Cognitive Representation of Phonological Categories: The Evidence from Mandarin Speakers' Learning of Cantonese Tones

Kaile Zhang[1], Yonghong Li[2], Gang Peng[1, 3]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
[2]Key Lab of China's National Linguistic Information Technology, Northwest University for Nationalities
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

kellerisapel@gmail.com, lyhweiwei@126.com, gpengjack@gmail.com

Abstract

Even when acoustic tokens vary substantially, they nevertheless can usually be recognized accurately. Two opposing models have been proposed to account for how the speech recognition mechanism works to achieve the perceptual consistency. The abstract model holds that there is a unitary cognitive representation for each phonological category. The speech signal, after having variations filtered out by a computational process, is matched to a particular representation. By contrast, the exemplar-based model holds that the previously encountered exemplars of a given speech category together form its mental representation. The speech recognition for this model involves searching for a match (based on similarity) between the incoming signal and stored exemplars. The present study tested which of these two models best fit data from second language acquisition. Mandarin speakers were trained with Cantonese tones that differed in acoustic variability. Results showed that training materials involving a large degree of within-class variability didn't produce a better learning outcome than those involving a small degree of variability, suggesting that the abstract model may provide a better fit for this data. The characteristics of Mandarin speakers' acquisition of Cantonese tones were also discussed.

Index Terms: speech recognition, the abstract model, the exemplar-based model, second language learning

1. Introduction

The acoustic features of given words produced by different speakers can vary dramatically ([1]-[4]). Variation also occurs in each individual under different conditions ([5]). The inter-talker and intra-talker variabilities represent significant difficulties for speech perception ([6], [7]). However, in the real communication, listeners can nevertheless in most cases overcome these obstacles and correctly categorize highly variable speech signals into their appropriate phonological categories.

Two models of speech perception provide contrasting hypotheses about how speech input from multiple talkers is integrated into the cognitive representation of a phonological category. The abstract model holds that whenever an individual perceives speech, the speech input is not recognized directly but undergoes online computation first. It is hypothesized that talker-specific information and variation in the speech are filtered out first, and then the speech signals are transformed into unitary abstract representations for each phonological category that are not talker-specific, thus achieving perceptual constancy (e.g., [8]-[10]). In support of the abstract model, [9] found that the F1-F0 and F3-F2 bark-difference dimensions provide robust representations of vowel height and vowel frontness/backness, thus reducing variability of the formant frequencies of vowels produced by different talkers. The resultant transformations allow vowels to be accurately classified, providing evidence for the online computation and transformation hypothesized by the abstract model for phonological category representation. On the contrary, the exemplar-based model holds that phonological categorization involves a direct match between the input signals and previously-encountered exemplars of this phonological category that are stored in memory. Based on this approach, all the speech signals listeners encountered previously are stored in the long-term memory along with talker-specific phonetic details. Once a new speech signal is received, it is categorized based on its perceptual similarity to the stored exemplars of each phonological category (e.g., [11]-[13]). The experimental evidence has shown that listeners detect a repeated word more accurately when the word is produced by the same talker than when it is produced by a different talker. This suggests that talker-specific information is stored implicitly together with each exemplar of a phonological category ([11]-[15]).

Because both of these models appear reasonable, it is unclear what cognitive mechanisms are used to represent and recognize phonological categories. We believe these models can be assessed against the predictions they make about how the variability of speech input affects second language (L2) learners' ability at forming mental representations of new phonological categories. According to the abstract model, no matter how the speech input varies, these details will be filtered out, such that after this computational adjustment all of them will belong to the same phonological category and will be mapped to a single cognitive representation. Therefore, speech variability will not be expected to impede the construction of the new phonological categories in their L2. Consistent with this model, [10] found that the acoustic stimuli that are closer to the grand mean of the speech community were in fact recognized more easily. They argued that listeners might have a prior expectation of the incoming speech signal and that the grand mean might be the expectation. Based on their experimental results and the logic of the abstract model, it is reasonable to further predict that the more congruent the speech inputs are with the unique mental representation, the

easier the construction of new phonological categories will be. By contrast, the exemplar-based model might provide different predictions. Since the recognition of a speech signal is based on the similarity between the speech input and stored exemplars, high variability of a phonological category's exemplars might make it easier for the incoming signals to be matched to a certain exemplar. From this view, the exemplar-based model suggests it is possible that forming cognitive representations of a new phonological category will be comparatively easier if the speech input is highly variable.

