The role of talker similarity in the perceptual learning of L2 tone categories

Jing Shao (jing.shao@polyu.edu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
Hung Hom, Hong Kong

Joanna Chor Yan Mak (joanna.mak@connect.polyu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
Hung Hom, Hong Kong

Caicai Zhang (caicai.zhang@polyu.edu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
Hung Hom, Hong Kong
Shenzhen Institutes of Advanced Technology, The Chinese Academy of Sciences
Shenzhen, China

Abstract

Different hypotheses were proposed concerning the role of talker variability in lexical learning. It remains unclear whether new phonetic categories are acquired as episodic memory traces with talkers' voice information preserved or as abstract categories. The current study investigated the role of voice similarity in perceptual learning of Cantonese tones. Six high-variability training sessions were given to 12 Mandarin speakers. Voice similarity was controlled in the training and pre-and posttests. Results indicate that the training positively transferred to both similar and dissimilar talkers. However, in the pretest, the performance was not significantly different between similar and dissimilar voices, whereas significant better performance was found in the similar voices in the posttest. These results suggest that learners retained speakers' information in the learning process and made use of such information for future perception. This implies that lexical tones are probably encoded episodically in the mental representation of Mandarin L2 learners.

Keywords: Talker similarity, high variability training, Cantonese lexical tones, Mandarin leaners of Cantonese, mental representation.

Introduction

When presented with a word auditorily, a listener receives a lot of information. For instance, the phonological form and meaning of the word, as well as the auditory information of the speaker. There existed a great amount of acoustic differences between speakers resulted from the differences in the shape and length of vocal tract, articulatory dynamics and native dialects (Goldinger, 1998). A long-documented problem for theories is how the speech perception and spoken word recognition achieve perceptual constancy in spite of the highly variable speech signals (Bradlow, Nygaard and Pisoni, 1999).

Two different accounts had been put forward concerning how the speaker's information is encoded in the memory system during auditory lexical learning, which are called abstractionist approach (Pisoni, 1997) and episodic theory (Nygaard, Sommers, & Pisoni, 1994). According to the speaker normalization hypothesis (Joos, 1948, as cited in Goldinger, 1998), acoustic variances produced by different speakers were considered as redundant information or noise that was quickly forgotten after lexical access and only lexical-semantic information was encoded in long-term memory. This allows listeners to understand new speakers instantly by only following the lexical-semantic content of speech and disregarding the superficial details such as speaker's information (Goldinger, 1998). In this abstractionist approach, a speaker-specific representation which contained all the lexical and speaker's information was first modified to a relatively speaker-neutral abstract representation prior to the encoding of lexical-semantic information (Johnson, 1997). The encoded lexical-semantic information was stored as an abstract representation, which then formed a template or prototype that could be used to match with the incoming speech signals to allow lexical retrieval (Goldinger, 1998). For instance, Peterson and Barney (1952) found that listeners could correctly perceive the target vowel despite the great variation in vowel formant frequencies produced by men, women and children. Evidence was also found in the perceptual normalization in consonants and prosody (Johnson, 2005), but normalization of such categories requires further intrinsic and extrinsic cues (Johnson, 2005; Moore, 1996; Strand & Johnson, 1996; Zhang & Chen, 2016). In a more recent study, it was also found that newly learned words could be sufficiently lexicalized, and be abstract with respect to talker voices (Kapnoula and McMurray, 2016).

On the other hand, a growing body of research suggested another direction. Episodic theory proposed that perceptual details including speaker's information were retained in memory during lexical learning and were integrated into perception later (Goldinger, 1996; 1998). All experienced instances were defined in a perceptual category and no abstract categories or prototypes were created. In accordance with this theory, Goldinger, Pisoni and Logan (1991) found that listeners made use of speaker variability to recall individual items in multiple-speaker word list, achieving higher accuracy than that in single-speaker word list. This result suggested that speaker variability in multiple-speaker word list facilitated word encoding and retrieval. Moreover, Bradlow, Nygaard and Pisoni (1999) presented a word list to a group of participants auditorily and asked them to judge whether they had heard the words before. They found that listeners made use of speakers' information as well as speaking rate and amplitude information in the tasks which also supported the episodic theory. In addition, Nygaard and Pisoni (1998) found that speech from familiar voices was more intelligible than from unfamiliar voices, which suggested speakers' information were related to linguistic processing at both word and sentence levels (see Souza et al., 2013 for similar findings). Altogether, these studies suggested that the talker information was stored as episodes in the long-term memory.

