Tonal Aspects of Languages 2016
24-27 May 2016, Buffalo, New York

The effects of talker variability and variances on incidental learning of lexical tones

Jiang Liu [1], *Jie Zhang* [2]

[1] Department of Asian Languages and Literatures, University of Minnesota, USA
[2] Department of Linguistics, University of Kansas, USA

liux2795@umn.edu, zhang@ku.edu

Abstract

Multi-talker variability has been found to be very effective in the perception and production training of nonnative sound categories in the past few decades. The phonetic training paradigms were mostly explicit learning in which learners received feedback of the categories when exposed to the training stimuli. More recently, studies have started to investigate how auditory categories are learned by adults incidentally during unsupervised training—a simulation of sound category learning in a natural environment without any experimenter-provided feedback. The stimuli used in incidental learning were mostly re-synthesized nonspeech sound categories due to the ease of manipulating the variance of different acoustic dimensions. Very few studies examined the effect of talker variability on incidental learning of novel speech categories. This study investigated whether American adults without any tone language experience can learn Mandarin Chinese lexical tones incidentally by playing a video game. We also examined the effects of carefully manipulated variability of the re-synthesized stimuli and the natural variability of multi-talkers on lexical tone category learning. In addition to tone discrimination and identification, we also examined the participants' cue-weighting change after the incidental learning. The result showed that novel speech categories, lexical tones in this case, can be learned incidentally. The results also showed that multi-talker stimuli not only led to better generalization for the identification of tones in stimuli not present in the training but also made learners have a more nativelike cue-weighting in tone perception. The results suggest that the manipulation of variance on significant acoustic dimensions such as pitch direction and height may not be as robust as talker variability in terms of learning lexical tones when incidental learning occurred.

Index Terms: phonetic training, lexical tones, incidental learning, acoustic variance, talker variability.

1. Introduction

In the study of nonnative/L2 sound categorization, non-native speakers do not always perceive L2 sound categories in the same way as the native speakers do. For example, Japanese speakers depend primarily on F2 rather than F3 for distinguishing English /r/ and /l/ whereas English speakers depend primarily on F3 for the /r/ and /l/ distinction [1]. For tone perception, American English speakers depend more on pitch height (average pitch) whereas native speakers of Mandarin Chinese depend more on pitch direction [2]. Based on these cross-linguistic perception studies, researchers developed phonetic training paradigms that aimed at training L2 learners to have more nativelike perception (e.g., [3], [4], [5]).This body of research has shown evidence of plasticity in the adult system to support non-native sound category learning even though the system is not as flexible as in earlier development [6]. As learners' L2 sound categorization improved, they learned to discriminate and perceptually weigh linguistically significant acoustic dimensions and to generalize across within-category acoustic variability in speech [7]. Different methods have been used to train learners to weigh more on linguistically relevant acoustic (see [7] for an overview). [8] found that by manipulating the variance on different acoustic dimensions, it is possible to shift listeners' cue-weighting from one acoustic dimension to another within a relatively short training period. The method of variance manipulation had been applied to the training of Japanese listeners' perception of L2 English /r/ and /l/ [9]. The result showed that making the variance on F2 larger than the one on F3 helped Japanese speakers shift cue-weights towards F3, which is the primary acoustic cue native English speakers use for the /r/ and /l/ distinction. With more cue-weighting shifted towards F3, the participants' identification of /r/ and /l/ also significantly improved.

Another important finding in the previous studies on sound category training, especially lexical tone training ([5], [10]) is that multi-talker stimuli are effective in terms of improving lexical tone categorization. Multi-talker training can also make learners' cue-weighting shift more towards pitch direction, the acoustic cue native Chinese speakers primarily rely on in tone perception [11]. If we take into account of the training paradigm, however, both manipulated variance and multi-talker variability have only been shown to be effective in the context of explicit learning. It is unclear whether they have the same effectiveness in an implicit learning environment.

There is growing evidence that overt and incidental learning paradigms draw upon neural substrates with distinctive computational specialties (e.g., [12], [13]). Based on the hypothesis of the reflexive and reflective learning model, [14] found that immediate and minimal feedback produced better perceptual learning of Mandarin lexical tones than the training with delayed and full feedback, suggesting different types of feedback trigger different neural learning mechanisms for tone categorization.

