9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Perception and production of Cantonese tones by South Asians in Hong Kong

Peggy P. K. Mok[1], *Crystal W.T. Lee*[1], *Alan C. L. Yu*[2]

[1]The Chinese University of Hong Kong, Hong Kong
[2]University of Chicago, U.S.A.

peggymok@cuhk.edu.hk, lwt1011@gmail.com, aclyu@uchicago.edu

Abstract

This study investigates the perception and production of Hong Kong Cantonese tones by South Asians residing in Hong Kong. Forty-three South Asian participants and twenty-six ethnic Chinese Hong Kong Cantonese speakers completed an AX discrimination task and a picture naming task. A series of regression analyses showed that, relative to the Chinese cohort, the South Asian cohort showed significant neutralization of tonal contrasts in production, as well as poorer tonal discrimination accuracy, especially among participants whose dominant language is Punjabi, which also has contrastive tones in its phonology. These findings are consistent with predictions of existing models of L2 phonetic acquisition, which argue that the hardest elements in L2 phonology for learners are those bearing similar features from their L1, rather than those that are different.
Index Terms: tone perception, tone production, L2 phonetics, Cantonese.

1. Introduction

Many studies have investigated the perception and production of tones of a second language (L2) by learners whose first language (L1) is either tonal or non-tonal [1, 2]. These studies have primarily focused on the acquisition of tone in a structured instructional environment where the participants in the study were taught to produce tones in the L2. Little is known regarding the perception and production of L2 tones by learners who learn Cantonese in an "organic" way, i.e. via general exposure in the environment without explicit instructions in the phonetics and phonology of the tone language.

The South Asian (SA) population in Hong Kong represents a unique situation in this respect in terms of L2 acquisition, particularly with respect to tone. There are more than 60,000 inhabitants in Hong Kong that are of South Asian descent, many of them are born and raised in Hong Kong. Yet, despite Cantonese being the dominant language of Hong Kong, a large percentage ($\sim$ 80%) of the South Asian inhabitants are illiterate in Chinese and a similarly large percentage ($\sim$60%) are reported to not speak Cantonese at all. Instead, in addition to English, they speak Urdu, Punjabi, Hindi, Gujarati, Sindhi, or Nepali, among others. While an increasing number of these ethnic minority parents in Hong Kong are sending their children to mainstream public sector schools, [3], given that the primary aims of the language education in Hong Kong are to develop biliterate competency in English and Standard Written Chinese, and trilingual competency in spoken Cantonese, Putonghua, and English, SA students often cannot rely on English as the sole medium of communication in school. Also, most primary schools in Hong Kong use spoken Cantonese and written standard Chinese as their medium of instruction. While some secondary schools are allowed to teach in English, most government and aided secondary schools in Hong Kong focus on Chinese medium of instruction. This sociological backdrop suggests a unique situation where SA students in school, to the extent they speak Cantonese at all, must acquire the language via general exposure, rather than through explicit classroom instructions geared toward teaching Cantonese as a second language. This study aims to examine the production and perception of Hong Kong Cantonese tones by South Asian secondary school students.

1.1. Tone systems of Cantonese, Urdu and Punjabi

Cantonese has six lexical tones (T): T1 [55] high-level, T2 [25] high-rising, T3 [33] mid-level, T4 [21] low-falling, T5 [23] low-rising, and T6 [22] low-level [4]. An example of the canonical tonal realization of these tones can be found on the left panel of Figure 1. Recent studies have found that younger Cantonese speakers are undergoing tonal changes (e.g.,[5]). These younger speakers tend to merge T2 and T5, as well as T3 and T6 respectively but the process is still incomplete. This paper compares the tone performance of Urdu and Punjabi speakers to that of teenage native Cantonese speakers, so it is expected that a portion of the recruited Cantonese participants were tone mergers.

Urdu is a non-tone language. Punjabi has three phonemically distinct tones: high-tone (rising-falling), low-tone (falling) and mid-tone (mid level) [6]. Punjabi possesses tonal and non-tonal words [7]. In tonal words, there are two tones i.e. high vs. low tones. Non-tonal words carry the mid-tone which is predicted by rules of redundancy [8]. Furthermore, it is argued that younger Punjabi speakers decreasingly rely on pitch contours to distinguish word meanings and tend to replace rising and level pitch by falling pitch, possibly due to the ease of articulation [9]. It is hypothesized that the complex Cantonese tone system may pose difficulties for Urdu and Punjabi speakers.