The two hypotheses mentioned above hold different views on the role of talker information in language processing/acquisition. Perceptual training in second language (L2) learners offers a scenario to test whether the new phonetic categories are learned as episodic memory traces with speaker's voice information preserved or as abstract categories, which is the aim of the current study.

High variability perceptual training (HVPT hereafter) was developed from the low variability training in which only one speaker was involved and very limited phonetic context was provided (Strange & Dittman, 1984). HVPT exposes subjects to a wider range of stimuli, including sounds produced by several speakers, in multiple phonetic contexts, and at multiple syllable positions (Bradlow, 2008). Several studies have adopted this approach to improve non-native speakers' identification of consonants (Bradlow et al., 1997; Bradlow et al., 1999;), vowels (Iverson & Evans, 2009; Iverson, Pinet & Evans, 2012), as well as lexical tones (Wang, Jongman & Sereno, 1999; Wang, Spence, Jongman & Sereno, 2003; Wang, 2013). These findings confirmed that HVPT was very effective in facilitating the acquisition of non-native phonetic contrasts.

Although HVPT was found to be effective in the generalization to novel stimuli and speakers, previous studies did not control the speaker voice similarity between training and pre-/posttests. The current study aims to control the voice similarity in the training sessions and pre-and posttest, and to compare the generalization effect to novel speakers whose voice was either similar or dissimilar to those speakers during training. Via this study, the question of whether L2 speech sounds are encoded as an abstract representation or an episodic representation during L2 perceptual learning will be investigated.

The current study focuses on the L2 lexical tone learning. Like consonants and vowels, lexical tones are important in differentiating word meanings in tonal languages. Mandarin and Cantonese are both tonal languages where pitch patterns of a syllable are crucial to its lexical meaning. Tone contours are commonly shown by numbers representing the pitch register according to a scale of five, 1 being the lowest and 5 being the highest. Usually two numbers, e.g., 55, indicate the pitch at the beginning and end of a syllable respectively (Chao, 1930). Mandarin has four lexical tones: Tone 1: high level (55); Tone 2: rising (35); Tone 3: falling-rising (214); and Tone 4: falling (51) (Norman, 1988). Cantonese has six regular tones, which are classified according to their register and contour (Bauer & Benedict, 1997). The six distinctive tones are: Tone 1: high level (55); Tone 2: high rising (25); Tone 3: mid level (33); Tone 4: mid-low falling (21); Tone 5: mid-low rising (23); and Tone 6: mid-low level (22). These six distinctive tones were included as the stimuli of the current study.

We aim to investigate the nature of mental representation of Cantonese tones in Mandarin L2 learners by controlling voice similarity of novel speakers. This allows us to examine the two hypotheses mentioned above. If Cantonese tones were encoded as an abstract representation, Mandarin listeners would ignore speaker variability and therefore generalize to novel speakers no matter whether their voices are similar or dissimilar to those in perceptual training. If Cantonese tones were encoded as an episodic representation, listeners would make use of speakers' information in lexical retrieval and therefore demonstrate better generalization to novel speakers whose voices are similar than those whose voices are dissimilar to the speakers in the perceptual training.

Method

General design

A pretest-training-posttest design was employed to assess the subjects' initial ability and the effects of training. Pretests and posttests consisted of two main parts: tone category identification and discrimination. The identification and discrimination tasks were designed to test whether identification training is effective and transfers to new talkers whose voice were similar and dissimilar to the talkers used in the training sessions.

Participants

Student participants were recruited in Hong Kong Polytechnic University (PolyU). 45 students responded to an online self-report questionnaire. 19 participants were selected based on the following criteria: (1) resided in Hong Kong for less than five months prior to the pretest session, (2) speaks Mandarin as mother tongue and did not speak any Southern dialect including Hakka and Southern Min dialect, (3) did not receive professional musical training. Seven participants withdrew from the study before the post-training test. A total of 12 participants completed the whole program.