The present study will assess which model provides a more reasonable fit to data regarding the learning of Cantonese tones by Mandarin participants, using training materials containing different amounts of variability. The tone systems of Mandarin and Cantonese differ substantially. First, Mandarin has only four lexical tones (high level /55/, high rising /35/, low falling-rising /214/, and high falling /51/), whereas Cantonese has six long lexical tones (high rising /25/, low rising /23/, high level/55/, middle level /33/, low level /22/and low falling /21/) ([16]). Second, the four Mandarin tones differ from each other mainly by pitch contour, whereas both pitch contour and pitch height are important in Cantonese tone recognition ([7]). Since F0 is the primary acoustic correlate for lexical tones perception ([17]), the variability of tones can be manipulated by adjusting F0. Based on the above-noted predictions, we argue that if Mandarin participants trained with materials of high variability can acquire Cantonese tone system better than those with materials of low variability, the result would support the exemplar-based model; otherwise, the abstract model would be favored.

2. Methods

2.1. Participants

Thirty-five right-handed Mandarin subjects from Northern China were paid to participate in the experiment. All these participants are college students, with no self-reported visual, audio, and cognitive problems, and with no previous knowledge of Cantonese. They were randomly divided into two groups based on the size of tone variation of training materials they listened to: large-variability group (L) vs. small-variability group (S). Twelve Cantonese participants from Hong Kong were also recruited as the control group.

2.2. Stimuli

Ten Hong Kong Cantonese speakers (five males) were recruited to make recordings in a sound-attenuated booth. They were asked to pronounce 36 Cantonese syllables covering the Cantonese six tones (see Table 1) ten times in a natural way. Two sets of audio stimuli containing different degrees of variation for training, and another set of stimuli for testing were obtained based on these recordings.

The raw F0 value of each utterance was transformed from Hertz to log-scale 5-level values based on the Formula 1 in [18]. The grand mean pitch height and grand mean pitch slope of each tone category were calculated based on 10 subjects' recordings to form the reference for pitch adjustment. Four samples from each of four speakers whose mean pitch heights and mean pitch slopes were closest to the grand mean were chosen to form the stimuli of high variability. To keep the naturalness of each recording, only the pitch heights were adjusted to random numbers generated for each tone category (see the ranges of pitch height A in Table 2). To ensure the stimuli for the small variability set was as close to the grand mean as possible, the generation of this set of stimuli was slightly different. First, only the sample from each of the four speakers which was closest to the grand mean was chosen. This sample was reduplicated four times with each copy changed to a pitch height listed in Table 2 (the ranges of pitch height B). One speaker's recording from the remaining six was used as the testing set. One sample of each syllable was chosen and their pitch heights were changed based on the ranges A. In sum, 1152 (36 tonal syllables x 4 speakers x 4 samples x 2 variabilities) utterances were used as the training materials and 36 (36 tonal syllables x 1 speaker x 1 sample) were used as the testing materials.

Table 1. *Thirty-six Cantonese syllables.*

	fan/fɐn/	fu/fu/	jan/jɐn/	ji/ji/	se/sɛ/	si/si/
55	婚	夫	因	醫	些	詩
25	粉	苦	隱	倚	寫	史
33	訓	富	印	意	卸	嗜
21	焚	扶	人	兒	蛇	時
23	奮	婦	引	耳	社	市
22	份	負	孕	二	射	事

Table 2. *The pitch range used to manipulate the audio stimuli.*

Tone category	55	33	25	23	22	21
The ranges of pitch height A*	4.75 ± 0.25	3.25 ± 0.15	3 ± 0.25	2.5 ± 0.25	2.75 ± 0.15	1.75 ± 0.25
The ranges of pitch height B*	4.75 ± 0.05	3.25 ± 0.05	3 ± 0.05	2.5 ± 0.05	2.75 ± 0.05	1.75 ± 0.05

*The ranges of pitch height A are used to manipulate the training materials of large variability and the testing materials, and the ranges of pitch height B applies to the training materials of small variability.

