



INTERNATIONAL JOURNAL OF
MODERN EDUCATION
(IJMOE)

www.gaexcellence.com/ijmoe



ENGLISH VOWEL CONTRASTS IN CANTONESE L1 LEARNERS: ACOUSTIC SEPARATION AND PROCESSING STRATEGIES

Ye Yizhou¹, Latisha Asmaak Shafie^{2*}, Razlina Razali³

¹Academy of Language Studies, Universiti Teknologi MARA Perlis Branch, Malaysia

 2023516063@uitm.edu.my

 <https://orcid.org/0009-0006-5863-1958>

²Academy of Language Studies, Universiti Teknologi MARA Perlis Branch, Malaysia

 ciklatisha@uitm.edu.my

 <https://orcid.org/0000-0003-1188-8095>

³Academy of Language Studies, Universiti Teknologi MARA Perlis Branch, Malaysia

 razlinarazali@uitm.edu.my

 <https://orcid.org/0000-0002-1671-1793>

*Corresponding Author

Article Info:

Article history:

Received date: 15.07.2026

Revised date: 05.08.2026

Accepted date: 20.09.2026

Published date: 28.09.2026

To cite this document:

Ye, Y., Shafie, L. A., & Razali, R.
(2026). English Vowel Contrasts
in Cantonese L1 Learners:
Acoustic Separation and
Processing Strategies.
*International Journal of Modern
Education*, 8(31), 1719-1745.

DOI: 10.35631/IJMOE.831100

Abstract:

Cantonese first-language learners of English often struggle with vowel contrasts not clearly represented in Cantonese. This exploratory mixed-methods study investigated the production and self-reported processing of three English contrasts: /i:/-/ɪ/, /æ/-/ɛ/, and /ʌ/-/ɑ:/. Acoustic data from seven speakers were analysed for F1/F2 values and Euclidean distances, and interviews with ten learners were coded for vowel-processing strategies. Results showed reduced vowel spacing and substantial spectral overlap, especially for /i:/-/ɪ/. The interviews revealed reliance on context clues and L1-based assimilation but limited articulatory awareness and formal phonetic training. These findings suggest that the difficulties Cantonese learners have with English vowels involve both acoustic crowding and conscious compensation.

Keyword:

Acoustic Production, Cantonese L1 Learners, English Vowel Contrasts, Formant Analysis, Learner Strategies, Second Language Pronunciation, Speech Learning Models



© The authors (2026). This is an Open Access article distributed under the terms of the Creative Commons Attribution (CC BY NC) (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact ijmoe@gaexcellence.com.

Introduction

English is widely taught as an additional subject in Cantonese-speaking areas, such as Hong Kong, Macao and some parts of Guangdong Province; most learners are educated in schools and do not live in English-speaking environments. Therefore, the learning of pronunciation may be significantly affected by how Cantonese phonological features relate to English sound distinctions in this environment. Cantonese-speaking learners of English often have trouble with vowel contrasts that are not phonologically contrastive in Cantonese. These contrasts include English /i:/-/ɪ/ as in beat-bit, /æ/-/ɛ/ as in bat-bet, and /ʌ/-/ɑ:/ as in cut-cart. These differences are regions in the vowel space that do not map directly to Cantonese vowel categories. Cantonese lacks a tense-lax vowel contrast such as English /i:/-/ɪ/, does not have a direct equivalent of the low-front /æ/, and does not contain a central /ʌ/ vowel (Bauer & Benedict, 1997; Chan & Li, 2000; Lee, 2016). As a result, Cantonese learners may have difficulty distinguishing these contrasts in perception and producing sufficiently different vowel qualities in English (Chan, 2012, 2013; Wong, 2012, 2013, 2014).

The two representative frameworks help explain why these contrasts (/i:/-/ɪ/, /æ/-/ɛ/, and /ʌ/-/ɑ:/) are difficult. The revised Speech Learning Model (SLM-r; Flege & Bohn, 2021) proposes that the formation of L2 phonetic categories is influenced by the perceived relationship between L1 and L2 sounds, and similar L2 sounds are often more difficult because they may be assimilated into existing L1 categories. The Perceptual Assimilation Model for L2 learners (PAM-L2; Best & Tyler, 2007) also holds that discrimination is based on how non-native contrasts are mapped onto L1 categories. Under PAM-L2, contrasts in which two L2 sounds are assimilated to a single L1 category are expected to be particularly difficult, and contrasts with clear category differences should be less problematic. For Cantonese learners of English, the above frameworks indicate that /i:/-/ɪ/ may be more prone to merger or overlap, while /æ/-/ɛ/ and /ʌ/-/ɑ:/ may have different types of mapping difficulty depending on their relationship with existing Cantonese vowel categories.

Previous studies have provided some evidence of Cantonese learners' perception and learning of English vowels. Research on perceptual assimilation has shown that Cantonese learners often map English vowels onto existing Cantonese categories in a way that aligns with PAM-L2 predictions (e.g., Chan, 2012, 2013; Wong, 2012). Phonetic training studies have also found that difficult English vowel contrasts, such as /i:/-/ɪ/ and /æ/-/ɛ/, can be improved through high-variability phonetic training. Therefore, the English vowel categories of Cantonese learners are flexible and can be modified by targeted input. However, less is known about the baseline acoustic realisation of multiple difficult vowel contrasts among learners who have not received such targeted training. In particular, relatively little research has combined contrast-specific acoustic production measures with learners' own reports on how they process difficult English vowels (e.g., Wong, 2013, 2014). The above deficiencies are problematic because, without a baseline of contrast-specific acoustic evidence linked to learners' reported processing

strategies, it is impossible to determine whether the learners' difficulties stem from L1 category assimilation, inadequate cue weighting, limited articulatory control, or compensatory reliance on spelling and context. Addressing this problem matters for both theory and practice. Theoretically, SLM-r and PAM-L2 make contrast-specific predictions about which L1–L2 mappings should be most difficult, and testing these predictions requires baseline production evidence from the very population the models are claimed to explain (Best & Tyler, 2007; Flege & Bohn, 2021). In practice, the majority of Cantonese-speaking learners of English are untrained classroom learners, and this is the group that pronunciation instruction aims to help; without contrast-specific acoustic and strategy data from this group, instructional priorities and future training interventions cannot be based on an evidence-based account of learners' starting points.

The two together are more effective because acoustic overlap alone cannot explain why learners find contrasts difficult. Overlapping productions may be due to perceptual assimilation, impaired articulatory control, reduced attention to contrast, or reliance on contextual and orthographic cues rather than vowel quality. Learners' self-reported strategies cannot solve these problems independently, but they can offer supplementary data on the types of cues and routines that learners consciously employ when facing difficult vowel contrasts. Therefore, by combining acoustic and interview data, a more all-encompassing study of how Cantonese learners handle English vowel contrasts in the absence of explicit pronunciation training can be carried out.

The current study examines Cantonese L1 learners' production and self-reported processing of three English vowel contrasts: /i:/–/ɪ/, /æ/–/ɛ/, and /ʌ/–/ɑ:/. These contrasts were selected because they represent different types of L1–L2 mapping difficulties: a tense–lax high-front contrast, a low-front contrast with /æ/, and a central-low contrast with /ʌ/. This paper addresses the following research questions.

- RQ1: How do Cantonese L1 learners acoustically realize /i:/–/ɪ/, /æ/–/ɛ/, and /ʌ/–/ɑ:/ in the F1–F2 vowel space?
- RQ2: What self-reported strategies do these learners use to deal with difficult English vowel contrasts?
- RQ3: What is the connection among group-level acoustic features and the distribution of self-reported vowel-processing strategies?

