Short-term exposure enhances perception of both between- and within-category acoustic information

Jessie S. Nixon[1], *Natalie Boll-Avetisyan*[2], *Tomas O. Lentz*[3], *Sandrien van Ommen*[4], *Brigitta Keij*[5], *Çağrı Çöltekin*[1], *Liquan Liu*[6], *Jacolien van Rij*[7]

[1]University of Tübingen, Germany, [2]University of Potsdam, Germany, [3]University of Amsterdam, The Netherlands, [4]LPP, Université Paris Descartes, France, [5]Utrecht University, The Netherlands, [6]Western Sydney University, Australia, [7]University of Groningen, The Netherlands

jessie.nixon@uni-tuebingen.de, nboll@uni-potsdam.de, j.c.van.rij@rug.nl

Abstract

A critical question in speech research is how listeners use non-discrete acoustic cues for discrimination between discrete alternative messages (e.g. words). Previous studies have shown that distributional learning can improve listeners' discrimination of non-native speech sounds. Less is known about effects of training on perception of within-category acoustic detail. The present research investigates adult listeners' perception of and discrimination between lexical tones without training or after a brief training exposure.

Native speakers of German (a language without lexical tone) heard a 13-step pitch continuum of the syllable /li:/. Two different tasks were used to assess sensitivity to acoustic differences on this continuum: a) pitch height estimation and b) AX discrimination. Participants performed these tasks either without exposure or after exposure to a bimodal distribution of the pitch continuum.

The AX discrimination results show that exposure to a bimodal distribution enhanced discrimination at the category boundary (i.e. categorical perception) of high vs. low tones. Interestingly, the pitch estimation task results followed a categorisation (sigmoid) function without exposure, but a linear function after exposure, suggesting estimates became less categorical in this task.

The results suggest that training exposure may enhance not only discrimination between contrastive speech sounds (consistent with previous studies), but also perception of within-category acoustic differences. Different tasks may reveal different skills.

Index Terms: categorical perception, psycho-acoustics, pitch, lexical tone, bimodal exposure, statistical learning

1. Introduction

One crucial question in speech research is how listeners use continuous (non-discrete) acoustic cues to discriminate between discrete alternative messages intended by a speaker (e.g. word meanings). One important mechanism seems to be the language-specific statistical distribution of acoustic cues. Listeners are highly sensitive to the statistical distribution of acoustic cues in the speech signal and brief statistical learning exposure can affect perception and categorisation of speech sounds [1–8]. The present research investigated listeners' statistical learning of a non-native tonal contrast. Specifically, we investigated how exposure to a bimodal distribution of training stimuli affected discrimination of a) lexical tone categories and b) detailed phonetic pitch differences.

Previous studies have shown that listeners are more likely to categorise two sounds as different when trained with a bimodal distribution compared to a unimodal distribution of sounds from a continuum [3–6]. While early accounts proposed that listeners were only able to detect acoustic differences that crossed category boundaries [e.g. 9], more recent evidence has demonstrated that listeners have a remarkable ability to detect fine-grained within-category acoustic information [e.g. 10, 11]. Despite this, statistical learning studies have tended to focus on effects of training on categorisation behaviour. Few studies have investigated effects of exposure to non-native speech sounds on perception of within-category acoustic information.

Although many of the world's languages are tonal, the majority of studies on statistical learning have focused on segmental phenomena in non-tonal, Indo-European languages. Similar studies on statistical learning of lexical tone are still scarce and most of these have investigated contour tones [2, 4, 5]. Two recent studies have used eye movements to investigate statistical learning of level tones by native [8] and non-native listeners [12]. The present study is, however, to the best of our knowledge, the first to study effects of exposure on both discrimination of lexical tone and fine-grained acoustic information by using both an AX discrimination task and a pitch height estimation task for testing.

A number of Chinese dialects (e.g. Cantonese, Southern Min) contain level tones, between which the main discriminative cue is pitch height. The present stimuli were based on Cantonese high-level and mid-level lexical tones [13, 14]. Discriminating between tones can present a challenge for native speakers of non-tonal languages beginning to learn a tonal language. Pitch does not discriminate between word meanings in non-tonal languages. Therefore, speakers have had a lifetime of experience learning to ignore pitch as a discriminative word-level cue during speech perception [15]. This provides an opportunity to test the effects of short-term exposure on learning to discriminate a non-native speech cue.