Talkers and Stimuli

Stimuli

The stimuli were 60 words contrasting six Cantonese tones (high level tone (T1)-/55/, high rising tone (T2)-/25/, mid level tone (T3)-/33/, extra low level/low falling tone (T4)-/21/, low rising tone (T5)-/23/, low level tone (T6)-/22/) on ten base syllables (/jan/, /ji/, /jau/, /jiu/, /fan/, /fu/, /ngaa/, /si/, /se/ and /wai/), all are meaningful in Cantonese. Each monosyllabic target word was embedded in a carrier phrase context "呢個係_ lei1 go3 hai6 [target word]" (this is [target word]). Six female and six male native Cantonese speakers recorded the stimuli. Each speaker recorded six repetitions of each target word. One token for each target word was chosen by the experimenters according to its intelligibility and tone accuracy. The carrier phrase was normalized in duration to 877 ms (mean value of all carrier phrases), and the target word was normalized to 631 ms (mean value of all target words). The mean intensity was scaled to 70 dB using Praat.

Voice Similarity Judgment

The voice similarity among the 12 talkers was rated by another 12 native Cantonese speakers who were blind to the purpose of the current study. One speaker for each gender (F01 and M01) were chosen as references. The other five talkers (F02, F03, F04, F05 and F06; M02, M03, M04, M05 and M06) in each gender were compared against the reference speaker in term of voice similarity. In the similarity judgment experiment, raters were asked to listen to the target words embedded in the context "呢個係… lei1 go3 hai6…" [This is …] spoken by the reference speaker and one other speaker of the same gender. They were asked to rate the voice similarity of the speakers on a scale of 1 (very dissimilar) to 9 (very similar). The 60 target words were included and each trial was repeated twice.

The similarity score was averaged across raters. Table 1 shows the similarity score of each talker. Three talkers who received highest similarity scores in each gender group were used in the training sessions (F02, F04 and F06; M03, M04 and M06). The talkers with lowest similarity rating score in each gender (F03 and M05) were included in the pre- and posttests as the speakers with dissimilar voices.

Table 1: Result of similarity judgment test averaged across 12 raters.

Female talkers	Similarity score	Male talkers	Similarity score
F02	**7.44**	M04	**7.49**
F04	**7.27**	M06	**7.73**
F06	**6.07**	M03	**7.38**
F05	5.90	M02	5.91
F03	*4.70*	M05	*5.30*

Procedure

The training programme consisted of a pretest, training, and posttest phase. All sessions were conducted in a soundproof room in PolyU.

Training

There were six sessions of HVPT (i.e., tone identification with feedback). The entire course of training for each subject was completed over 1-2 weeks, and each session lasted about 1 hour. There was a different talker each session, as is typical of HVPT procedures. Female and male stimuli were trained alternatively. Moreover, our design was different from previous studies in that the six talkers in the training sessions (3 female and 3 male) were similar to two of the talkers who were presented in the pre- and posttest. Although the talkers used in the training phase were judged to have similar voices, it is still a HVPT, as the subjects were exposed to a wide range of stimuli, produced by several speakers, and the tones were carried by multiple syllables.

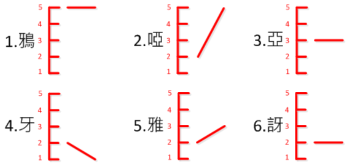

Figure 1. An example of identification choice in the training sessions.

The 60 stimuli were repeated twice, which gave 120 basic trials in each training session. On each trail, participants were presented with a target word with the context "呢個係 … lei1 go3 hai6…" [This is …]. Six choices with minimal contrast of tones were presented on the computer screen (see an example of "ngaa" in Figure 1). As can be seen in Figure 1, visual demonstration of each of the six tones were provided. The subjects were then asked to indicate the tone of the target word with the number keys 1-6 corresponding to the tones one to six. After each response, a feedback screen was presented. The feedback information includes: (1) if the response made by the participant was correct in blue text or incorrect in red text, (2) the cumulative accuracy of the current training session and (3) the correct response, its corresponding tone contour and its corresponding word written in traditional Chinese. If the response made was correct, the participant was proceeded to the next trial, if the response was incorrect, the incorrect trial was repeated until a correct response was indicated.

Pre- and Posttest

The pretest and posttest consisted of identification and discrimination tasks. In both identification and discrimination task, there were four talkers separated in four blocks. Two talkers (female F01 and male M01) were *similar* to those used in the training sessions, while two talkers were of *dissimilar* voice to the talkers used in the training (F03 and M05).