This study combines F1–F2 acoustic analysis and semi-structured interview data to provide an initial understanding of how Cantonese learners produce and process English vowels in this study. Unlike most previous acoustic research on this group, which has studied one contrast at a time (Wong, 2012, 2013, 2014), the present study provides baseline production data and strategy data for classroom-taught learners in mainland China across three contrasts and aligns the two types of evidence at the group level. It provides contrast-specific production evidence for a learner group that has been neglected in acoustic L2 vowel research and shows how learners' conscious strategy reports can supplement phonetic models such as SLM-r and PAM-L2. The results will provide a descriptive foundation for future studies with larger samples, linked perception–production measures, and targeted pronunciation interventions.

Literature Review

L2 Vowel Learning and Cross-linguistic Mapping

Models of second-language speech learning agree that the learning of L2 vowels is influenced by the connection between learners' first-language phonetic categories and the target-language vowel system. The revised Speech Learning Model, or SLM-r, holds that L2 sounds are difficult to learn when they are perceived as being the same as existing L1 categories rather than as separate phonetic units (Flege, 1995; Flege & Bohn, 2021; Flege et al., 1997). Similarly, the Perceptual Assimilation Model for L2 learners, or PAM-L2, also holds that the difficulty of discrimination is related to how two L2 sounds are assimilated into L1 categories. When two L2 vowels are assimilated to a single L1 category, discrimination is expected to be particularly difficult; when they differ in category goodness or map to different L1 categories, discrimination is expected to be easier (Best, 1993, 1994; Best & Tyler, 2007).

These models provide helpful predictions for the learning outcomes of the three English contrasts studied by Cantonese speakers. These predictions are also supported by descriptions of the Cantonese vowel system, which show that Cantonese lacks a tense–lax high front vowel contrast similar to English /i:/–/ɪ/, does not have a direct equivalent of English /æ/ or /ʌ/, and contains a low /a:/ vowel that may partially correspond to English /ɑ:/ (Bauer & Benedict, 1997; Zee, 1999, 2003). Previous research on the English vowel perception and training of Cantonese learners has also found that English contrasts involving non-equivalent or weakly differentiated vowel categories, especially tense–lax contrasts such as /i:/–/ɪ/, are difficult but learnable for this group of learners (Chan, 2012, 2013; Wong, 2012, 2013).

A further model, Language-Specific Linguistic Perception (L2LP), is concerned with how learners use and reweight acoustic cues in the development of L2 perceptual categories (Escudero, 2007). Therefore, the problem of L2 vowel learning is not whether a new category can be formed, but whether learners pay attention to the acoustic cues that distinguish the relevant categories. English vowel contrasts that require learners to use spectral cues, such as F1 and F2, are particularly susceptible to this problem, and previous studies have shown that L2 listeners give different weights to spectral and durational cues in perceiving English vowels (Bohn, 1995; Escudero, 2005; Escudero & Boersma, 2004; Morrison, 2008).

SLM-r, PAM-L2 and L2LP together suggest that Cantonese-speaking learners' English vowel production may exhibit both L1-based category assimilation and incomplete cue reweighting, as previously shown in general studies of L2 vowel perception and production and in particular studies on Cantonese-speaking learners of English (Best & Tyler, 2007; Chan, 2012; Escudero, 2005; Flege, 1995; Morrison, 2008; Wong, 2013).

Cantonese English Vowel Mismatch and Acoustic Evidence

The Cantonese and English vowel systems differ substantially in both inventory size and contrast structure. These differences have long been linked to pronunciation problems for Cantonese-speaking learners of English (Chan & Li, 2000). Cantonese generally has about seven or eight monophthongs, while English varieties, such as Southern British English, have a larger inventory of monophthongs (Bauer & Benedict, 1997; Deterding, 1997; Hillenbrand et al., 1995; Lee, 2016). Three differences are particularly relevant to the present study for

Cantonese-speaking learners of English: Cantonese lacks an English-like tense–lax contrast, has no direct equivalent of /æ/, and does not contain a vowel close to English /ʌ/.

Therefore, the /i:/–/ɪ/ contrast is expected to be more susceptible to overlap because both vowels are relatively close in the high front area of the vowel space. Previous studies on Cantonese learners have reported difficulties with this contrast in both perception and production, indicating that learners may not consistently maintain the spectral distinction between the two vowels (Chan, 2012, 2013; Wong, 2012, 2013, 2014). There is also a problem with the /æ/–/ɛ/ contrast; although /ɛ/ has a closer Cantonese equivalent, /æ/ is more peripheral and less directly represented in the Cantonese vowel system. Therefore, there may be some assimilation, unstable category formation, or reliance on non-spectral cues.

/ʌ/ and /ɑ:/ are close in sound and are often confused with the Cantonese /ɐ/–/a:/ contrast. Cantonese /a:/ is generally considered a low central-to-back vowel, and /ɐ/ is a shorter and slightly higher central vowel (Lee, 2016). At first glance, this contrast may seem to provide a good basis for English /ʌ/–/ɑ:/. The Acoustic Alignment is not ideal. English /ɑ:/ is generally lower and more back than /ʌ/, and Cantonese /ɐ/ and /a:/ differ in both vowel quality and duration. As a result, Cantonese learners may only partially map English /ʌ/–/ɑ:/ onto Cantonese /ɐ/–/a:/, and produce an unstable or compressed distinction in the F1–F2 space.

Empirical studies of Cantonese English vowel production support a contrast-specific approach. Wong's research on high-variability phonetic training indicated that Cantonese ESL learners had a foundational difficulty with contrasts such as /ɪ/–/i:/ and /æ/–/ɛ/, but targeted training could enhance both their perception and production (Wong, 2012, 2013, 2014). More recently, some studies have shown that Cantonese English bilingual vowel production is affected by vowel-specific, speaker-specific and sociolinguistic factors rather than a simple global L1–L2 distance effect (Cao, 2024; Law et al., 2021; Mok et al., 2023). Therefore, based on the above results, it is necessary to study the contrasts of individual English vowels separately and not treat them uniformly.

Studies of other Chinese L1 learners provide additional context but cannot be directly generalised to Cantonese learners. Mandarin-speaking learners, for example, have been shown to rely heavily on durational cues in distinguishing some English vowel contrasts, such as /i:/–/ɪ/ and /ɑ:/–/ʌ/ (Feng & Wang, 2024). Cantonese and Mandarin have different vowel systems and cue structures; therefore, findings from Mandarin learners cannot be assumed to transfer to Cantonese learners. Ji and Jiang (2022) found that Cantonese and Mandarin speakers showed different uses of spectral cues in their discrimination of English vowels. Therefore, Cantonese-specific acoustic data need to be collected to learn how Cantonese speakers produce the target contrasts.

Although previous research has studied the English vowel perception and production, perceptual training and bilingual vowel convergence of Cantonese-speaking learners (Baker & Trofimovich, 2005; Chan, 2012, 2013; Georgiou & Savva, 2026; Wong, 2012, 2013), few studies have jointly investigated contrast-specific F1–F2 patterns for /i:/–/ɪ/, /æ/–/ɛ/, and /ʌ/–/ɑ:/ in the same sample of Cantonese-speaking learners.

Cognitive-Perceptual Strategies in Vowel Processing

L2 learners do not rely only on stable phonetic categories when processing unfamiliar vowel contrasts. When perceptual categories are unavailable, people may use compensation mechanisms, such as focusing on vowel duration, relying on spelling, using lexical or sentence context, or imitating articulatory patterns taught in class. These strategies are feasible because they can help learners complete a task without learning native-like phonetic categories.