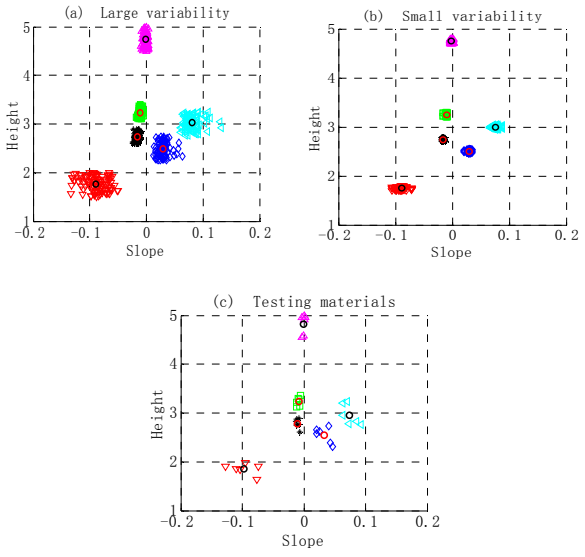

Figure 1: *The tone charts for all the speech utterances used as (a) training materials of high variability, (b) training materials of low variability, and (c) testing materials*

Figure 1 demonstrates the pitch information of each stimulus used as training materials for the large variability set (a), training materials for the small variability set (b), and the

materials for the testing set (c). The black or red circles in the middle of each tone category represent the grand mean of that tone category calculated based on 10 speakers' recordings. As can be seen, the testing set tokens [Figure 1. (c)] can be covered by the training tokens from the large variability set [Figure 1. (a)] but cannot be covered by those of the small variability set [Figure 1. (b)], thus the testing set tokens are more similar to the stimuli in the large variability set. However, the stimuli of small variability are closer to the grand mean, which might make the forming of the unique mental representation easier.

2.3. Procedures

The whole experiment contained two sessions of Cantonese tone training and three sessions of tone identification task that were carried out 1) before, 2) in the middle of, and 3) after the training. The training sessions used the speech shadowing techniques, that is, the stimuli (36 tonal syllables x 4 speakers x 4 samples x 4 repetitions) were played by Praat [19] and Mandarin participants were asked to imitate the recordings they heard. The corresponding traditional Chinese character, the jyutping, and the tone letter of each utterance were also shown on the screen for learners' reference. Each training session contained 6 sets of training and each set lasted about 30 minutes. Subjects finished one training sets every two days.

The tone identification task was used to evaluate their Cantonese tone learning performance. Thirty-six testing materials were randomly repeated five times. In each trial, an audio stimulus was played bilaterally to subjects, and subjects were instructed to identify the target word as any of the six Cantonese words which share the same base syllable as the audio stimulus but each with a different tone. Subjects were asked to press a button from 1 to 6 on the keyboard to indicate the tone perceived. Once a choice was detected, next trial was proceeded automatically.

3. Results

3.1. The accuracy of tone identification task

The accuracy (Acc) of Mandarin participants in tone identification task was submitted to a three-way repeated-measures analysis of variance (ANOVA) with *Experimental session* [session 1 (S1), session 2 (S2) and session 3 (S3)], *Base syllable* (/fɐn/, /fu/, /jɐn/, /ji/, /sɛ/, and /si/), and *Tone* (/55/, /33/, /22/, /21/, /23/, /25/) as the within-subject factors and *Group* (large variation and small variation) as the between-subject factors. There were significant main effects of *Experimental session*, F (2, 62) = 27.712; P<0.001, *Tone*, F (5, 155) =111.311; P<0.001, and *Base syllable*, F (5, 155) = 15.653; P<0.001. In addition, there was a significant two-way interaction: *Tone by Base syllable*, F (25, 775) = 18.718; P<0.001. No significant three-way interactions were found. The factor of *Group* was not involved in any significant effects, indicating the variability of speech input did not affect tone learning.