In the current study, we attempt to fill a research gap by investigating whether manipulated variance and talker variability are still effective for the perceptual learning of Mandarin tones in the context of incidental learning. An incidental learning involves no instructions to search for category-diagnostic dimensions, no overt category decisions, and no explicit categorization-performance feedback. Learners' goals and attention are not directed to sound categorization. However, participants quickly learn the sound categories and generalize to novel exemplars [15]. We examined both the tone categorization performance and the cue-weighting on pitch direction and pitch height before and after the training.

http://dx.doi.org/10.21437/TAL.2016-5

2. Incidental Learning—Video Game Playing

To create an incidental learning environment, we designed a video game and embedded the training of Mandarin lexical tones in the video game. Participants did not necessarily need to pay attention to the lexical tones. But being able to detect these tones can help them improve their video game performance.

2.1. Animal feeding video game

We created a video game in a 2D space. In the game, the participants needed to select the correct food to feed four different animals. There were four animals—1) cat; 2) monkey; 3) dog; and 4) rabbit. The animals' favorite food items were shown at the top of the screen—1) fish; 2) banana; 3) bone; and 4) carrot. The animal appeared one at a time and ran across the computer screen. Each animal was associated with a specific lexical tone: cat—T1; monkey—T2; dog—T3; rabbit—T4. For each lexical tone, there were 72 exemplars/tokens. During each trial, an exemplar of a lexical tone was randomly selected and played repeatedly to the participant together with the appearance of the animal. At the beginning, the animals were clearly visible, as shown in Fig.1. As the game progressed, it became more and more difficult to identify the animal visually, as shown in Fig.2. To make it difficult to identify the animals visually, we only showed part of the animal (e.g., only the head) in a vehicle. We used 7 different speed levels. The higher the game level, the faster the animal moved across the screen. The lexical tone information was available auditorily throughout the game. In other words, at the beginning levels of the game, players could simply depend on the visual information to feed the animal the correct food; for later levels, however, the players needed to rely more on the auditory information to identify the animal and feed it.

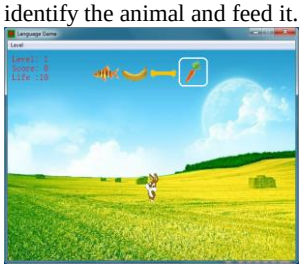


Figure 1. Game at level 1

Figure 2. Game at level 5

The participants needed to use their left hand to press keys 1, 2, 3 and 4 to choose the food. The selected food was highlighted. Then the participants needed to use their right hand to move the mouse over the running animal and left click to feed it.

In order to track the individual differences when playing of the game, each participant's correct and incorrect responses were recorded for each level of the game. In the training period, each participant played the video game for 4 sessions, each of which lasted 30 mins except the last session, which lasted 15 mins because the participants needed to do post training tone discrimination and identification tasks. This video game training paradigm was an implicit learning of lexical tone in nature as no explicit feedback or information about the tones was provided to the participants during the game.

2.2. Experiments

Using this video game, we conducted two experiments that used variance manipulated and multi-talker training stimuli for native English speakers without any tone language experience to learn L2 Chinese tone categories. In addition, a native and a non-native control conditions were included as well. The participants did two AX discrimination tasks before and after the training. The participants also did a word/tone identification task after the training.

2.2.1. Participants

For naïve listeners, there were one control condition and two training conditions. Ten participants were recruited and tested for each condition. None of the participants learned any tone languages before. We also tried to recruit participants who had as little formal music training as possible.

2.2.2. Experiment 1a—Native control

Experiment 1a aimed to examine native speakers' cue-weighting of different acoustic cues for lexical tone perception. Ten native speakers of Mandarin Chinese were recruited. However, they did not participate in any training. They did a single time tone discrimination task and we used INDSCAL [17] to calculate their cue-weightings based on their discrimination result.

2.2.3. Experiment 1b—Non-native control

Experiment 1b aimed to establish a baseline for naïve listeners' tone categorization. There is a small possibility that the participants' tone discrimination can improve after the pre-test without any tone training. We used four monosyllables /sa/, /fa/, /ma/ and /na/ recorded by a male native English speaker as training stimuli. The vowel duration of the tokens was normalized to be 300ms. Each monosyllable had 4 tokens. In total, there were 16 monosyllable tokens without lexical tones.