Cue-weighting research has shown that L2 learners may use acoustic dimensions differently from native speakers in their learning. Morrison (2008) found that L1 Spanish learners of English paid more attention to duration when perceiving the English /i:/–/ɪ/ contrast, even in the presence of spectral cues. This line of research is relevant to Cantonese-speaking learners because it shows that improving production ability does not necessarily lead to native-like perception cue weighting; that is to say, even after enhancing production skills, L2 learners may continue to use non-native cue-weighting strategies (Georgiou, 2021; Georgiou & Savva, 2026; Llompert, 2021; Nagle, 2018).

Phonetic training literature also indicates that perception and production are enhanced through organised exposure to variable input for learners. Bradlow et al. (1997, 1999) have determined that perceptual training can boost both the perception and production of difficult L2 contrasts. Wong's work with Cantonese ESL learners also showed that high-variability phonetic training improved their performance in English vowel contrasts (Wong, 2012, 2013, 2014). Recently, research in training has also shown that production practice with direct articulatory feedback, such as real-time ultrasound imaging of the tongue, can enhance both the production and perception of English tense–lax contrasts, and these benefits continue after the training ends (Chang, 2023). Training studies have shown what learners can achieve in a guided-learning environment. They do not necessarily show what strategies learners spontaneously use in their daily learning without explicit phonetic instruction.

Recent research on phonological representations has also shown that learners' production and perception do not always develop simultaneously. Some learners can produce discernible acoustic differences without fully forming stable phonological representations, and others may perceive a contrast but fail to apply it consistently in speech (Cook et al., 2016; Simonchyk & Darcy, 2023). Given this possible mismatch, it is necessary to investigate how learners themselves identify and recall English vowel contrasts, as well as how they produce them.

The Production-Perception Interface

The connection between L2 vowel perception and production is not necessarily direct. Some models and empirical studies suggest that perceptual category formation often supports later production development, while other studies show that production and perception can be partially dissociated (Casillas, 2020; Simonchyk & Darcy, 2023; Song & Eckman, 2024). This means that acoustic production data alone may not fully explain how learners process L2 vowel contrasts.

For Cantonese-speaking learners of English, the production–perception relationship may vary by contrast. A learner can make a relatively clear /i:/–/ɪ/ distinction in production but reports using spelling or memorised words to indicate the difference. Another learner may be aware

that /æ/ and /ɛ/ are different but still produce them close together in the F1–F2 space. These patterns are suitable for combining with the interview data.

Both qualitative and quantitative research will be used in combination for this study. Acoustic analysis is employed to describe how learners achieve the three vowel contrasts in F1–F2 space, and interviews are conducted to identify the strategies learners report using during the processing of English vowels. The two data sources are compared to determine whether reduced acoustic separation is associated with a particular strategy profile, such as dependence on L1 categories, spelling, context or limited articulatory awareness.

Summary and Rationale

The three problems in the literature identified above provide the motivation for this study. First, the learning of Cantonese-English vowels is strongly influenced by cross-linguistic category mapping, but the three contrasts examined here differ in their predicted difficulty and in the clarity of their L1–L2 mappings. Second, previous studies have shown that Cantonese-speaking learners may have difficulty with contrasts such as /i:/–/ɪ/ and /æ/–/ɛ/, but little is known about how /i:/–/ɪ/, /æ/–/ɛ/, and /ʌ/–/ɑ:/ patterns are realised acoustically in the same learner sample. Third, the learners' own strategies for handling these contrasts are not known, especially in terms of acoustic production patterns.

Therefore, the current study integrates F1–F2 acoustic analysis with semi-structured interview data. This design can help the study describe learners' production of the target contrasts and examine how this production patterns are related to reported strategies, such as reliance on spelling, context, articulatory awareness, and first-language categories. Addressing these gaps is not merely descriptive; we need to know which contrasts have been merged spectrally and what strategies learners actually employ in order to design targeted pronunciation instruction and perception–production training for Cantonese classroom learners.

Methods

Research Design

Based on the theory-informed mixed-methods design, acoustic phonetic analysis of Cantonese-speaking learners' English vowel production was conducted in conjunction with qualitative interview data on their self-reported vowel-processing strategies. Quantitatively, F1–F2 Euclidean distance was used to examine the acoustic separation of selected English vowel contrasts, and qualitatively, learners' reported reliance on phonological, orthographic, durational, contextual, and articulatory cues was investigated.

The design is based on cross-linguistic phonological comparisons of Cantonese and English vowel systems, as well as three models of second-language speech learning: SLM-r, PAM-L2, and L2LP. In short, the above models indicate that Cantonese-speaking learners' English vowel production may show both L1-based category assimilation and incomplete cue reweighting. Figure 1 shows the theory-based analytical framework that guides the current study, connecting cross-linguistic vowel mapping, perceptual categorisation, acoustic production patterns, and learners' self-reported processing strategies.

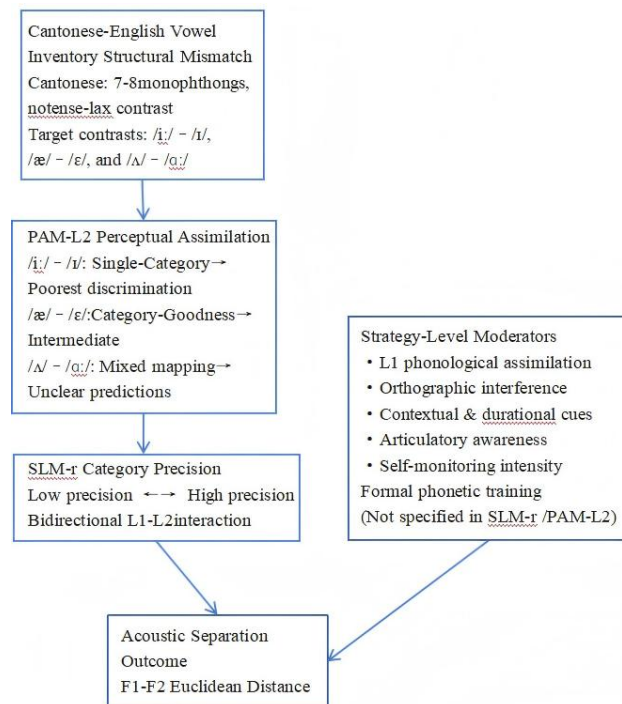


Figure 1: Theory-Informed Analytical Framework Of The Present Study

An interview section was added to provide the context for the acoustic results rather than directly presenting perception data. Interview responses were used to explore learners' metalinguistic awareness, perceived difficulty of vowels, and reported reliance on spelling, context, duration, articulatory knowledge or first-language categories. This part addresses the second research problem directly.

Participants and Sampling

A total of ten participants were recruited from a comprehensive university in Guangdong Province, mainland China, and formed the entire sample. All were native Cantonese speakers enrolled in non-English major programs and had received at least ten years of formal EFL education. None reported speech, language or hearing impairment, and none had lived in an English-speaking country for more than three months. English proficiency was not used as a selection criterion; this was done intentionally because the study aimed to characterize the pronunciation of learners who had acquired English primarily through classroom exposure, and screening for high proficiency would have restricted the very range of performance the study intended to explore. The two subsets of participants were partially overlapping. The interview subset included all ten participants, aged 19–23 years ($M = 20.8$), six female and four male participants, coded SA-SJ. The production subset included seven of these participants, aged 19–22 years ($M = 20.6$), four female speakers (S1, S5, S6, and S7) and three male speakers (S2, S3, and S8). All ten participants completed the interview; participation in the production recording was voluntary and required a separate session under controlled recording conditions. Eight people were present at the recording session; one additional speaker (S4) was recorded but excluded due to poor recording quality, so seven interviewees participated in the production. The seven retained speakers were in the same gender distribution and age range as the full interview sample. Because the production and interview data were anonymised using separate coding schemes, individual-level links between the two coding systems could not be

established; therefore, alignment analysis for RQ3 had to be conducted at the group level instead of the individual level. Written consent was obtained from all participants before the start of data collection, and they were informed in writing about the purposes and procedures of the study. Participation was voluntary, and participants are identified in this paper only by code. The steps of the study comply with the ethical provisions of the host institution.