The accuracy of S2 (0.685) and S3 (0.686) is significantly better than that of S1 (0.623) (Ps<0.001). However, the accuracy of S3 is not significantly different from that of S2, The Accuracy among tones is also significantly different. Tone /21/ is the easiest tone to be identified, with the highest accuracy of 0.952, which was followed by the high level tone /55/ (Acc = 0.899). The low level tone /22/ is most difficult to identify (Acc = 0.279). The middle level tone /33/ (Acc = 0.635), the low rising tone /23/ (Acc = 0.634), and the high rising tone /25/ (Acc = 0.588) are almost equally difficult for Mandarin learners. Regarding the base syllables, utterances whose base syllables are /jɐn/ (Acc = 0.619), /sɛ/ (Acc = 0.62), and /fɐn/ (Acc = 0.664), /si/ (Acc = 0.675) are significantly difficult to identify than /fu/ (Acc = 0.702) and /ji/ (Acc = 0.707).

3.2. Confusion analysis

Table 3 illustrates the confusion of the six Cantonese tones. As can be seen, both Mandarin and Cantonese subjects seldom confuse the low falling tone with other tones. For two rising tones, they are easily confused with each other for both Mandarin and Cantonese speakers. But for the three level tones, Mandarin and Cantonese listeners show notably different perceptual preferences. Except for misidentifying the low level tone as middle level tone (56%), Mandarin subjects also perceived it as a high level tone (12%), whereas Cantonese subjects seldom did so (1%). In addition, Mandarin speakers frequently misperceive the middle level tone as high level tone (24%), while Cantonese speakers almost never confuse it with the high level tone (1%) but instead are more likely to confuse it with the low level tone (14%). Cantonese speakers rarely misidentify the high level tone (2%) but Mandarin speakers sometimes confuse it with middle level tone (7%). In sum, Cantonese speakers can differentiate each tone much better than Mandarin speakers, as would be expected. The three level tones were often confused with each other, as were the two rising tones. Cantonese and Mandarin speakers show different confusion patterns, especially for the three level tones.

Table 3. *Confusion matrix of tone identification for (a) Mandarin and (b) Cantonese participants.*

(a)

	R21	R22	R23	R25	R33	R55
T21	95%	2%	1%	1%	1%	0%
T22	1%	28%	2%	1%	**56%**	12%
T23	1%	1%	64%	**32%**	1%	1%
T25	1%	0%	**39%**	59%	0%	1%
T33	1%	**9%**	1%	1%	64%	**24%**
T55	0%	1%	0%	2%	**7%**	90%

(b)

	R21	R22	R23	R25	R33	R55
T21	98%	1%	2%	0%	0%	0%
T22	1%	49%	1%	1%	**47%**	1%
T23	2%	1%	79%	**17%**	1%	0%
T25	1%	1%	**22%**	76%	1%	1%
T33	1%	**14%**	2%	1%	82%	1%
T55	0%	0%	0%	1%	1%	98%

*The letter T in the table refers to the target responses and the letter R refers to the responses given by the participants.

4. Discussion

4.1. Cognitive mechanisms of constructing new phonological categories in L2

The fundamental difference of two opposing models regarding speech recognition lies in their hypotheses about the precise nature of mental representation. The abstract model holds that there is only one cognitive representation for each phonological category. To match individual tokens to this

unitary representation, the speech variability is argued to be filtered out first. Therefore, no matter how the speech tokens vary, they will be recognized successfully. By contrast, the exemplar-based model holds that the mental representation of each phonological category is actually a cluster of exemplars of that phonological category. The process of speech recognition under this model involves the assessment of the similarity between the incoming speech signal tokens and previously-experienced exemplars in certain dimensions. In the present study, the large-variation group, whose training materials cover all the testing materials in the dimension of both the pitch height and the pitch slope [see Figure 1 (a) and 1(c)], did not achieve a significantly better result in a Cantonese tone identification task. At the same time, although the small-variation group had never heard several exemplars [see Figure 1(b) and 1(c)] during their entire training sessions, they still were able to identify these exemplars as correctly as the large group did, again suggesting that the exemplar-based model does not explain these results as well.