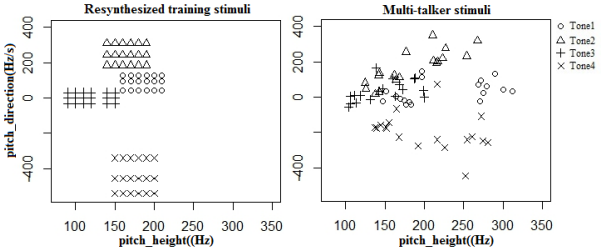
2.2.4. Experiment 2—Resynthesized stimuli training

Experiment 2 aimed to test the effectiveness of resynthesized stimuli on training lexical tone categorization and examine whether it can shift cue-weighting towards pitch direction in the context of incidental learning.

The stimuli consisted of four Mandarin lexical tones. The tones' pitch direction was quantified as (pitch offset-pitch onset)/duration whereas the pitch height was quantified as the f0 value averaged across 11 time normalized pitch values using Yi Xu's TimeNormalize Praat script [16]. A male native Chinese speaker who had a middle-range fundamental frequency recorded the four lexical tones on a monosyllable 'yu' /y/ (a high front rounded vowel) in citation form. We selected three tokens of each lexical tone with a slight pitch direction difference. We made sure the tokens of each tone had no overlap with tokens of any other tones on the pitch direction dimension. We then used PSOLA in Praat to shift the pitch tracks of each lexical tone so that six different pitch heights of each lexical tone were derived. Therefore, each lexical tone had 18 tokens (3 pitch directions x 6 pitch heights). We superimposed the different pitch tracks on a single 'yu' token to resynthesize the four tones so that all acoustic cues were controlled except tones. The duration of the vowel was normalized to 300ms. The amplitude was normalized too.

Five native speakers of Mandarin Chinese listened to the resynthesized tone stimuli and did a tone labeling task. All tone tokens were correctly identified by native Chinese speakers. The T3 that we used for the training was a low dipping tone. Following [11], we simplified the tone direction calculation of T3 by subtracting the pitch offset from the pitch onset divided by the vowel duration. Fig. 3 shows the distribution of the 18 exemplars for each lexical tone in the pitch height and direction space. As pitch height is correlated to the identification of

speaker's gender, we made the range of the pitch height of the resynthesized tone stimuli within the pitch range of the male voice. In this way, we can tease apart the effect of talker variability and variance manipulation on shifting cue-weighting in tone perception. All resynthesized stimuli sounded like male voice.



Figures 3 & 4. Distribution of resynthesized and multi-talker exemplars of four lexical tones in the pitch height and direction space.

2.2.5. Experiment 3—Multi-talker stimuli training

Experiment 3 aimed to test the effectiveness of multi-talker training stimuli on tone categorization and cue-weighting in the context of incidental learning.

Nine native speakers of Mandarin Chinese (5 males and 4 females) recorded their pronunciation of the four lexical tones. Each person produced each lexical tone twice. Therefore, each lexical tone had 18 exemplars. Fig. 4 shows the distribution of the multi-talker exemplars of the four lexical tones in the pitch height and direction space. Similar to the resynthesized training tokens, the multi-talker tokens had larger variance on pitch height than that on pitch direction but certain tones had overlap in the space. The amplitude and duration of the tone tokens were normalized.

2.2.6. Tasks

A word/tone identification task was used after the video-game training for the two training conditions. The stimuli used in the identification task was the same syllable 'yu' with four lexical tones. It included both training and new stimuli. The new talker stimuli were used to test whether word identification generalizes to new talkers. For the non-native control group, the test stimuli were the same four syllables with monotones used in the video game.

A speeded AX discrimination task was used as the pretest and posttest for all experiments (Native Chinese speakers in the native control condition only participated in the pretest AX discrimination tasks). The reaction time recorded in the speeded AX discrimination task was used to calculate the cue-weighting of pitch height and pitch direction using INDSCAL, a multidimensional scaling method [17].