Data Collection Procedures

The target words were embedded in the carrier sentence "Say-again" and shown visually in a random order in the three blocks. After some practice, participants could read each sentence at a normal speaking rate. Individually in a quiet room, recordings were made with a digital recorder at 44.1kHz and 16-bit resolution, and an external condenser microphone was placed approximately 15 cm away from the participant's mouth.

Praat is a speech analysis program (Boersma & Weenink, 2023). Target vowels were manually segmented following standard criteria based on the onset and offset of periodic voicing and visible formant structure. F1 and F2 were taken at the vowel midpoint for relatively stable formant structure. Formant ceilings were set to 5000 Hz for male speakers and 5500 Hz for female speakers, and five formants were tracked using the Burg LPC algorithm. All formant tracks were visually compared with wideband spectrograms. Eleven tokens, accounting for 6.5% of the dataset, required manual correction due to formant-tracking errors; no planned tokens were excluded.

Speaker-level mean F1 and F2 values for each vowel category were calculated by averaging over tokens. In total, the seven speakers produced 24 target words each: four /ʌ/ words, four /ɑ:/ words, three /i:/ words, five /ɪ/ words, four /æ/ words, and four /ɛ/ words. One lexical item, 'velvet', contains both /ɛ/ and /ɪ/ in consecutive syllables and thus contributes one token to each of these two categories. Three or five lexical items were selected for each vowel category to reduce word-specific effects. The full target word list is presented in Appendix A.

For each contrast, the spectral separation is taken as the Euclidean distance between the vowel centroids in the raw F1–F2 space: $ED = \sqrt{[(F1_1 - F1_2)^2 + (F2_1 - F2_2)^2]}$. A relatively large ED value indicates that the two sets of vowels are spectrally separated. The analysis focused on F1 and F2 only; duration and F3 were not extracted. Therefore, the acoustic results should be viewed as spectral patterns rather than as a full-fledged system for analysing contrastive vowel realisation.

Deterding's Standard Southern British English monophthong data were used as the reference values and served as descriptive benchmarks. As no native-speaker control group was recorded under the same circumstances, these reference values could not be used for inference. Given the small production sample, only descriptive statistics and gender-stratified comparisons of acoustic data were performed to mitigate the effect of anatomical formant differences. No vowel normalisation was carried out.

For visualisation and overlap quantification, confidence ellipses were generated based on the token-level F1 and F2 values of each vowel category. The centre of each ellipse was set to the mean F1–F2 centroid of the category, and the orientation and axis lengths of the ellipse were obtained from the covariance matrix of F1 and F2 under a bivariate normal approximation. Figure 2 is the 68% confidence ellipse; the central part of the token distribution for each

category and a typical location and spread of the category in the F1–F2 space are contained within this ellipse. The 95% ellipses in the overlap-ratio analysis were also generated using the same method. For each contrast, the overlap ratio was calculated as the area of intersection of the 95% ellipses for the two categories divided by the area of the smaller ellipse; thus, a ratio of 1.000 indicates complete overlap, and a lower value indicates greater separation. The ellipses, overlap ratios and the corresponding figures were all generated by custom Python scripts (matplotlib for plotting).

Given that Chinese was the participants' convenient functional bilingual language, semi-structured interviews were conducted in Chinese to explore their awareness of and sense for English vowels and sounds more deeply. Interviews lasted 20–35 minutes ($M = 28.0$, $SD = 6.2$) and were audio-recorded and transcribed verbatim. The interview protocol listed the strategies learners reported for processing English vowel contrasts, such as L1-based mapping, awareness of category differences, contextual reliance, guessing patterns, spelling effects, articulatory awareness, self-monitoring and formal phonetic training. The complete list of guiding questions for the interviews is provided in Appendix B. The protocol included prompts with /i:/–/ɪ/ and /æ/–/ɛ/ minimal pairs but deliberately omitted the dedicated /ʌ/–/ɑ:/ prompts. After piloting, learners rarely noticed the /ʌ/–/ɑ:/ contrast spontaneously, and when directly prompted, they tended to give recognition-based guesses rather than reports of an existing processing strategy. However, the contrast was still covered by the acoustic production task (target words in Appendix A), and any spontaneous mention of /ʌ/–/ɑ:/ during the interviews was coded; such mentions were treated as informative evidence of how seldom this contrast was consciously noticed, rather than as a measure of learners' ability to distinguish it.

Directed content analysis (Hsieh & Shannon, 2005) was used for the transcript. Based on the strategy categories in the Literature Review, an initial coding scheme was developed deductively and then refined through iterative reading of the transcripts. Eight dimensions were coded: L1 Assimilation Strength, Category Differentiation Awareness, Contextual Reliance Level, Default Guessing Bias, Orthographic Interference, Articulatory Awareness, Self-Monitoring Intensity, and Formal Phonetic Training. Each dimension was rated on a 0-3 qualitative scale, and 0 indicated the absence of the feature while 3 indicated a strong or organised presence. The coding dimensions and anchors are presented in Table 1.

The first author coded all of the transcripts. To improve the consistency of coding, a coding manual with anchor examples was prepared before the official coding; analytic memos were collected during the coding, and the entire dataset was re-coded after a two-week interval. Intra-coder agreement reached more than 90%. Borderline cases were handled by the more conservative score, and representative excerpts and coded matrices were checked for peer debriefing with the co-authors. Since no independent second coder was hired, the qualitative coding should be considered exploratory.

Table 1: Coding Dimensions, Codes and Anchors

Dimension	Code	0	1	2	3
L1 Assimilation Strength	L1AS	No mapping	L1Occasional reference	L1Frequent mapping	L1Systematic substitution
Category Differentiation Awareness	CDA	Perceives sounds same	Knows as difference but vague	Can describe exists acoustic difference	Can explain articulatory distinction

Contextual Reliance Level	CRL	Relies on phonetic detail	Sometimes context	uses Strong reliance	context Cannot discriminate without context
Default Guessing Bias	DGB	No guessing pattern	Occasional guessing	Clear bias	Fixed systematic guessing
Orthographic Interference	OI	No spelling effect	Minor interference	Clear interference	Strong spelling reliance
Articulatory Awareness	AA	No articulatory reference	General awareness	Specific articulatory description	Technical articulatory explanation
Self-Monitoring Intensity	SMI	No monitoring	Occasional correction	Frequent monitoring	Continuous deliberate monitoring
Formal Phonetic Training	FPT	None	Basic phonics instruction	Systematic pronunciation training	Technical phonetic instruction

According to the principal patterns of the eight dimensions after coding, the participants were grouped into qualitative strategy profiles. A score of 2 or more was considered to indicate a substantial feature of a participant's profile, and scores of 0-1 were treated as absent or weak. Profiles were used descriptively and not converted into composite numerical indices.

Data Analysis

To answer RQ3, acoustic and interview data were combined in a group-level joint display. As the individual links between production and interview codes were not stored, it could not be determined whether a particular learner's acoustic pattern matched the strategies they reported. Instead, the distribution of Euclidean distances in the production subset was compared with the distribution of strategy profiles in the interview subset. Based on this comparison, descriptive group-level patterns were identified, such as the convergence of reported strategy resources and acoustic differentiation, the divergence of reported awareness and acoustic overlap, and cases where acoustic separation occurred without strong reported awareness. No inferential statistics were used. Therefore, the alignment analysis should be considered as an exploratory, group-level phenomenon rather than proof of individual production-perception links.