1.2. Speech learning mechanisms

Similarities and differences between native (L1) and second (L2) languages play an important role in speech learning. The Perceptual Assimilation Model (PAM; [10]) argues that, if naïve learners assimilate two contrastive sounds in L2 into different L1 categories, the contrast will be successfully discriminated; if the two contrastive sounds are assimilated into one single L1 category, the discrimination will be inaccurate. Non-native contrasts are thus not equally difficult for listeners to perceive and the difficulty depends on how they perceive the L2 sounds in relation to L1 categories.

While PAM focuses on speech perception between L1 and L2, the Speech Learning Model (SLM: [11]) connects speech perception and production in L2 phonology. It claims that the learner's phonetic system reorganizes in response to L2 by adding new categories or modifying old ones. SLM classifies the L2 sounds, in relation to L1, as "identical", "similar", or "new". Similar sounds often result in inaccurate pronunciation,

10.21437/SpeechProsody.2018-93

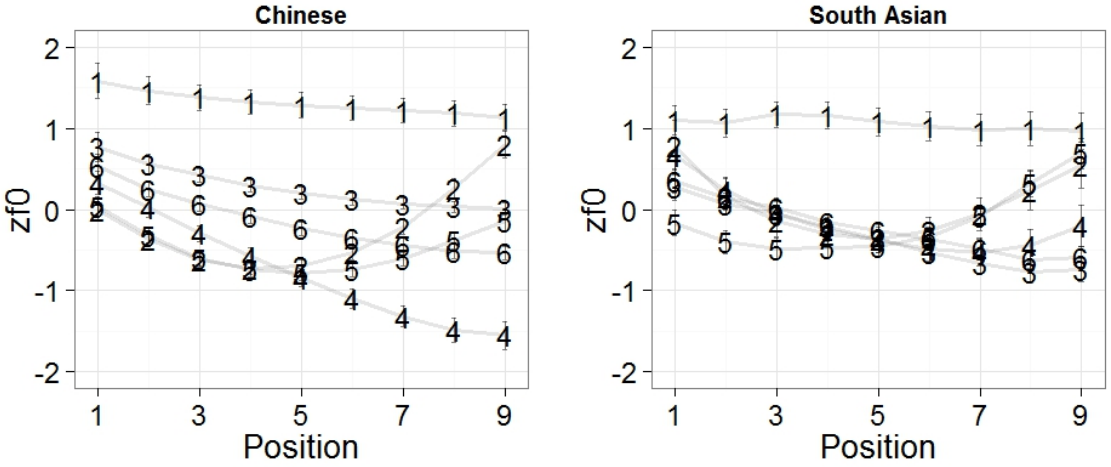

Figure 1: *Average z-transformed f0 of the six Cantonese tones across nine measurement positions produced by Chinese (left) and South Asian (right) Cantonese speakers.*

as failure to detect the difference between an L1 sound and an L2 sound leads to "equivalence classification", i.e., a single phonetic category used to process the linked L1 and L2 sounds. Then the L2 production will be influenced by features of the L1 category. By contrast, a "new" L2 sound, due to its disparate features from L1, will not be analogized to an existing L1 sound and will cause the formation of a new category for this L2 sound. However, this new category may still be dissimilar from L1 to ensure a maximal dispersion of L1 and nearby L2 sounds. The SLM has been applied to production studies of non-native sounds in diverse L1-L2 pairs (e.g. [12, 13]). L2 learners often produce L2 sounds with similar L1 equivalents ("similar") differently from native speakers, while they can produce L2 sounds with no L1 equivalents ("new") more accurately.

Both PAM and SLM suggest that the hardest elements in L2 phonology for learners are not the ones that are very different from their L1, but rather those bearing similar features as their L1. Therefore, as Punjabi is a tone language, Punjabi-speaking L2 Cantonese learners may have greater difficulties over Cantonese lexical tones than Urdu-speaking learners, especially with the interruption of a different tone assignation rule in their L1 Punjabi. Apart from L1 influence, this study also examines how the demographic backgrounds of the South Asian speakers in Hong Kong affect their Cantonese tone learning. We test this prediction in this study.