The empirical data seemingly fit the abstract model better, but the results are not conclusive. Under the abstract model, to form new phonological categories a computational process would be triggered first to filter out the speech variation and then to extract the unitary representation of the speech samples belonging to the same phonological category. Speech samples of small variability theoretically might exert less pressure on the computation process than the samples of large variability, and thus the small group should form the Cantonese tonal categories easier and faster than the large group. However, the results of the present study didn't show that the small group is better than the large group. It is possible that different speech inputs do affect the construction of cognitive representations early in the learning process, but the learners' performance was tested too late in the present study to catch the between-group difference. That is, it is possible the stable mental representations had already formed before the test between two training sessions, and because the identical testing materials made the complexity of computation the same for both groups, the large group and small group could both recognize the Cantonese tones equally well in the tone identification task. This might also explain why the learners' tone identification didn't improve after six more training sets. Therefore, further research needs to be conducted to trace the learning dynamics with a smaller time window.

4.2. The influence of a Mandarin linguistic background on forming the Cantonese tone system

The present study also reveals some special features of Mandarin learners' acquisition of Cantonese tones. Generally speaking, Mandarin learners can hardly tell the tones with similar pitch slope (i.e. three level tones /22/, /33/, and /55/ and two rising tones /23/ and /25/), but tones of distinct pitch contour can be recognized easily (the falling tone /21/ is seldom misidentified), indicating that Mandarin speakers might be more sensitive to the change of the pitch direction but comparatively be less sensitive to the change of the pitch height [20]. This perceptual pattern could be partly explained by the characteristics of Mandarin tone system. The four Mandarin tones can be identified almost solely by the pitch contours. Therefore, the subtle change of the pitch height, which is less frequently used to distinguish Mandarin tones, may not as easily catch Mandarin speakers' attention. This is also consistent with [20] who reported that the language-specific weighting of pitch height and direction of change may explain perceptual discrepancies between different lexical tones.

The perceptual difference between native speakers and second language learners is mainly shown in the perception of three level tones. The middle level tone /33/ is easily misidentified by both Mandarin and Cantonese learners, but people with different language backgrounds show significantly different perceptual preferences. Mandarin speakers tended to perceive the middle level tone /33/ as the high level tone /55/ (which also exists in the Mandarin tonal system), while Cantonese speaker frequently misidentified it as the low level tone /22/ (which is acoustically closer to the Cantonese middle level tone /33/ than the high level tone /55/). Apparently, such perceptual preferences reflect that the native speakers rely more on the acoustic property to differentiate similar tone pairs, but when nonnative speech signals come into the auditory system of Mandarin learners, rather than being recognized directly, these acoustic signals are affected by expectations from their native Mandarin phonological system. As the Perceptual Assimilation Model for suprasegmentals (PAM-S; e.g., [21], [22]) suggested, Mandarin learners might treat the Cantonese middle level tone /33/ as an exemplar of Mandarin high level tone /55/ and then misidentify it as Cantonese high level tone /55/.

The acquisition of the tonal system in L2 is affected not only by the suprasegmental knowledge of one's native language but also by the segmental components. The result of the tone identification task shows that syllables /jɐn/, /sɛ/, /fɐn/, and /si/ (which only exist in Cantonese) are significantly more difficult to identify for Mandarin speakers than /fu/ and /ji/ (which are shared by both Mandarin and Cantonese). It seems the speech recognition mechanism processes the segmental and suprasegmental components of the speech signals simultaneously. The familiarity to the segmental components makes the recognition task relatively easier and thus leaves more cognitive resource to process the suprasegmental components, yielding a better identification of tones ([23]).