The present study investigated how listeners of a non-tonal language perceive the pitch cue in a level tone contrast and how exposure to a bimodal distribution affects perception. Based on previous statistical learning studies [2, 4, 5], we predicted that participants would be better at detecting differences across category boundaries after training exposure than without training exposure (AX discrimination task). Secondly, we predicted that perception of tokens on the continuum would be more 'categorical', or cluster together more into two groups, following training, compared to without training (pitch height estimation task).

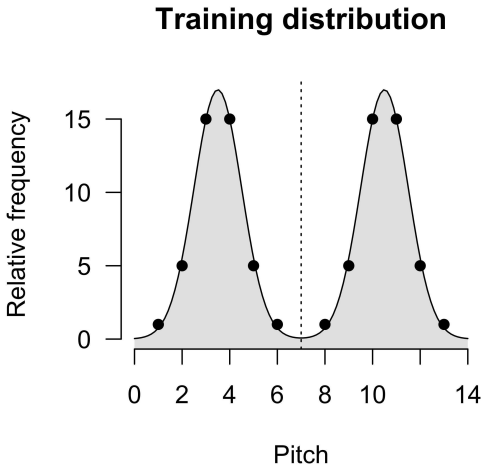

Figure 1: *Relative presentation frequency of the 13 steps on the pitch continuum during the training phase. The boundary between distributions was defined to be at step 7, which was not present in the training exposure.*

2. Method

2.1. Participants

Participants were 39 native speakers of German. Twenty participants received short-term training exposure prior to testing (exposure group). Nineteen participants received no training exposure prior to testing (no-exposure group). Testing involved two tasks: an AX discrimination task and a pitch height estimation task. In the no-exposure group, three participants completed only the AX task and two only the pitch height task.

2.2. Stimuli

We recorded the syllable /li:/ (duration = 285 ms), produced with a flat contour (level tone, F0 = 258 Hz) by a female speaker with phonetics training. The CV syllable /li/ was selected because it is meaningless in German and sonorant /l/ is a good carrier of prosodic information. The syllable was then manipulated (in Praat [16]) to have a level pitch on the vowel. A 13-step pitch continuum was then created from the syllable, ranging from 260 Hz to 287 Hz with steps of 0.14 semitones. Semitones take into account that perceptual differences between tones are not linear.

2.3. Procedure

Participants were tested either without prior exposure (no-exposure group) or after exposure (exposure group) to a bimodal distribution of the 13 tokens of the-pitch continuum (Figure 1). An oddball design was used, in which a series of four tokens from one tone (e.g. mid tone) were presented, followed by one token of the other tone (e.g. high tone). EEG was recorded as the exposure group listened to eight blocks of 42 such trials, each containing five tokens, sampled from the two Gaussian distributions in Figure 1. The exposure phase lasted approximately 25–30 minutes. We will discuss the EEG data in a separate publication.

Participants sat at a laptop in a quiet room (no exposure group) or a desktop computer in a sound-attenuated booth (exposure group). Stimuli were presented over Sennheiser HD 280 pro, 64 Ω headphones. Participants received a chocolate bar (no exposure group) or payment (exposure group) for their participation.

2.3.1. AX discrimination task

The AX discrimination task tested whether participants were sensitive to the difference between two similar pitch values and whether their sensitivity was affected by the position on the pitch continuum. Participants were presented with two sounds (AX pairs) and asked to decide whether the sounds were the same or different. The two sounds always differed by two steps on the continuum (e.g. AX = 1–3, or 8–6; i.e. 0.28 semitones). There were 88 trials for the exposure group and 44 trials for the no-exposure group. The number of trials presented to the exposure group was kept to a minimum to reduce potential effects on the following pitch height estimation task of 'unlearning' caused by the flat distribution. Trials were divided into blocks, with 11 AX pairs tested twice within each block, once with the lower tone presented before the higher tone (direction 'up'), and once with the higher tone presented before the lower tone (direction 'down'). Within each block, the order of trials was randomised for each participant.