Tone identification. In the identification task, the 60 stimuli words were presented with the context "呢個係… lei1 go3 hai6…" [This is …]. Each trial was repeated twice and presented randomly to the participants. Six choices with minimal contrast of tones were presented on the computer screen with the tone, its corresponding tone contour and its corresponding character in traditional Chinese as shown in Figure 1. Participants were asked to indicate the tone of the target word by pressing the number keys 1-6 corresponding to the tones one to six. No feedback was given.

Category Discrimination. In the discrimination task, the syllable /ji/ was selected as the stimuli. Fifteen different tone pairs were presented in both directions of comparison (T1-T2, T2-T1, T1-T3, T3-T1, T1-T4, T4-T1 etc.) and then repeated twice, making up 60 "different" pairs. Six same tone pairs (T1-T1, T2-T2, T3-T3 etc.) were repeated 10 times to make up 60 "same" trials in order to balance the number of "same" and "different" trials, which gave a total of 120 trials in the discrimination test. Participants were asked to discriminate whether the two syllables were of the same or different tones by pressing the left (same) or right (different) arrow. No feedback was given.

For both discrimination and identification tasks, a short practice session was given before the first set of stimuli. Participants were allowed to take a break every 20 trials in both tasks. The responses of participants were recorded and coded. Response time was also collected.

Results

In both discrimination and identification tasks, for the accuracy analysis, mixed-effects logistic regression models were conducted, with the response to each trial as the input, *training* (pretest, posttest) and *voice similarity* (similar, dissimilar) as two fixed effects, and subjects as a random effect. For the response time analysis, two-way repeated measures ANOVAs were conducted, with the response time as the dependent variable and *training* (pretest, posttest) and *voice similarity* (similar, dissimilar) as independent variables. Figures 2 and 3 showed the mean accuracy of all participants achieved in identification and discrimination tasks respectively, in pretest and posttest for similar and dissimilar scenarios. Figure 4 and 5 showed the mean response times of all participants in identification and discrimination task.

For the identification accuracy, mixed-effects logistic regression model revealed significant main effect of *training* ($\chi2(1) = 315.81$, $p < 0.001$), and significant two-way interaction between *training* and *voice similarity* ($\chi2(2) = 5.185$, $p < 0.05$), while the effect of voice similarity alone was insignificant ($\chi2(1) = 2.267$, $p = 0.132$). Post-hoc tests showed that the accuracy in the posttest was significantly higher than the pretest in both similar ($z = -13.936$, $p < 0.001$) and dissimilar voices scenarios ($z = -10.962$, $p < 0.001$). Within the pretest, there was no significant difference between similar and dissimilar voices ($z = -0.389$, $p = 0.697$). However, in posttest, the accuracy in the similar voices scenario was significantly higher than that in dissimilar voices scenario ($z = 2.713$, $p < 0.01$).

For the response time in the identification task, there were significant main effects of *training* ($F(1, 11) = 20.471$, $p = 0.001$), *voice similarity* ($F(1, 11) = 25.73$, $p < 0.001$), as well as significant two-way interactions between *training* and *voice similarity* ($F(1, 11) = 8.552$, $p = 0.014$). Independent sample t-tests were then conducted within pretest and posttest to test the effects of speaker similarity. The results showed that in the pretest, response time in the similar voice scenario was marginally significantly longer than the dissimilar voice scenario ($t(22) = 1.899$, $p = 0.071$), but the difference was not significant in the posttest ($t(22) = 0.563$, $p = 0.579$), implying that training improved the response time in the similar voice condition. Within the similar speaker condition, the response time in the posttest was significantly shorter than the pretest ($t(22) = 2.708$, $p = 0.013$, suggesting the effects of training. While in the dissimilar voice condition, there was no significant difference between the pretest and posttest ($t(22) = 1.479$, $p = 0.153$), indicating that training had very little impact on the response time in the dissimilar voice scenario.

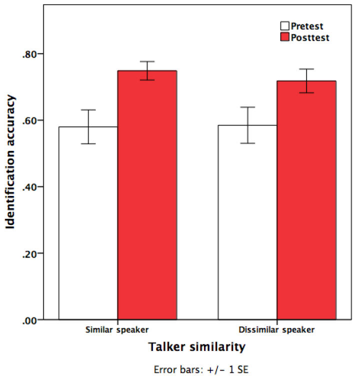

Figure 2: Mean accuracy in the identification task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.