3. Results

3.1. Tone identification result

As shown in Fig. 5, both the resynthesized and multi-talker training groups reached identification accuracy rate well above the chance level for the identification of the four lexical tones used in the training. The result suggests that the naïve listeners successfully associated the lexical tones to the animals in the video game after the video game play. Therefore, it suggests native English speakers can learn lexical tones incidentally. We conducted a 2x4x2 factorial ANOVA (within-subject: Block—old stimuli vs. new stimuli; Tone—4 lexical tones; between-subject: Training—resynthesized vs. multi-talker training), using a transformed tone identification accuracy rate arcsin as DV. The result showed a significant main effect of Block (old talker: 80% vs. new talker: 67%, $F(1,17)=6.5$, $p<.05$), a main effect of Tone (T1: 76%, T2: 75%, T3: 74%, T4: 65%, $F(1,17)=5.5$, $p<.05$) and a significant Tone x Training interaction ($F(2,17)=7.15$, $p<.01$). The main effect of Tone suggested the identification performance varied depending on the tones. The main effect of Block suggests that the overall tone identification accuracy rate for the new talker stimuli was significantly lower than that for the old talker stimuli. In the old talker stimuli identification, the multi-talker training had higher accuracy rates than the resynthesized training for T3 and T4 whereas the resynthesized training had higher accuracy rates for T1 and T2. In the new talker stimuli identification, multi-talker training had much higher accuracy rates for T3 and T4. It also had a higher accuracy rate for T2, but on a smaller scale. Making generalization is a crucial criterion of successfully learning new speech categories. The result suggests that multi-talker training is more robust in terms of training tone categorization especially for T3 and T4 as their identification can be better generalized to new test stimuli.

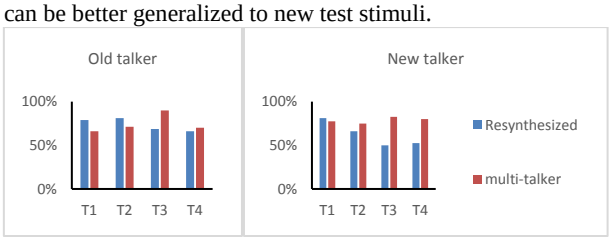


Figure 5. Identification accuracy rates of lexical tones as a function of training in the old and new test stimuli identification.

The better generalization of tone identification by the multi-talker training group could possibly come from its naturalistic variability in contrast to the synthetic one as in the resynthesized training condition or its talker variability whereas the resynthesized training lacks as it only had male voice. To address this question, we split the data into male and female test stimuli datasets and conducted two 4x2 factorial ANOVA (within-subject: Tone—4 lexical tones; between-subject: Training—resynthesized vs. multi-talker training) for the male and female test stimuli respectively. For the male test stimuli, there was a significant Tone by Training interaction ($F(3,68)=3.1$, $p<.05$). Shown as in Fig. 6, for the male voice tone identification, the multi-talker training group had higher accuracy rate than the resynthesized training group in identifying T2, T3 and T4 but the reverse happened in T1 identification. For the female test stimuli, there was also a significant Tone by Training interaction ($F(3,68)=3.0$, $p<.05$). As shown in Fig. 6, the resynthesized training group had higher accuracy rates whereas the multi-talker training group performed better for T3 and T4, especially T3.

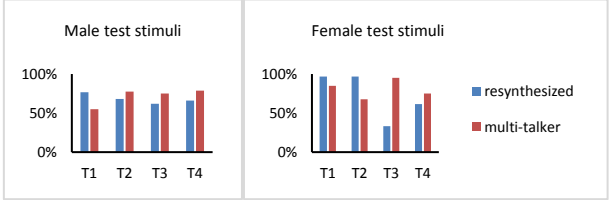


Figure 6. Identification accuracy rates of lexical tones as a function of training in the male and female test stimuli identification.

3.2. Cue-weighting result

The INDSCAL analysis generated two-dimensional configurations that reflected the perceptual distance among the

four lexical tones for each group. First, we replicated the results of previous cross-linguistic studies (e.g., [2], [11]) on cue-weighting for tone perception. Fig. 7 shows the group configuration of native Chinese speakers' cue-weighting on the two dimensions. On Dimension 1 (Dim 1), T2 and T4 were judged to be the most distant whereas on Dimension 2 (Dim 2), T1 and T3 were judged to be the most distant. As T2 being a rising tone and T4 being a falling tone, Dim 1 can be interpreted as pitch direction; as T1 being a high level tone and T3 being a low dipping tone, Dim 2 can be interpreted as pitch height. Since the native control group did not play the video game, there was no posttest cue-weighting result. For native English speakers' cue-weightings in the pretest, the left panels in Fig. 8, 9 and 10 show an opposite pattern where T1 and T3 were judged to be the most distant on Dim 1 whereas T2 and T4 the most distant on Dim 2. Therefore, Dim 1 was interpreted as pitch height while Dim 2 was interpreted as pitch direction. In INDSCAL, Dim 1 carries more weight. Therefore, Dim 1 is the primary acoustic cue whereas Dim 2 is the secondary cue.