2.3.2. Pitch height estimation task

The pitch height estimation task tested participants' ability to perceive and identify the relative pitch height of the syllables along the continuum. Before the task began, participants heard each of the 13 sounds of the continuum played once in random order. At test, on each trial, participants heard one token from the continuum. Their task was to place a visually presented slider on a vertical bar with the mouse to indicate their estimate of the pitch height (with the top reflecting highest pitch and the bottom lowest pitch). There were 52 trials organised into 4 blocks. In each of the 4 blocks, the 13 different continuum steps were presented once each, with the order randomised for each participant. The position of the slider (in pixels) was recorded when the participant released the mouse button.

3. Analysis and Results

3.1. Data and analysis

Generalised Additive Mixed Modeling (GAMM) [17,18] as implemented in the R package `mgcv` version 1.8-22 [18, 19] was used to analyse the data. GAMM is a mixed effects regression method that does not assume a linear relationship between dependent and independent variables. This makes it suitable for modeling the hypothesised non-linear relationship between participants' discrimination or pitch estimation decisions on the one hand and the pitch continuum step on the other. In addition, it allows for including (non-linear) random effects to account for variability between participants. The R package `itsadug` version 2.3.2 [20] was used for visualisation and interpretation of the model predictions. Note that, in addition to model comparisons and model summary statistics, visualisations play an integral role in significance testing of GAMM models.

3.1.1. Discrimination task

Participants' decisions ('same', 1 vs. 'different', 0) were modelled with a logistic GAMM. The effect of training exposure was tested with a two-level factor (exposure vs. no exposure). To test for the effect of the position on the pitch continuum, a continuous predictor of pitch was included, which was the po-

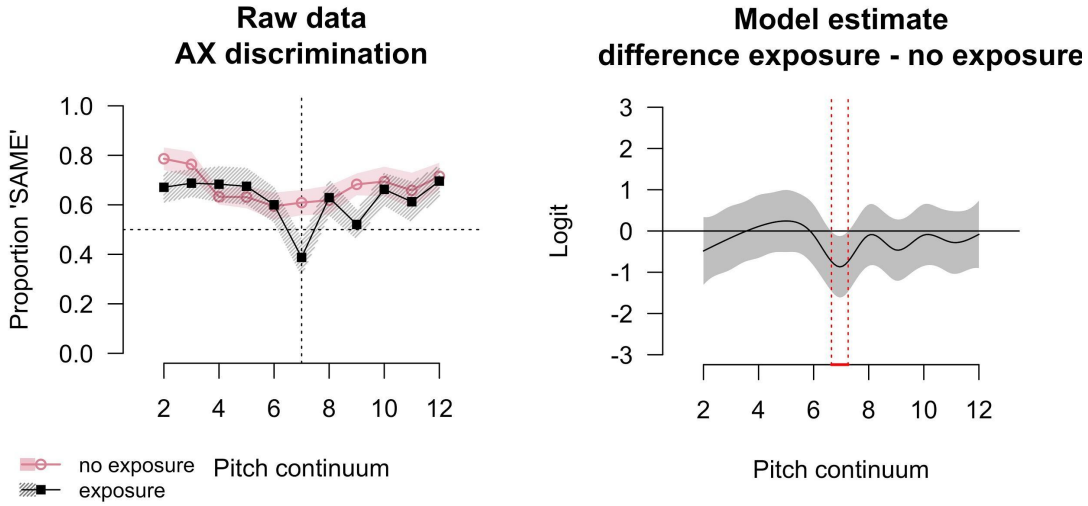

Figure 2: *Left: Raw data from the AX discrimination task in the no-exposure (red circles) and exposure groups (black squares). The pitch step is on the x-axis. The proportion of 'same' responses is on the y-axis. Right: GAMM model estimate of the difference in responses between the two groups on logit scale. Vertical dashed lines indicate where the lines significantly diverge. (The divergence is only significant at pitch step 7).*

sition between the two pitch values being discriminated (e.g. 7 for the tokens 6–8 and 8–6). Shrunk factor smooths (nonlinear random effects) for pitch value per participant and trial per participant were included to account for differences between participants. The smoothing parameter estimation method was set to the default (UBRE).