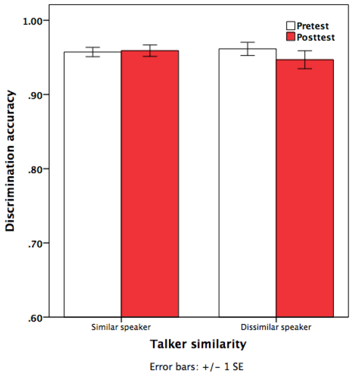

Figure 3: Mean accuracy in the discrimination task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.

In discrimination tasks, mixed-effects logistic regression model revealed a significant two-way interaction between *training* and *voice similarity* ($\chi2(2) = 4.362$, $p < 0.05$), while the effects of training ($\chi2(1) = 2.875$, $p = 0.090$) and voice similarity ($\chi2(1) = 1.110$, $p = 0.292$) were not significant. Post hoc tests showed that accuracy in pretest was significantly higher than that in posttest ($z = 2.665$, $p < 0.01$) with dissimilar voices. However, no significant difference was found between pretest and posttest in the similar voices scenario (z = -0.333, p = 0.740). The accuracy in similar voices scenario was significantly higher than that in dissimilar voices scenario within posttest (z = 2.197, $p <$ 0.05), while the difference in accuracy with similar and dissimilar voices was found to be insignificant within pretest (z = -0.809, p = 0.419).

For the response time in the discrimination task, no effects were significant.

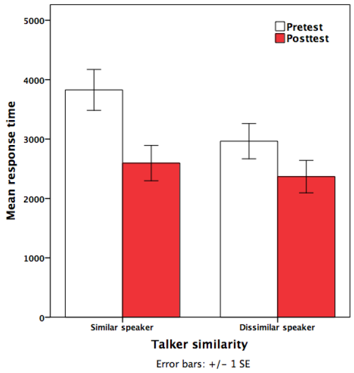

Figure 4: Mean response time in the identification task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.

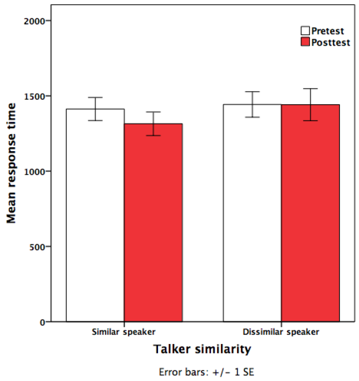

Figure 5: Mean response time in the discrimination task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.

Discussion

The current study provided HVPT to Mandarin speakers who had minimal exposure to Cantonese prior to the study. The generalization to novel speakers who had similar and dissimilar voices to the training stimuli was investigated to explore the role of talker voice similarity in the learning of Cantonese tones and henceforth to shed light on the nature of mental representation of Cantonese tones in L2 learners. If Cantonese tones were encoded as an abstract representation, no significant difference would be found in the generalization to similar and dissimilar novel voices. On the other hand, if Cantonese tones were encoded as an episodic representation, generalization to similar voices would be significantly better than to dissimilar voices.

We found that in the identification task, the effect of training was transferred successfully to both the similar and dissimilar voice scenarios. This finding fell in line with previous studies in that exposing the listeners to multiple talkers and phonetic contexts would enhance the phonemic categorization of the trained speech sounds. Miller, Zhang & Nelson (2016) investigated whether adult listeners who became deaf postlingually and had cochlear implant (CI) could benefit from multiple-talker category identification training. They found that the perception performance was significantly improved for the CI listeners for the familiar talkers (i.e., talkers used in the training sessions) and also generalized to the unfamiliar talkers (i.e., talkers not included in the training sessions). There was also evidence that talker variation aids young infants' phonotactic learning (Seidl, Onishi & Cristia, 2014). Together with these previous studies, our study provided extra evidence that the multiple-talker training was highly successful in learning to categorize speech sounds.

In both discrimination and identification posttests, the effect of voice similarity alone was significant. However, significantly better performance was found with similar voices than dissimilar voices in posttest but not in pretest and hence significant interaction between training and voice similarity was revealed. These findings echoed with previous studies that the speech from familiar voices was easier to identify than unfamiliar voices (Nygaard and Pisoni, 1998; Souza et al., 2013), supporting the hypothesis that Cantonese tones were encoded episodically in the mental representation of Mandarin L2 speakers. L2 learners retained perceptual details, speakers' voice characteristics, when encoding the tonal information into their mental representation. These perceptual details were integrated into later perception when the learner encountered with novel speakers. It is likely that the acoustic/phonetic representations are stored during the training stage, which facilitates the identification of tones in the similar voice context. In contrast, no matching representation is available in the dissimilar voice scenario, and thus the identity of the tone has to be construed from scratch.