2. Methods

2.1. Participants

Forty-eight South Asian secondary students, and thirty three Hong Kong Chinese participants took part in both the production and perception experiments. As the first language of some of the SA participants is a language other than Urdu, Punjabi, or Hong Kong Cantonese, their data were excluded in the analysis. In the end, the final data set includes production and perceptual responses from 43 SA participants (24 males, 19 females) and 26 Chinese participants (11 males, 15 females). While all SA participants, aged from 12 to 18, are speakers of Urdu and/or Punjabi, 8 are English-dominant, 14 Punjabi-dominant, 15 Urdu-dominant, and 6 claimed to be fluent in multiple (non-Cantonese) languages. The Chinese participants, aged from 15 to 18, were either secondary students or first-year undergraduate students at a university in Hong Kong. The participants, who were paid a nominal fee to take part in the experiment, reported no hearing or speech problems.

2.2. Materials

Production data was elicited via a picture-naming task, which included 84 pictures. Each picture was accompanied by the corresponding Chinese character and English gloss to facilitate production. The tonal production was part of a larger study examining SA Cantonese. Of the 84 pictures, twelve were intended to elicit the tone contrasts. The target words were the six tonal variants of the syllables [ji] and [si]. The other pictures were designed to elicit segmental contrasts.

The perception task consists of an AX discrimination task with 150 AA pairs and 150 AB pairs. Five syllables, namely [fa:n], [fan], [si], [jan], and [ji], each carrying 6 tones in Cantonese, were chosen. Therefore, there were 30 target monosyllables. They were paired up with monosyllables having the same segment and same tone to form the AA pairs, and with monosyllables with the same segment and different tones (e.g. T1/T2, T2/T5) to form the AB pairs. Since there were 15 possible tone combinations to form the AB pairs, and the order of the AB pairs was counterbalanced, there were 150 AB pairs altogether (15 tone combinations x 2 orders x 5 syllables). The AA pair of each syllable appeared five times in order to the balance the number of the AB pairs (6 tone combinations x 5 syllables x 5 repetitions = 150). The stimuli were produced by a phonetically-trained female native speaker of Hong Kong Cantonese of ethnic Chinese background. The two stimuli in the AA pairs were not from the same recording. All pairs were randomized in the perception task.

2.3. Procedure

All subjects participated in both the production and perception tasks, and finished a language background questionnaire in one sitting. Half of the participants did the production task first, while the remaining half did the perception task first. The experiments were conducted in quiet classrooms in the respective schools (for the secondary school students) or a recording booth (for the university students). In the production experiment, all subjects were instructed to say the monosyllabic words in the

pictures naturally. They were given plenty of time to pronounce the words themselves. When needed, the experimenters would provide hints if they had difficulties in recalling the pronunciation. If a student knew the word, they pronounced the target monosyllabic word by themselves for at least 2 times (self-attempted). If they really did not know the word, they were prompted to repeat after the experimenters for at least twice (shadowed). The two repetitions with the best quality were chosen for analysis. Their speech was recorded using a digital audio recorder placed approximately 20 cm away from them. Only self-initiated production responses were included in the analysis. The final dataset in the production analysis consists of productions from 27 SA speakers (11 females; 6 English-dominant, 4 Punjabi-dominant, 13 Urdu-dominant, and 4 with multiple dominant languages) and 26 Chinese speakers. Missing f0 measurements were smoothed and the f0 measurements were z-transformed by participant prior to further analysis.

In order to quantify the degree of difference between tones across cohorts, we rely on a calculation of distance based on parameterizing the f0 trajectory using the discrete cosine transition over the rhyme of the vowel [14, 15]. The DCT allows the reduction of the complexity of the f0 trajectory into a triplet of coefficients that are proportional to the mean, linear slope, and curvature of f0 respectively. Taking the coefficients of the mean f0 trajectories of each tone of the CC speakers as the centroids (the mean of each tone in this three-dimensional space), we calculate the Euclidean distance from each tone production of each SAC speaker to the respective tonal centroid. For example, the first 3 DCT coefficients for Tone 1 for the CC speakers are 1.859, 0.181, and 0.037 respectively and the 3 DCT coefficients for the same tone for SA participant #1 were 1.442, -0.745, -0.853. The Euclidean distance between the two points in the three-dimensional space is 1.35.

In the AX discrimination task, each subject listened to a randomized list of 300 monosyllabic word pairs and was asked to judge if the tones of the two monosyllables were the same. The stimuli were presented using E-Prime 2.0 on a laptop computer via headphones. They were instructed to use the index and the middle fingers of their dominant hand to press the keys for the 'same' and 'different' responses on the keyboard. Also, they were required to respond as quickly (no more than 5000 ms) and as accurately as possible. Participants were given a short practice session prior to the actual task. The task was divided into six blocks, with short breaks scheduled between blocks. Prior to each trial, a fixation point appeared on the screen. Participants were given feedback only in the practice session. The inter-stimulus interval was 500 ms. Reaction time was measured from the offset of the second monosyllable.