3.1.2. Pitch height estimation task

The pitch height estimation data (the y-position on the screen) was modelled with Gaussian GAMMs. The effect of training exposure was tested with a two-level factor (exposure vs. no exposure). To test for the effect of the position on the pitch continuum, a continuous predictor of pitch value was included. Shrunk factor smooths for participant and trial per participant were included to account for differences between participants. The slider y-position was scaled between 0 and 1 for the plots. The smoothing parameter estimation method was set to "ML".

3.2. Results

3.2.1. Discrimination task

A visualisation of the raw data for the AX discrimination task is shown in the left panel of Figure 2. The data suggest that participants most often perceived the pairs as the same. However, there seems to be a non-linear effect of pitch value. The no-exposure group (no-fill red circles) shows a gradual downward slope from the outer values to the central values. For the exposure group (solid black squares), the data shows a sharp downward peak at the central point (pitch 7). That is, the proportion of 'same' responses to stimuli steps 6 and 8 appears to be lower after exposure.

The right panel of Figure 2 shows the model estimate of the difference between groups. Vertical dashed lines indicate where the lines significantly diverge. The model plot shows that participants' performance in the AX discrimination task did not differ between exposure and no-exposure group for the outer values on the continuum. However, at position 7 (i.e. when participants heard 6–8 or 8–6 trials), the exposure group gave significantly fewer 'same' responses than the no exposure group. This difference was confirmed by model comparisons (i.e., the model with interaction between pitch and training exposure included had lower AIC value than the model without interaction; ΔAIC = 2.594) and by a pointwise permutation test (at position 7, the proportion of 'same' responses was significantly lower for the exposure group than for the no-exposure group; n=10000, z=3.178, p=0.014).

3.2.2. Pitch height estimation task

The raw average pitch height estimates per pitch value from the no-exposure (red circles) and exposure group (black squares) are shown in the left panel of Figure 3. The no-exposure group data suggests a steeper slope near the central pitch values (6–9) and a relatively flat effect of pitch near the edges of the continuum (1–5 and 10–13). This contrasts with the exposure group, for which the effect of pitch appears linear. Note that is is the opposite of the expected direction of effects, in that responses appear less rather than more clustered after training.

The right panel of Figure 3 shows the model estimates for the exposure and no-exposure groups in the pitch height estimation task. In the no-exposure group (red dashed line) pitch estimation follows a subtle sigmoid shape: the slope is flatter towards the edges of the pitch distribution and becomes steeper around the central values. This shape corresponds to the categorisation function that occurs in category discrimination tasks - although it is weaker than typical categorisation results. In contrast, in the exposure group (solid black line), this nonlinearity is absent: the model summary shows a linear effect of pitch in the exposure group. While the difference between the exposure and no-exposure group model did not reach significance, the estimated degrees of freedom for the model terms (edf) for the regression line fitting the exposure group data as 1.00 (i.e. a straight line), and the edf for the regression line fitting the non-exposure group data as 4.62 (i.e. a wiggly line). Model comparisons suggested that the data was significantly better accounted for when a non-linear smooth was allowed than when only linear regression lines were included (χ^2(4.0)=373.1; p<.001; ΔAIC=-838.30). These results indicate that the nonlinear pattern that occurred in the non-exposure group data was absent