5. Conclusion

The present study examined two models regarding the speech recognition — the abstract model and the exemplar-based model — from the perspective of second language learning. Learners whose training materials share more similarity with the testing materials didn't achieve a better result than those of less similarity, suggesting that the speech recognition might not be a direct match between the incoming signals and the exemplars stored mentally. Considering the fact that the construction of the mental representation was not significantly affected by the variabilities of learning materials, a computation process proposed by the abstract model must occur. However, further studies need to be carried out to test the fit of the abstract model with a more detailed tracking of the formation of the new phonological categories. In addition, familiarity to the segmental components may also affect the process of formation of a new tonal system.

6. Acknowledgements

This study was supported in part by a grant from Research Grant Council of Hong Kong (GRF: 14411314).

7. References

[1] G. E. Peterson and H. L. Barney, "Control methods used in the study of vowels," *Journal of the Acoustical Society of America*, vol. 24, pp. 175-184, 1952.

[2] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological Review*, vol. 74, pp. 431-61, 1967.

[3] K. Johnson, "Speaker normalization in speech perception," In D. B. Pisoni and R. E. Remez (eds.), *The Handbook of Speech Perception* (pp. 363-89). Blackwell Publishing. 2005.

[4] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara and T. Irino, "The processing and perception of size information in speech sounds," *Journal of the Acoustical Society of America*, vol. 117, pp. 305-18, 2005.

[5] Y. Xu, "Sources of tonal variations in connected speech," *Journal of Chinese Linguistics*, Monograph Series vol.17, pp. 1-31, 2001.

[6] P. C. M. Wong and R. L. Diehl, "Perceptual normalization for inter-and intra-talker variation in Cantonese level tones," *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 413–421, 2003.

[7] G. Peng, C. C. Zhang, H. Y. Zheng, J. W. Minett and W. S. Y. Wang, "The effect of intertalker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems," *Journal of Speech, Language, and Hearing Research*, vol. 55, pp. 579–595, 2012.

[8] L. Gerstman, "Classification of self-normalized vowels," *IEEE Transactions of Audio Electroacoustics*. AU-16, pp. 78-80, 1968.

[9] A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *Journal of the Acoustical Society of America*, vol. 79, pp. 1086-100, 1986.

[10] C. C. Zhang, G. Peng, and W. S. Y. Wang, "Unequal effects of speech and non–speech contexts on the perceptual normalization of Cantonese level tones," *Journal of the Acoustical Society of America*, vol. 132, pp.1088–1099, 2012.

[11] D. L. Hintzman, R. Block and N. Inskeep, "Memory for mode of input." *Journal of Verbal Learning and Verbal Behavior*, vol. 11, 741-749, 1972.

[12] S. D. Goldinger, "Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, pp.1166-83, 1996.

[13] K. Johnson, "Decisions and mechanisms in exemplar-based phonology," In Sole, M. J., Beddor, P., Ohala, M., (Eds.) *Experimental approaches to phonology: In honor of JohnOhala* (pp. 25-40). Oxford University Press, 2007.

[14] T. J. Palmeri, S. D. Goldinger, and D. B. Pisoni "Episodic encoding of voice attributes and recognition memory for spoken words," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, pp.309-28. 1993.

[15] S. D. Goldinger, "Echoes of echoes? An episodic theory of lexical access," *Psychological Review*, vol. 105, pp. 251-79, 1998.

[16] M. Yip, *Tone* Cambridge: Cambridge University Press, pp. 17-208, 2002.

[17] W. S. Y. Wang, "The many uses of F0," in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, edited by A. Valdman (Mouton, The Hague), pp. 487–503, 1972.

[18] G, Peng and W. S.-Y. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines", *Speech Communication,* vol. 45 49–62, 2005.

[19] P. Boersma, and D. Weenink, "Praat: Doing phonetics by computer (Version 5.3.23) [Computer program]," http://www.praat.org (Last viewed on August 7, 2012), 2012

[20] J. Gandour, "Tone perception in Far Eastern languages", *Journal of Phonetics*, vol.11, pp.149-175, 1983.

[21] C. K. So and C. T. Best, "Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences," *Language and Speech*, vol. 53, pp. 273 – 293, 2010.

[22] C. K. So and C. T. Best "Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories," *Studies in Second Language Acquisition*, vol.36, pp.195–221. 2014.

[23] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, pp. 159-177, 2002.