3. Results

Figure 1 summarizes the tonal production results. Qualitatively, there are less tonal distinctions in the SA productions compared to the productions of the Chinese cohort. To begin with, the SA tone space appears to be much smaller than that of the Chinese cohort. Also, certain tonal distinctions are neutralized. In particular, T2 and T5 among the SA cohort are mainly differentiated at the tonal onset, rather than at the offset, as in the case of the Chinese cohort. Also, among the SA productions, the contrast between T3 and T6, and possibly their contrasts with T4, appear to have been neutralized.

The distance-from-centroid measurements were modeled in terms of linear regression. Besides TONE (6 levels), other predictors tested included subject AGE, GENDER, years of RESIDENCE in Hong Kong, dominant language (English, Punjabi, Urdu, and multiple languages) and EDUCATION level. Both TONE and GENDER were treatment-coded with T6 and Female set as the reference levels respectively. All other variables were treated as continuous variables and were z-scored and centered. The final model includes TONE and years of residence in Hong Kong. The inclusion of the other predictors, including their two-way interactions with TONE, did not significantly improve model likelihood. Table 1 summaries the regression model for the centroid distance analysis. As T6 was set as the reference level, the model shows that the other tones show significantly higher distances from the CC centroid. Years of residence in Hong Kong also significantly affects the centroid distance. The longer the SA participant resides in Hong Kong, the smaller the distance her tones are from the Chinese Cantonese tones (β=-0.09, t = -2.53, $p < 0.05$). The effects of language dominance were not analyzed since many of the Punjabi-dominant participants were excluded.

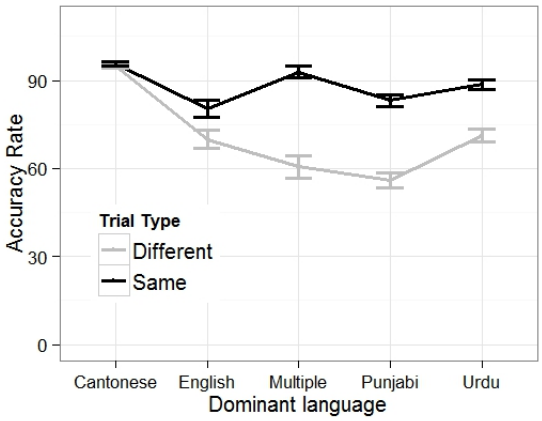

Figure 2: *Averaged accuracy rates by language dominance. Error bars indicate 95% confidence intervals.*

Table 1: *Summary of regression model for centroid distance.* * $= p < 0.05$, ** $= p < 0.01$, *** $= p < 0.001$

	Coef (SE)	t value
INTERCEPT	0.72 (0.09)	8.34 ***
T1	0.24 (0.12)	1.95
T2	0.46 (0.12)	3.78 ***
T3	0.37 (0.12)	3.08 **
T4	0.59 (0.12)	4.90 ***
T5	0.25 (0.12)	2.08 *
RESIDENCY	-0.09 (0.04)	-2.53 *

Response accuracy for the stimulus pairs were modeled using a series of logistic mixed-effects regressions fitted in `R`, using the `lmer()` function from the `lme4` package. Unlike the production data, which included only a subset of data, discrimination responses from all 43 SA participants were included. Responses made with less than 100 ms (10% of the trials) were excluded in the regression analysis. The first regression model (Model 1) includes the following predictors: log-transformed reaction time (logRT), DOMINANT language (5 levels), and tone PAIR TYPE (same vs. different), as well as the interaction between DOMINANT and PAIR TYPE. The model also includes by-subject random intercepts as well as by-subject random slopes for logRT and PAIR TYPE. DOMINANT was contrast-coded in such a way to yield four contrasts: Cantonese-dominant vs. the rest, Punjabi- vs. Urdu-dominant, Punjabi- vs. English-/Urdu-dominant, Cantonese-/Punjabi-dominant (tonal

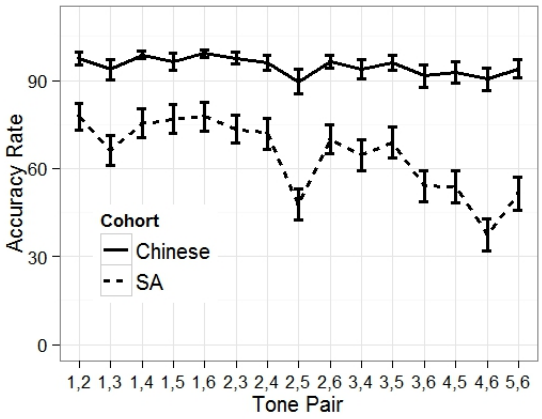

Figure 3: *Averaged accuracy rates for each tone pair by the two cohorts. Error bars indicate 95% confidence intervals.*

languages) vs. the rest. PAIR TYPE was sum-coded and logRT was scaled and centered.