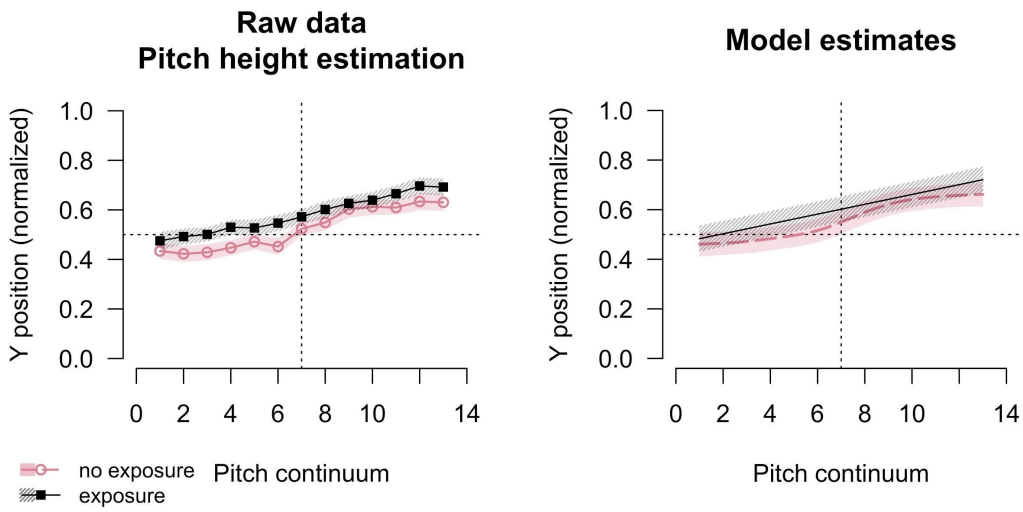

Figure 3: *Left: Plot of the raw data from the pitch estimation task for the exposure (black squares) and no-exposure groups (red circles). The pitch continuum is on the x-axis; the pitch estimate (y-position on the screen) is on the y-axis. Right: The GAMM model estimates of the pitch height estimation responses in the no-exposure group (dashed red line) and exposure group (solid black line). Summed effects (incl. intercept), random effects are excluded.*

after exposure.

4. Discussion

The goal of the present study was to investigate the effects of short-term bimodal training on discrimination and identification of pitch in two non-native level tones. Perception was assessed by means of two tasks: (1) AX discrimination and (2) pitch height estimation. Two groups of participants were tested: one with bimodal training exposure to the test stimuli and the other with no training exposure. Based on previous literature [2, 4, 5], we expected perception to become more categorical after training.

Results of the two tasks seem to tap into different effects of the training. The AX discrimination task provided some evidence in support of this prediction. For this forced choice task, discrimination of tokens near the category boundary of the high and low tones was enhanced following training exposure, compared to the no-exposure group. This was evidenced by a sharp discrimination peak - fewer 'same' responses - at the category boundary in the exposure group. This is consistent with previous studies [2, 4, 5] and extends this finding to include level tones, discriminated by pitch height.

The pitch height estimation task investigated effects of training on the ability to place the acoustic cue (pitch) of each token relative to the other tokens on the continuum. Interestingly, the results of this task were the *opposite* of our predictions. That is, estimates of pitch height were more evenly spread, with less clustering into two groups, following training, compared to the no-exposure group. Without exposure, there was a nonlinear relationship between pitch step and participants' pitch estimates, indicating grouping of tokens into two (high and low) clusters. In contrast, estimates of the exposure participants were linear, indicating improved ability to identify the tokens' pitch relative to the other tokens.

The ability to discriminate speech sounds is often seen as being in contrast with the ability to detect within-category acoustic detail [9, 21]. For instance, while infants of a few months of age are able to detect changes between many non-native speech categories, this ability seems to decrease with age as experience with the native language grows [21]. However, counter to expectations, exposure seemed to improve both skills simultaneously in the present study. In same-different decision task, participants who had been exposed to a bimodal distribution of tokens showed an increase in 'different' responses near the distribution boundary. However, in the pitch estimation task, which requires identification of the position of the current token relative to the whole distribution, they show an increased capability to perceive pitch differences, even within the categories.

Nonlinear sensitivity to within-category phonetic detail has previously been demonstrated for native listeners [8, 10]. Given the probabilistic and predictive nature of speech comprehension in which cues differ with context, speaker and so on [22, 23], better discrimination of small acoustic differences within a cue dimension has advantages. A recent visual world eyetracking study tracked native English speakers' emergence of nonlinear sensitivity to pitch with bimodal distributional exposure [12]. The study found that cue sensitivity depended on the statistical variance in the stimuli. The present results suggest that between- and within-category perceptual learning may be related in a different way than previously thought. Potentially, within- and between-category sensitivity are two manifestations of learning: more accurate identification of the relevant acoustic cue - in this case pitch - may enhance discrimination of lexical contrasts.