Results

This section presents the findings in three parts, corresponding to the three research questions. The first section presents learners' acoustic realisations of the three target vowel contrasts in the F1–F2 vowel space. The second part is the learners' self-reported vowel-processing strategies from the interview data. The third is the group's acoustic pattern in relation to the distribution of these strategies.

Findings of Research Question

1: How do Cantonese L1 learners acoustically realize /i:/-/ɪ/, /æ/-/ɛ/, and /ʌ/-/ɑ:/ in the F1-F2 vowel space?

The production data showed that learners placed the six target vowels in the generally expected areas of the F1-F2 space, but the contrastive categories had reduced separation. Figure 2 shows the 68% confidence ellipses for learner and native-speaker reference values by gender. The learner /i:/ category formed a relatively compact cluster, while /ɪ/ was more dispersed and extended to both /i:/ and /ɛ/. The front low-to-mid region had significant overlap between /æ/ and /ɛ/. The /ʌ/ category had the largest spread among the six vowels and extended both towards /ɑ:/ and /ɛ/.

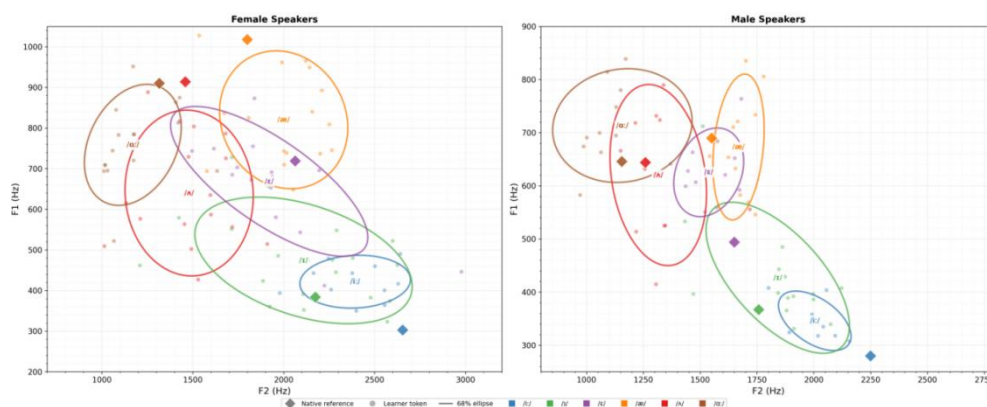


Figure 2: F1 - F2 68% Confidence Ellipses for Learners and Native-speaker Reference Values by Gender

Overlap Ratios based on 95% confidence ellipses are shown in Figure 3. For both female and male speakers, there was complete overlap of /i:/ and /ɪ/, and the overlap ratio in both groups was 1.000. The overlap ratio for /æ/ and /ɛ/ was relatively large, at 0.529 for female speakers and 0.601 for male speakers. There is a relatively large overlap between /ʌ/ and /ɑ:/, and the proportion of female speakers to male speakers is 0.730 and 0.653, respectively. These results show that /i:/-/ɪ/ was the most crowded contrast in F1-F2 space, followed by /ʌ/-/ɑ:/ and /æ/-/ɛ/.

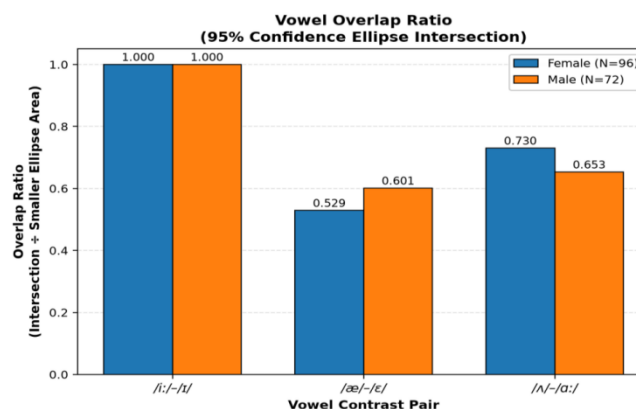


Figure 3: 95% Confidence Ellipse Overlap Ratios by Contrast and Gender

Dimension-wise comparisons showed a compressed F1 range for the learner vowel system. As shown in Figure 4, learner /i:/ and /ɪ/ had higher F1 values than the native-speaker reference values and were less closely aligned. /ɛ/, /æ/, /ʌ/ and /ɑ:/ generally had lower F1 values than the reference values and were thus considered raised productions. In this way, the entire range of the learners' vowel space was reduced. F2 patterns were not uniform. As shown in Figure 5, /i:/ showed full backing, /ɪ/ and /ɛ/ showed slight backing, and /æ/ and /ʌ/ showed fronting. Therefore, the learner system exhibited vertical compression and contrast-specific horizontal rearrangement simultaneously.

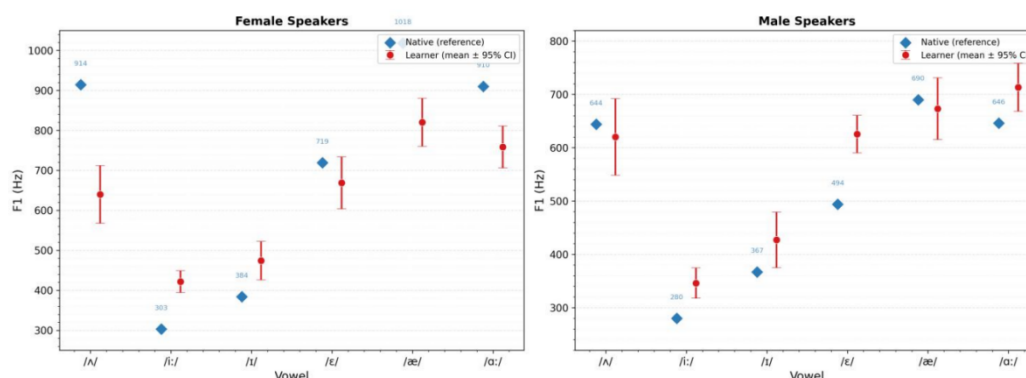


Figure 4: F1 Values for Learner and Native-speaker Reference Vowels by Gender

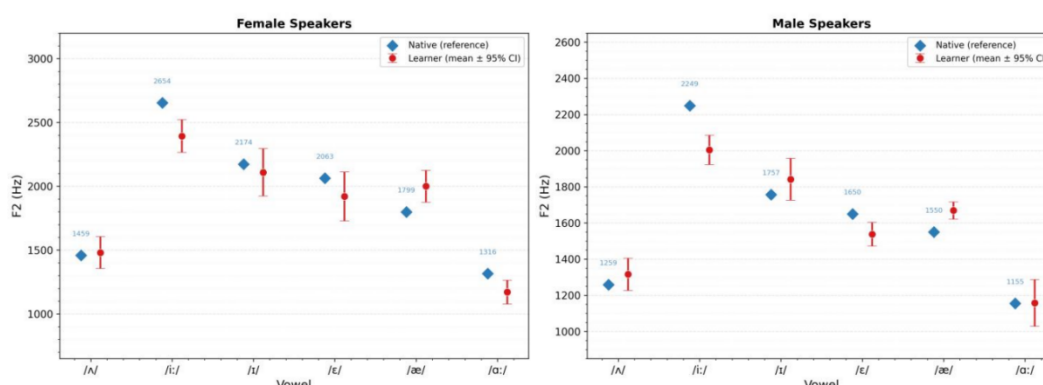


Figure 5: F2 Values for Learner and Native-speaker Reference Vowels by Gender

Speaker-level Euclidean Distances are shown in Table 2. The /i:/–/ɪ/ contrast had the largest range of individual differences among female speakers. Female EDs were 66 Hz for S6 and 507 Hz for S1, and male EDs were 167 Hz and 209 Hz. Generally speaking, the Euclidean distances (EDs) for the /æ/–/ɛ/ contrast ranged from 130 Hz to 332 Hz for female speakers and from 140 to 156 Hz for male speakers. The range of the /ʌ/–/ɑ:/ contrast for female speakers and male speakers was 168 Hz–455 Hz and 141 Hz–219 Hz, respectively.