Table 2: *Summary of Model 1.* $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$

	Coef (SE)	z value
INTERCEPT	1.86 (0.11)	17.26 ***
LOGRT	-0.26 (0.04)	-6.04 ***
PAIR TYPE	-0.57 (0.07)	-7.87 ***
DOMINANT1:Chinese vs. SA	1.29 (0.16)	8.12 ***
DOMINANT2:Punjabi vs. Urdu	-1.29 (0.22)	-5.75 ***
DOMINANT3:Punjabi vs. English/Urdu	-1.00 (0.17)	-5.87 ***
DOMINANT4:Tonal vs. nontonal	-0.72 (0.21)	-3.39 ***
PAIR TYPE x DOMINANT1	0.35 (0.11)	3.13 **
PAIR TYPE x DOMINANT2	-0.34 (0.15)	-2.23 *
PAIR TYPE x DOMINANT3	-0.48 (0.12)	-4.09 ***
PAIR TYPE x DOMINANT4	-0.33 (0.15)	-2.27 *

Model 1 is summarized in Table 2. There is a main effect of PAIR TYPE (β=-0.57 z = -7.87, $p < 0.001$), suggesting that the accuracy rates for the "different" trials are lower than the "same" trials in general. LogRT also came out significant (β=-0.26 z = -6.04, $p < 0.001$), suggesting accuracy suffers when logRT is longer. Language dominance also affects tonal discrimination accuracy. To begin with, participants who are Cantonese-dominant (i.e the Chinese cohort) has a better discrimination accuracy than the SA cohort (β=1.29, z = 8.12, $p < 0.001$). Participants who are Punjabi-dominant are significantly worse than the Urdu-dominant participants in terms of tonal discrimination (β=1.29, z = 8.12, $p < 0.001$). Punjabi-dominant participants also performed less well relative to the English-dominant and Urdu-dominant participants (β=-1, z = -5.87, $p < 0.001$). Speakers of a tone-language are better at tonal discrimination, although this effect is likely to be driven by the performance of the Cantonese-dominant speakers in light of the poor performance of the Punjabi-dominant speakers (β=-0.72, z = -3.39, $p < 0.001$). There are also significant interactions between PAIR TYPE and DOMINANCE. As illustrated in Figure 2, while the discrimination accuracy among the SA participants are generally poorer than those of the Cantonese-dominant (i.e. the Chinese cohort) participants, the accuracy of the "different" trials is much worse among the SA cohort compared to that of the Chinese cohort.

To further explore the effects of individual tone pair contrasts, a second regression model (Model 2) focused on the "different" trials only. Like Model 1, Model 2 also included logRT and DOMINANT language as predictors, as well as TONE PAIR (15 levels). The inclusion of the interaction between DOMINANT language and TONE PAIR did not improve model likelihood and was therefore not included in the final model. The model also includes by-subject random intercepts and by-item random intercepts, as well as by-subject random slopes for logRT. A model including by-subject random slopes for TONE PAIR did not converge. The model formula in `lme4` format is `Accuracy ~ LOGRT + TONE PAIR + DOMINANT + (1 + LOGRT | Participant) + (1|Word)`.

Figure 3 illustrates the discrimination pattern by tone pair between the two cohorts. Like Model 1, Model 2 also showed a significant effect of logRT (β=-0.14 z = -2.80, $p < 0.01$) and the the effects of language dominance are also consistent with those found in Model 1. Of particular interest is the fact that tonal discrimination is poorest between T2/T5 and between T3/T6, T4/T5 and T4/T6. Given that the addition of an interaction between TONE PAIR and DOMINANCE did not improve model likelihood, it suggests that the pattern of discrimination between tone pairs do not differ in significant ways across participants with different dominant languages.