As mentioned in the result, a significant training effect was found in identification tasks but not discrimination tasks. It is probably due to ceiling performance in discrimination tasks, for the reason that speakers without knowledge of Cantonese tones could also discriminate different tones merely by relying on psychoacoustic differences of the stimuli (Qin & Mok, 2011). In addition, since Mandarin is a tonal language, participants already had some tonal categories in their mental representation within their L1, although it is not as complex as Cantonese categories. Therefore, participants could make use of psychoacoustic differences and their L1 knowledge to tell apart perceptually different tones in discrimination tasks even they had no knowledge of Cantonese tones. Moreover, since only the syllable /ji/ was used in the discrimination task, lack of syllable variability also reduced the cognitive loading of the tasks.

It should be noted that our study focused on L2 learners with limited Cantonese exposure. The episodic encoding of lexical tones might paly a role in early stage of learning, when a new speaker' voice counts as a distinct learning episode. It is unclear whether the episodic representation of L2 phonetic categories will change in late stages of learning. Future studies may include the experienced L2 learners to test the scope of the episodic learning.

Conclusion

In the present study, perceptual trainings were given to native Mandarin speakers who had minimal exposure to Cantonese, and the voice similarity among the talkers in training and test phases was controlled. We found that the HVPT was highly effective, for the training effects generalized to both similar and dissimilar voices. However, the degree of generalization was significantly different between similar and dissimilar voices, which supported the hypothesis that learners retained speaker information during learning and made use of such information in future tone perception. This implies that newly learnt tones are encoded episodically in the mental representation of L2 learners. Future studies may explore the performance on the individual tones, so as to reveal the relationship between L1 and L2.

Acknowledgments

This paper was supported partly by grants to the last author from the Research Grants Council of Hong Kong (ECS: 25603916), and the PolyU Start-up Fund for New Recruits (Project account no.: 1-ZE4Y). We thank Kelly Kwong for the help with data collection.

References

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology.* Berlin: Mouton de Gruyter.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977-985.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & psychophysics*, *61*(2), 206-219.

Chao, Y.R. (1930). A system of tone letters. *Le Maitre Phonetique*. *45*: 24-27.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166-1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 152-162.

Kapnoula, E. C., & McMurray, B. (2016). Newly learned word forms are abstract and integrated immediately after acquisition. *Psychonomic bulletin & review*, *23*(2), 491-499.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (eds.), *Talker Variability in Speech Processing* (pp. 145–66). San Diego: Academic Press.

Johnson, K. (2005). Speaker Normalization in Speech Perception. In Pisoni, D. B., & Remez, R. E. (eds.), *The handbook of speech perception* (pp.363-389). Malden, MA: Blackwell Pub.

Miller, S. E., Zhang, Y., & Nelson, P. B. (2016). Efficacy of Multiple-Talker Phonetic Identification Training in Postlingually Deafened Cochlear Implant Listeners. *Journal of Speech, Language, and Hearing Research*, *59*(1), 90-98.

Moore, C. (1996). Speaker and rate normalization in the perception of lexical tone by Mandarin and English listeners. PhD Dissertation, Cornell University, Ithaca, NY.

Norman, J. (1988). *Chinese.* Cambridge University Press.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355-376.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, *24*(2), 175-184.

Qin, Z. & Mok, P. P. K. (2011). Discrimination of Cantonese tones by Mandarin, English and French speakers. In *The Psycholinguistic Representation of Tone, 2011* (pp. 50-53). Hong Kong: Causal Production.

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker variation aids young infants' phonotactic learning. *Language Learning and Development*, *10*(4), 297-307.

Strand, E. A. & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (ed.), *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld* (pp. 14–26). Berlin: Mouton de Gruyter.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, *24*(8), 689-700.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America J. Acoust. Soc. Am., 106*(6), 3649.

Zhang, C. & Chen, S. (2016). Toward an Integrative Model of Talker Normalization. *Journal of Experimental Psychology: Human Perception and Performance*.