As shown in Table 2, the directional F1 and F2 differences do not indicate that learners have learned the contrasts uniformly. /i:/–/ɪ/ differences in ED were mainly due to F2 shifts by most speakers, though the magnitude varied considerably. For /æ/–/ɛ/, most speakers produced /æ/ with a higher F1 than /ɛ/, consistent with a height-based distinction, but S8 showed a near-zero difference in the opposite direction. For /ʌ/–/ɑ:/, both F1 and F2 were required for

differentiation, but the proportion of these two dimensions varied among speakers. Generally speaking, the acoustic data have both group-level spectral crowding and substantial individual differences.

Table 2: Speaker-Level Euclidean Distances (Hz) by Contrast

Speaker (Sex)	/i:/-/ɪ/ ED	F1	F2	/æ/-/ɛ/ ED	F1	F2	/ʌ/-/ɑ:/ ED	F1	F2
S1 (F)	507	+62	-504	332	+227	+243	366	-125	+345
S5 (F)	379	+70	-373	212	+100	-188	455	-102	+444
S6 (F)	66	+61	-24	130	+130	+1	168	-62	+156
S7 (F)	236	+15	-236	303	+150	+263	348	-186	+294
S2 (M)	183	+58	-173	145	+55	+134	219	-47	+214
S3 (M)	209	+138	-157	156	+99	+120	208	-134	+159
S8 (M)	167	+47	-159	140	-12	+139	141	-99	+101
F M	297	+52	-284	244	+152	+80	334	-119	+310
M M	186	+81	-163	147	+47	+131	189	-93	+158

Note. ED = Euclidean distance in raw F1–F2 space, in Hz. F1 and F2 are the directional differences for each contrast, as shown in the column headings. A positive $\Delta F1$ value indicates that the first vowel in the directional label has a higher F1 than the second vowel in the label. Female and male values cannot be directly compared because formant normalisation was not performed.

Findings of Research Question 2: What self-reported strategies do Cantonese L1 learners use to process difficult English vowel contrasts?

According to the interview data, the learners' self-reported vowel-processing strategies mainly included context-based reliance, L1-based mapping and spelling-related interference; articulatory awareness and formal phonetic training were relatively weak. Table 3 shows the group medians and representative excerpts for the eight coded dimensions.

Table 3: Coding Dimension Medians and Representative Excerpts (N = 10)

Dimension	Mdn	IQR	Representative excerpt (translated from Mandarin)
CRL	3.00	0.50	"I judge word meaning mainly by language intuition and context. Without context, I really can't tell similar-sounding words apart." (SH)
DGB	2.00	1.00	"I always guess wrong between /æ/ and /ɛ/. It is basically fifty-fifty." (SE)
CDA	2.00	1.50	"I would start from the most basic vowel oral cavity positions, like thoroughly mastering the tongue height of /æ/ (cat) and /ɑ:/ (car)." (SB)
L1AS	2.00	2.50	"I map English sounds onto approximate Cantonese sounds to help with memory." (SI)
OI	2.00	1.50	"Spelling often misleads me. For 'debt', at first I would pronounce the 'b'." (SH)
SMI	2.00	2.50	"I specifically went to learn pronunciation, to imitate and correct myself. After practicing, I could feel the difference." (SF)
AA	1.00	2.00	"The difficulty mainly lies in controlling mouth opening and tongue position — quite different from our usual Chinese pronunciation." (SC)

FPT 1.00 1.50 "We had maybe two or three classes where the teacher showed the vowel chart. But there was no practice, no feedback." (SH)

Note. The grade of the code was 0 to 3. CRL = Contextual Reliance Level; DGB = Default Guessing Bias; CDA = Category Differentiation Awareness; L1AS = L1 Assimilation Strength; OI = Orthographic Interference; SMI = Self-Monitoring Intensity; AA = Articulatory Awareness; FPT = Formal Phonetic Training. Excerpts were translated from Mandarin Chinese.

The highest median score was found for Contextual Reliance Level (Mdn = 3.00), indicating that participants commonly reported relying on lexical or sentence context when processing difficult vowel contrasts. L1 Assimilation Strength, Orthographic Interference, Default Guessing Bias, Category Differentiation Awareness and Self-Monitoring Intensity each had a median score of 2.00, indicating moderate use or awareness among the group. Articulatory Awareness and Formal Phonetic Training had lower median scores of 1.00, respectively, indicating a lack of detailed articulatory knowledge and systematic phonetic instruction.

The three target contrasts had different patterns in the contrast-specific comments. For /i:/-/ɪ/, most participants reported that vowel duration was used as the main cue, especially for isolated words such as ship and sheep. For /æ/-/ɛ/, participants often described the difference in terms of mouth opening but also reported unstable discrimination. For /ʌ/-/ɑ:/, spontaneous comments were infrequent: only one participant mentioned this contrast without prompting, and none of the participants described a specific strategy for distinguishing it.

Findings of Research Question 3: How do the group-level acoustic patterns relate to the distribution of self-reported vowel-processing strategies?

The five qualitative strategy profiles of the participants are shown in Table 4. The most frequent profile was P1, L1-Context-dominant (n=4), which had high contextual dependence, strong L1 assimilation and low articulatory awareness. Two participants were assigned to P3, Articulatory-Monitoring, and thus had higher articulatory awareness, self-monitoring, and category differentiation awareness. Two participants were assigned to P4 (Undifferentiated), and one participant each was assigned to P2 (Orthography-Dependent) and P5 (Low-Awareness).

Table 4: Strategy Profile Distribution (N = 10)

Profile	Defining features	n	Participants
P1: L1-Context-Dominant	CRL ≥ 2, L1AS ≥ 2, AA ≤ 1, FPT ≤ 1	4	SA, SF, SI, SJ
P2: Orthography-Dependent	OI ≥ 2, elevated CRL and DGB, low AA	1	SH
P3: Articulatory-Monitoring	AA ≥ 2, SMI ≥ 2, CDA ≥ 2, FPT ≥ 1	2	SB, SD
P4: Undifferentiated	Low-to-moderate scores across all dimensions, no dominant strategy	2	SC, SE
P5: Low-Awareness	CRL = 3, AA = 0, SMI = 0, no spontaneous /ʌ/-/ɑ:/ mention	1	SG

Note. Profile Assignment was qualitative and many-faceted. Participants could show features of more than one profile.

Since the production and interview codes were not linked individually, acoustic-strategy alignment was only evaluated at the group level. The three groups did not differ significantly. First, the prevalence of L1-context-oriented profiles was in line with the general acoustic pattern of reduced spectral separation across the three contrasts. Second, articulatory awareness did not need to replace L1-based or context-based strategies; both participants in the

Articulatory-Monitoring profile also showed signs of L1-oriented processing or contextual reliance. Third, the alignments were not the same. The /i:/-/ɪ/ contrast showed complete F1-F2 ellipse overlaps but wide individual ED variation, and participants frequently reported duration-based strategies. The /æ/-/ɛ/ contrast had relatively little spectral overlap and was often expressed by mouth opening. /ʌ/ and /ɑ:/ are close in pronunciation, so this contrast was rarely mentioned spontaneously.