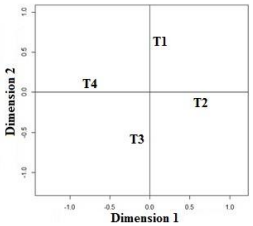


Figure 7. Native control group's cue-weightings.

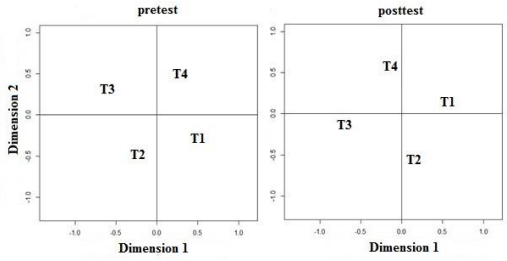


Figure 8. Non-native control group's cue-weightings in pre- and posttest.

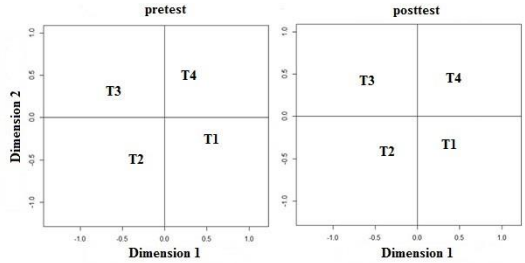


Figure 9. Resynthesized training group's cue-weightings in pre- and posttest.

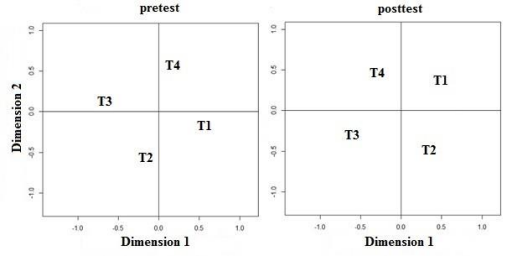


Figure 10. Multi-talker training group's cue-weightings in pre- and posttest.

After four days of incidental learning of lexical tones, in the posttests, the perceptual distance between T2 and T4 increased on Dim 1 among the resynthesized and multi-talker training groups but not in the non-native control group. The perceptual distance between T1 and T3 increased on Dim 2 among the resynthesized and multi-talker training groups but not in the non-native control group. The increased distance between T2 and T4 on Dim 1 means that the variance on Dim 1 was accounted for more by the pitch direction difference between T2 and T4. The increased distance between T1 and T3 on Dim 2 means that the variance on Dim 2 was accounted for more by the pitch height difference between T1 and T3. As the overall variance was primarily accounted for by Dim 1 and secondarily by Dim 2, these results suggest that cue-weighting has been shifted towards pitch direction in both resynthesized and multi-talker training groups.

The individual cue-weightings were also calculated. We conducted a 2x3 repeated measures ANOVA (within-subject: Test (pretest vs. posttest); between-subject: Experiments, using cue-weighting values on Dim 1 (pitch height) and Dim 2 (pitch direction) as DV). In terms of cue-weights on Dim 1 (pitch height), the result showed a main effect of Test ($F(1,27)=6.28$, $p<.05$) and a significant Test by Training interaction ($F(2,27)=1.77$, $p<.05$). Only the multi-talker training group had a significant cue-weighting decrease on the pitch height dimension ($F(1,27)=7.1$, $p<.05$) whereas the resynthesized training group and the non-native control group did not have a cue-weighting decrease on the pitch height dimension. In terms of cue-weights on Dim 2 (pitch direction), the result showed a main effect of Test ($F(1,27)=5.7$, $p<.05$) and a significant Test by Training interaction ($F(2,27)=2.82$, $p<.05$). Both the resynthesized training and multi-talker training groups had a significant cue-weighting increase on the pitch direction dimension (resynthesized training group: $F(1,27)=1.44$, $p<.05$; multi-talker training group: $F(1,27)=5.28$, $p<.05$) whereas the non-native control group did not have a cue-weighting increase on pitch direction.

