008.

[6] J. C. Toscano and B. McMurray, "Cue-integration and context effects in speech: Evidence against speaking-rate normalization," *Attention, Perception, & Psychophysics,* vol. 74, pp. 1284-1301, 2012.

[7] S. Rosen, A. Faulkner, and L. Wilkinson, "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," *The Journal of the Acoustical Society of America,* vol. 106, pp. 3629-3636, 1999.

[8] J. D. Harnsberger, M. A. Svirsky, A. R. Kaiser, D. B. Pisoni, R. Wright, and T. A. Meyer, "Perceptual "vowel spaces" of cochlear implant users: Implications for the study of auditory adaptation to spectral shift," *The Journal of the Acoustical Society of America,* vol. 109, pp. 2135-2145, 2001.

[9] S. Rosen and P. Iverson, "Constructing adequate non-speech analogues: what is special about speech anyway?," *Developmental Science,* vol. 10, pp. 165-168, 2007.

[10] D. Dahan and R. L. Mead, "Context-conditioned generalization in adaptation to distorted speech," *Journal of experimental psychology: Human perception and performance,* vol. 36, p. 704, 2010.

[11] J. C. Toscano and B. McMurray, "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," *Cognitive science,* vol. 34, pp. 434-464, 2010.

[12] M. B. Winn, M. Chatterjee, and W. J. Idsardi, "The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearings)," *The Journal of the Acoustical Society of America,* vol. 131, pp. 1465-1479, 2012.

[13] J. Rönnberg, T. Lunner, A. Zekveld, P. Sörqvist, H. Danielsson, B. Lyxell, et al., "The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances," *Frontiers in systems neuroscience,* vol. 7, 2013.

[14] A. C. Moberly, J. H. Lowenstein, E. Tarr, A. Caldwell-Tarr, D. B. Welling, A. J. Shahin, et al., "Do adults with cochlear implants rely on different acoustic cues for phoneme perception than adults with normal hearing?," *Journal of Speech, Language, and Hearing Research,* vol. 57, pp. 566-582, 2014.

[15] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical society of America,* vol. 97, pp. 3099-3111, 1995.

[16] K. Wanrooij and P. Boersma, "Distributional training of speech sounds can be done with continuous distributions," *The Journal of the Acoustical Society of America,* vol. 133, pp. EL398-EL404, 2013.

[17] J. H. Ong, D. Burnham, and P. Escudero, "Distributional learning of lexical tones: A comparison of attended vs. unattended listening," *PloS one,* vol. 10, 2015.

[18] K. Wanrooij, P. Boersma, and T. L. van Zuijen, "Distributional vowel training is less effective for adults than for infants. A study using the mismatch response," *PLoS one,* vol. 9, 2014.