These patterns indicate that the connection between sound production and reported processing strategies is contrast-specific and does not follow a single uniform pattern across the three contrasts. However, because no standardised proficiency measure was administered, a general proficiency effect cannot be ruled out; as the alignment analysis was also conducted at the group level, these results should be considered descriptive rather than evidence of individual-level production-strategy fit.

Discussion

Discussion of Research Question 1: Acoustic Differentiation and Spectral Crowding

Research Question 1 is how Cantonese L1 learners acoustically realize /i:/-/ɪ/, /æ/-/ɛ/, and /ʌ/-/ɑ:/ in the F1-F2 vowel space. The results showed that the learners produced the six target vowels in the general expected areas, but the three vowel contrasts were not as well differentiated. Across the dataset, the most noticeable phenomenon was spectral crowding; learners' vowel categories occupied overlapping areas in the F1-F2 space.

The most crowded group was /i:/-/ɪ/, and the 95% ellipse overlap ratio for both female and male speakers reached 1.000. This shows that the two high front vowels are almost identical in the F1-F2 space. Generally speaking, the results support perceptual assimilation theories, such as the revised Speech Learning Model (SLM-r) and the Perceptual Assimilation Model for L2 speech learning (PAM-L2); these models predict that there will be persistent difficulties when two L2 categories are perceptually close or have already been assimilated by an existing L1 category (Best & Tyler, 2007; Flege & Bohn, 2021). The current results also extend the previous work on the difficulty of Cantonese learners with English /i:/-/ɪ/ by showing that the problem is not limited to perception or learner intuition but also appears in production as reduced acoustic separation.

The /æ/-/ɛ/ and /ʌ/-/ɑ:/ contrasts had significant but partial overlap. It can be seen that the learners have created a certain degree of spectral differentiation for these contrasts, but not a fully separate category space. This is in line with the previous studies that have found difficulties for Cantonese-speaking learners in distinguishing English vowel contrasts that do not directly correspond to Cantonese vowel categories. However, the current results also indicate a gradient pattern of difficulty rather than a uniform problem for all target vowels. Whereas /i:/-/ɪ/ showed near-complete spectral crowding, /æ/-/ɛ/ and /ʌ/-/ɑ:/ were more partially differentiated. In this regard, the study supports the previous claim of cross-linguistic influence in Cantonese learners' English vowel production but refines it by showing that different vowel contrasts may be at different positions along a continuum of merger-like overlap to partial category separation.

A second pattern concerns the F1 dimension. Learners produced high vowels with higher F1 values than the native-speaker reference values, and mid and low vowels generally had lower F1 values. This two-way compression reduced the total range of the learners' vowel system. Importantly, this pattern cannot be described as simple centralisation because the F2 shifts were contrast-specific: /i:/ was backed, /ɪ/ and /ε/ showed slight backing, and /æ/ and /ʌ/ showed fronting. A possible explanation is that Cantonese L1 learners organise English vowels in a reduced vowel-height space formed by their L1 vowel system.

The interpretation is consistent with SLM-r and PAM-L2 in that both models give more weight to the function of cross-linguistic phonetic similarity in L2 category formation (Best & Tyler, 2007; Flege & Bohn, 2021). However, the current results also indicate that L1 influence may occur in a non-one-to-one manner. The learner's vowel system appears to have been rearranged in a compressed acoustic space, particularly along the F1 dimension, in this study. Therefore, L1 influence may affect not only how individual English vowels are mapped to Cantonese categories but also the overall structure of the developing L2 vowel system.

The /ʌ/–/ɑ:/ contrast is a slightly different problem. Acoustically, there was significant spectral overlap in both cases, and the available data do not provide strong evidence that the learners failed to perceive the contrast. The overlap may be due to a lack of differentiation in production; at the same time, it could also indicate that the contrast is less prominent in classroom teaching or less frequently observed by learners. These two explanations are not the same; previous research on L2 speech has shown that production accuracy, perceptual sensitivity and conscious awareness do not always develop together.

The /ʌ/–/ɑ:/ finding should therefore be interpreted cautiously. It indicates acoustic overlap in production, but not directly whether perception has failed. This view is consistent with the notice-based accounts of second-language learning that distinguish between conscious awareness of a language problem and implicit sensitivity to input (Schmidt, 1990, 2001). Controlled perception tasks, such as AXB discrimination, identification and goodness-rating tasks, would be required to determine whether learners perceptually distinguish /ʌ/–/ɑ:/ and how they weigh F1, F2 and duration cues.

In short, the results of the research for Question 1 show that Cantonese L1 learners' English vowel production has significant but contrast-specific spectral crowding. The most severe overlap was for /i:/–/ɪ/, and /æ/–/ε/ and /ʌ/–/ɑ:/ showed partial differentiation. The extra pattern of F1 compression suggests that learners' vowel categories may be forming in a reduced vowel-height space rather than moving towards a centralised target. The above results support cross-linguistic models of L2 speech learning and also indicate that the organisation of the learner's vowel space should be viewed as a system rather than a collection of independent vowel categories.

Discussion of Research Question 2: Self-Reported Processing Strategies

The second research question is how Cantonese ESL learners themselves think about and use strategies to learn English tense–lax vowel contrasts. As shown in the results above, the learners employed various cues, such as spelling, vowel length, articulatory awareness and familiar lexical examples. The above strategies indicate that the learners were not ignorant of the target contrasts but actively tried to manage the difficult L2 vowel distinctions by combining auditory, visual, articulatory and lexical information. At the same time, the self-report data also show

that such conscious strategies do not always correspond to accurate acoustic differentiation. This pattern is in line with previous studies that have shown that the perception, production and conscious awareness of L2 learners do not always develop simultaneously (Bradlow et al., 1997; Casillas, 2020; Simonchyk & Darcy, 2023; Song & Eckman, 2024).

A typical way was to use orthographic cues. Some students reported that they have linked vowel sounds with spelling patterns, such as a silent 'e', double consonants, or familiar written words. This is understandable in an ESL context where English is often learned through reading and classroom teaching. English spelling does not always correspond to vowel sounds. Therefore, orthographic strategies can help learners remember individual words but do not support the formation of stable phonetic categories. Cook et al. (2016) believe that fuzzy nonnative phonolexical representations may lead to an unstable form-to-meaning mapping; this is also true here. Learners may have known that the two words were written differently in the past, but this knowledge did not necessarily help them distinguish between them accurately in speaking.

The second is to use a longer vowel sound. Many learners described tense vowels as "longer" and lax vowels as "shorter" and were particularly good at distinguishing between /i:/–/ɪ/ and /u:/–/ʊ/. This strategy is a typical classroom explanation for English "long" and "short" vowels. While vowel length is relatively easy to observe, it is not the only characteristic of English vowel contrasts; tense–lax differences also include spectral attributes such as vowel height and backness. Morrison (2008) showed that the acquisition of the English /i:/–/ɪ/ contrast by L2 learners cannot be reduced to an initial duration-based stage. The duration-based strategies reported here therefore help explain why some learners could describe the contrast explicitly but still produced spectrally overlapping vowels.

The problems of /i:/ and /ɪ/ were particularly severe. Learners believed that these vowels had different lengths, but the acoustic results showed that the two groups were significantly overlapping. Therefore, there is a discrepancy between explicit knowledge and phonetic realisation. Previous research has shown that it is particularly difficult for Cantonese learners to master English vowel contrasts due to differences in vowel systems and phonological structures between the two languages (Chan, 2012, 2013; Chan & Li, 2000). Wong (2013) further found that Cantonese ESL learners benefited from high-variability phonetic training in perceiving /ɪ/–/i:/, but production-only training did not lead to significant perceptual gains. The current results extend this line of research by showing that learners' self-reported length-based strategies may exist in conjunction with limited spectral differentiation in production.