4. Discussion

This study first showed lexical tone learning can happen incidentally by playing a video game. The learning outcome, the tone identification result suggests that the multi-talker training is still more robust in the incidental learning in terms of making generalization to new talker stimuli. Although overall the multi-talker training group performed better than the resynthesized training group, the tone identification accuracy rate is tone specific. The multi-talker training had much higher accuracy rate for T3 and T4. Such result suggests an interaction between the input and the acoustics of L2 sound categories that affect the ultimate sound categorization [19]. In terms of the cue-weighting result, both the naturalistic and resynthesized variance on the pitch height and direction lead to cue-weighting shifted towards the pitch direction as it had smaller variance relative to the pitch height. The result suggests the variance manipulation approach [8] can be applied to train cue-weighting for speech sound categories as well, lexical tones in this case. The multi-talker training also reduced the cue-weighting on pitch height. Based on the speech normalization model proposed by [20], we argue that the pitch height in the multi-talker stimuli that was correlated with gender identity made learners decrease their dependence on pitch height for tone categorization; rather, they learned to use it as a non-linguistic acoustic cue for gender identification. The more native-like cue-weighting, especially in the multi-talker training group, also helps the participants achieve a better tone identification in new test stimuli.

5. References

[1] Yamada, R. A., "Age and acquisition of second language speech sounds: perception of American English /r/ and /l/ by native speakers of Japanese." In W. Strange (Ed.),Speech perception and language experience: issues in cross-language research (pp. 305 – 320). Baltimore, MD: York Press, 1995.

[2] Gandour, J. "Tone perception in Far Eastern languages," Journal of Phonetics 11, 149–175, 1983.

[3] Bradlow, A. R., Pisoni, D. B., Yamada, R. A., and Tohkura, Y., "Training the Japanese listener to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production," The Journal of the Acoustical Society of America. 101, 2299–2310, 1997.

[4] Lively, S. E., Logan, J. S., and Pisoni, D. B., "Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environ-ment and talker variability in learning new perceptual categories," J. Acoust. Soc. Am. 94, 1242–1255, 1993.

[5] Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A., ''Training American listeners to perceive Mandarin tones,'' The Journal of the Acoustical Society of America, 106, 3649–3658, 1999.

[6] Goudbeek, M., Cutler, A., & Smits, R., "Supervised and unsupervised learning of multidimensionally varying non-native speech categories." Speech Communication, 50, 109–125, 2008.

[7] Iverson, P., Hazan, V., & Bannister, K., "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching /r/ -/l/ to Japanese adults.", The Journal of the Acoustical Society of America ., 118, 3267 –3278, 2005.

[8] Holt, L. L. & Lotto, A. J., "Cue weighting in auditory categorization: Implications for first and second language acquisition." The Journal of the Acoustical Society of America., 119, 3059-3071, 2006.

[9] Lim, S.–J. Lim & Holt, L. L., "Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization." Cognitive Science, 35, 1390-1405, 2011.

[10] Wong, P. C., Perrachione, T. K., and Parrish, T. B. "Neural characteristics of successful and less successful speech and word learning in adults," Hum. Brain Mapp 28, 995–1006, 2007.

[11] Chandrasekaran, B., Sampath, P. D., and Wong, P. C., "Individual variability in cue-weighting and lexical tone learning." The Journal of the Acoustical Society of America. 128 (1), 456-465, 2010.

[12] Doya, K., "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?", Neural Networks, 12, 961–974, 1999.

[13] Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A., "Performance feedback drives caudate activation in a phonological learning task." Journal of Cognitive Neuroscience, 18, 1029 –1043, 2006.

[14] Chandrasekaran, B., Yi, H.-G., & Maddox, W. T., "Dual-learning systems during speech category learning." Psychonomic Bulletin & Re-view, 21, 488 – 495, 2014.

[15] Wade, T., & Holt, L. L., "Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task". The Journal of the Acoustical Society of America, 118, 2618 – 2633, 2005.

[16] Xu, Y. "Contextual tonal variations in Mandarin.." Journal of Phonetics 25: 61-83, 1997.

[17] Carroll, J. D., and Chang, J. J. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," Psychometrika35, 283–319, 1970.

[18] Chandrasekaran, B., Krishnan, A., and Gandour, J. T. "Mismatch negativity to pitch contours is influenced by language experience," Brain Res. 1128, 148–156., 2007b.

[19] Escudero, P & Boersma, P. "Bridging the gap between L2 speech perception research and phonological theory." Studies in Second Language Acquisition, 26, 4: 551-585, 2004

[20] McMurray, B., and Jongman, A. "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations." Psychological Review 118, 219-246, 2011