4. Discussion

Overall, our findings show that the SA cohort exhibits different perception and production patterns relative to the Chinese cohort. In particular, the SA cohort appears to have neutralized the six-way tonal contrast to a three-way contrast in production. Specifically, The two rising tones (T2 and T5) are neutralized and so are the contrasts between the lower register tones (T3, T4, T6). In general, regardless of cohort, tonal discrimination is poorest between contour tones (i.e. T2/T5) and between the lower register tones (i.e. between T3/T6, T4/T5, T4/T6, T5/T6). The SA participants show poorer tonal discrimination accuracy relative to the Chinese cohort. Among the SA participants, the Punjabi-dominant participants exhibited the most difficulty with tonal discrimination relative to other SA participants, a finding consistent with the predictions of PAM and SLM. That is, the fact that Punjabi has contrastive tones might have hindered the learning of the Punjabi speakers' learning of Cantonese tones.

The fact that the SA cohort exhibits instances of tonal mergers brings to mind the abovementioned on-going tonal mergers that are happening in Cantonese among ethnic Chinese speakers (e.g., [5]). In particular, the two rising tones T2/T5, the two level tones (T3/T6), and the low falling and low level tones (T4/T6) are all undergoing mergers, although the changes appear to be in its early stage of development. To be sure, as [5] noted, the ongoing tone mergers in Hong Kong Cantonese is still at the beginning stage and there are still six tone categories even among young speakers who are merging T2 with T5, T3 with T6, and T4 with T6 in production. The tonal mergers observed among the SA cohort are much more complete. Also, as the tonal discrimination data suggests, the SA cohorts exhibit much poorer discrimination accuracy than the Chinese cohorts

5. Acknowledgments

We are grateful to the students and staff at Delia Memorial School (Glee path) and FDBWA Szeto Ho Secondary School for their participation in this study and the Standing Committee on Language Education and Research (SCOLAR) for funding this project.

6. References

[1] A. J. Yue Wang and J. A. Sereno, "L2 acquisition and processing of Mandarin Chinese tones," in *Handbook of Chinese psycholinguistics*, P. Li, L. anad E. Bates, and O. Tzeng, Eds. Cambridge: Cambridge University Press, 2006, pp. 250–257.

[2] Y. Hao, "Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers," *Journal of Phonetics*, vol. 40, pp. 269–279, 2012.

[3] S. Carmichael, *Language rights in education: a study of Hong Kongs linguistic minorities*, ser. Occasional Paper. Hong Kong: Centre for Comparative and Public Law, the University of Hong Kong, 2009, vol. 19.

[4] R. S. Bauer and P. K. Benedict, *Modern Cantonese Phonology*, ser. Trends in Linguistics: Studies and Monographs 102. Berlin and New York: Mouton de Gruyter, 1997.

[5] P. Mok, D. Zuo, and P. Wong, "Production and perception of a sound change in progress: tone merging in Hong Kong Cantonese," *Language Variation and Change*, vol. 25, pp. 341–370, 2013.

[6] S. Gill, "Punjabi tonemics," *Anthropological Linguistics*, vol. 2, no. 6, pp. 11–18, 1960.

[7] S. Lata, *Challenges for design of pronunciation lexicon specification (PLS) for Punjabi language*. Department of Information Technology, Govt of India, 2011.

[8] A. Singh, D. Pandey, and S. S. Agrawal, "Analysis of punjabi tonemes," in *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference, Mar 1113, New Delhi, India*, 2015, pp. 1694–1697.

[9] M. S. Rafi, "Semantic variations of Punjabi toneme," *Language in India*, vol. 10, pp. 56–65, 2010.

[10] C. Best, "The emergence of native-language phonological influences in infants: A perceptual assimilation model," in *The development of speech perception: The transition from speech sounds to spoken words.*, J. C. Goodman and H. C. Nusbaum, Eds. Cambridge: MIT Press, 1994, pp. 167–224.

[11] J. E. Flege, "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issue in Cross-Language Research*. Timonium, MD: New York Press, 1995, pp. 233–277.

[12] O. S. Bohn and J. E. Flege, "Perception and production of a new vowel category by adult second language learners," in *Second-Language Speech: Structure and Process*. Berlin, Germany: Mouton de Gruyter, 1996, pp. 53–73.

[13] C. B. Chang, Y. Y. E. F. Haynes, and R. Rhodes, "Production of phonetic and phonological contrast by heritage speakers of Mandarin," *Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 3964–3980, 2011.

[14] C. I. Watson and J. Harrington, "Acoustic evidence for dynamic formant trajectories in Australian English vowels," *Journal of the Acoustical Society of America*, vol. 106, pp. 458–468, 1999.

[15] J. Harrington, *Phonetic analysis of speech corpora*. Chichester, UK: Wiley–Blackwell, 2010.