5. Acknowledgements

We would like to thank the organisers of the "Modelling meets Infant Studies in Language Acquisition: A Dialogue on Current Challenges and Future Directions" Workshop (2013), Lorentz-Center, Leiden, The Netherlands. This paper is part of a collaborative project that started at this workshop. We would also like to thank Harald Baayen for making the Quantitative Linguistics Group EEG lab available for this research. This research was supported by a Research Networking grant (ESF) NetwordS No. 6609 to NB and a Leiden University AMT Individual Researcher Grant to JSN.

6. References

[1] D. B. Pisoni, R. N. Aslin, A. J. Perey, and B. L. Hennessy, "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants." *Journal of Experimental Psychology: Human perception and performance*, vol. 8, no. 2, p. 297, 1982.

[2] P. A. Hallé, Y.-C. Chang, and C. T. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *Journal of phonetics*, vol. 32, no. 3, pp. 395–421, 2004.

[3] 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, no. 5, pp. EL398–EL404, 2013.

[4] J. H. Ong, D. Burnham, and P. Escudero, "Distributional learning of lexical tones: A comparison of attended vs. unattended listening," *PloS one*, vol. 10, no. 7, p. e0133446, 2015.

[5] L. Liu and R. Kager, "Perception of tones by infants learning a non-tone language," *Cognition*, vol. 133, no. 2, pp. 385–394, 2014.

[6] 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, no. 10, p. e109806, 2014.

[7] P. Escudero, T. Benders, and K. Wanrooij, "Enhanced bimodal distributions facilitate the learning of second language vowels," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL206–EL212, 2011.

[8] J. S. Nixon, J. van Rij, P. Mok, R. H. Baayen, and Y. Chen, "The temporal dynamics of perceptual uncertainty: eye movement evidence from Cantonese segment and tone perception," *Journal of Memory and Language*, vol. 90, pp. 103–125, 2016.

[9] M. Schouten and A. J. van Hessen, "Modeling phoneme perception. I: Categorical perception," *The Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 1841–1855, 1992.

[10] B. McMurray, M. K. Tanenhaus, and R. N. Aslin, "Gradient effects of within-category phonetic variation on lexical access," *Cognition*, vol. 86, no. 2, pp. B33–B42, 2002.

[11] ——, "Within-category vot affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition," *Journal of memory and language*, vol. 60, no. 1, pp. 65–91, 2009.

[12] J. S. Nixon and C. T. Best, "Acoustic cue variability affects eye movement behaviour during non-native speech perception: a GAMM model," in *Proceedings of Speech Prosody 9*, 2018, p. (accepted).

[13] R. S. Bauer and P. K. Benedict, *Modern Cantonese phonology*. Walter de Gruyter, 1997, vol. 102.

[14] P. P.-K. Mok and P. W.-Y. Wong, "Production and perception of the rising tones in Hong Kong Cantonese," in *The 9th phonetics conference of China, Tianjin*, 2010.

[15] M. Ramscar, E. Suh, and M. Dye, "For the price of a song: How pitch category learning comes at a cost to absolute frequency representations," in *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, vol. 33, no. 33, 2011.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program] version 5.3.74, retrieved 30 april 2014, from http://www.praat.org," 2014.

[17] T. Hastie and R. Tibshirani, *Generalized additive models*. CRC Press, 1990.

[18] S. N. Wood, N. Pya, and B. Säfken, "Smoothing parameter and model selection for general smooth models," *Journal of the American Statistical Association*, vol. 111, no. 516, pp. 1548–1563, 2016.

[19] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed., ser. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2017.

[20] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2017, R package version 2.3.

[21] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant Behavior and Development*, vol. 7, no. 1, p. 49–63, 1984.

[22] J. S. Allen and J. L. Miller, "Listener sensitivity to individual talker differences in voice-onset-time," *The Journal of the Acoustical Society of America*, vol. 115, no. 6, pp. 3171–3183, 2004.

[23] T. Kraljic and A. G. Samuel, "Perceptual adjustments to multiple speakers," *Journal of Memory and Language*, vol. 56, no. 1, pp. 1–15, 2007.