Some students reported that they used articulation, such as changing mouth shapes, moving tongues, rounding lips, or feeling muscle tension. The above responses show that the learners are somewhat aware of how English vowels are articulated. This awareness can help learners focus more on the articulation of vowel quality and avoid rote memorisation. In line with previous studies on pronunciation, instruction and explicit attention have been shown to help develop L2 pronunciation (Saito & Plonsky, 2019; Thomson & Derwing, 2015). Not all Articulatory Awareness changes are reflected in speech. From the perspective of SLM-r and PAM-L2, this may be because learners need to create sufficiently distinct L2 phonetic categories, especially when two L2 vowels are perceptually close to each other or assimilated to similar L1 categories (Best & Tyler, 2007; Flege & Bohn, 2021). Therefore, one can describe an articulatory difference, but it does not mean that the contrast has been fully mastered in production.

Another way is lexical familiarity. Some learners reported comparing the target vowels with familiar words they had learned before. Therefore, they may have learned vowels by memorising exemplars rather than through abstract phonetic categories. Such a strategy may be suitable for familiar words, but it will not help with new words or new speakers. In line with Wong's (2013) study on how high-variability phonetic training helps learners apply perceptual learning to new words and speakers, this interpretation is appropriate. Reliance on familiar lexical examples may be the reason why learners could describe some contrasts but produced inconsistent acoustic patterns across items in the present study.

In short, the self-reported strategies are generally associated with a degree of metalinguistic awareness. Learners could discuss English vowel contrasts according to concepts such as length, spelling and mouth shape, and demonstrated a certain level of metalinguistic awareness. However, this consciousness did not always result in high-precision production. Schmidt (1990, 2001) thought that noticing and attention are required for L2 learning; however, it has also been shown that awareness alone does not automatically lead to stable phonetic categories. Verbal reports should also be interpreted cautiously, as learners may not have full access to the perception and motor processes that underlie their speech behaviour (Nisbett & Wilson, 1977). Therefore, the strategies reported in this study should be viewed as learners' conscious explanations of their behaviour rather than direct evidence of their actual online processing.

Discussion of Research Question 3: Acoustic-Strategy Alignment

The third research problem is how to connect the learners' self-reported vowel-processing strategies with their acoustic-production data. RQ1 collected the learners' actual vowel acoustics, and RQ2 documented their metalinguistic approaches to tense-lax contrasts; together, these two datasets offer an all-around picture of Cantonese ESL vowel acquisition. In short, the strategies did not improve the acoustic results; conscious awareness of vowel differences was often not reflected in clear spectral separation, and thus, explicit knowledge did not lead to enhanced perception or articulation.

First, length-based strategies showed a possible explanatory mismatch with the F1–F2 results. Most participants used vowel duration to differentiate between tense-lax pairs, especially /i:/ and /ɪ/. However, since the duration was not recorded in this acoustic analysis, it cannot be directly compared with the production pattern of duration. Instead, it is proposed that learners may consciously conceptualise tense-lax contrasts in terms of length but still fail to establish clear spectral separation in F1–F2 space. The difference was most prominent for /i:/–/ɪ/; although learners often described a length-based difference, the acoustic results showed significant formant overlap.

Second, articulatory awareness did not match the acoustic result. Tongue height, lip shape and mouth opening were used to distinguish vowels, and thus directly affected F1 and F2 values. Learners who actively observed articulation occasionally produced clearer vowel separation, particularly for pairs with large openness differences; however, many still showed compressed vowel spaces and overlapping categories. General articulatory knowledge cannot compensate for a lack of distinct phonetic categories, and thus declarative awareness fails to provide consistent acoustic differentiation.

Third, orthographic and lexical familiarity strategies were not related to production accuracy. Spelling patterns or memorised familiar word templates can help with word recognition and memorisation, but they do not guarantee the correct pronunciation of vowels. Cook et al. (2016) have shown that fragmented L2 phonological storage results in unstable sound-form mappings; written and lexical memory provide only indirect access to phonetic targets and cannot reliably generate distinct F1–F2 vowel categories.

Layered dissociation is also a way to combine acoustic data with self-reported data. Overlapping formant patterns alone would incorrectly assume that learners are unaware of vowel contrasts, and exclusive reliance on metalinguistic reports would overestimate their productive proficiency. In line with previous research (Bradlow et al., 1997, 1999; Wong, 2013), perception and production develop asynchronously; In line with this, the present study has found that learners' declarative strategic knowledge is often not reflected in real-time phonetic execution.

In short, self-reported vowel strategies only partially aligned with Cantonese learners' actual speech. Length-based approaches achieved the highest congruence, articulatory awareness mixed congruence, and orthographic/lexical tactics had almost no congruence. Although learners had gained some explicit knowledge of English tense–lax contrasts, this declarative knowledge had not been converted into automatic, stable phonetic control. The above results support the main content of this paper: learning L2 vowels is closely related to acoustic differentiation, perception, physical production and conscious learner strategies.

Pedagogical Implications

The results of this study provide some implications for teaching English tense–lax vowel contrasts to Cantonese ESL learners. Therefore, pronunciation teaching should help students connect what they know about vowel contrasts with what they hear and produce in real life.

First, teachers should not present tense–lax vowels merely as long–short distinctions. Many learners reported using duration to distinguish pairs such as /i:/–/ɪ/ and /u:/–/ʊ/, and this strategy was partially reflected in their production. However, the acoustic results showed an overlap in F1 and F2 for the difficult contrast of /i:/–/ɪ/. Therefore, alone is not sufficient. Teachers should therefore show students how to produce different vowels with their tongue, lips and throat. For example, /ɪ/ should not be taught solely as a reduced form of /i:/, but rather as a relatively low and central vowel.

Second, instruction should include perceptual training and production practice. Some learners could explain the differences in vowels but still produced overlapping categories; therefore, their perceptual targets had not yet been stabilized. High-variability phonetic training can be used to expose learners to various speakers, accents, words and sentence contexts (Bradlow et al., 1997, 1999; Wong, 2013). Do not use only the minimal pairs from one teacher's model; instead, have learners hear tense–lax contrasts in various materials and then practice producing them.

Third, teachers can give visual and auditory cues to help students distinguish between the vowels more easily. To confirm that the different productions are indeed different, vowel charts, spectrograms, formant plots and other speech analysis tools can be used. For example, comparing 'seat' and 'sit' on a vowel chart can show whether the two vowels are different in F1

and F2. This feedback can help learners understand that changes in articulation alter sound production; learners may otherwise be unable to hear these differences in their own speech.

Fourth, instruction can improve the sense of articulation. Some students reported using mouth shape, tongue position, lip rounding or muscular tension to differentiate vowels. Awareness alone, however, does not guarantee accurate production. Therefore, teachers can offer contrastive instruction by focusing on jaw opening and tongue height for /æ/–/ɛ/, or lip rounding, tongue backness and centralization for /u:/–/ʊ/. Instead of telling students vaguely to "open their mouths more", a teacher can use a mirror, a diagram, guided repetition and feedback for more precise correction.

Fifth, teachers should reduce learners' dependence on orthography. Some students used spelling patterns to recognise vowel sounds, but English spelling is often irregular. Sound sorting, phonemic transcription and listening discrimination are all ways to help learners focus on the categories of sounds rather than written forms. This is particularly true for Cantonese ESL learners who have received relatively little English input from other sources.

Finally, pronunciation teaching should no longer use isolated-word exercises. Familiar anchor words are convenient for learners, but they also need to expand their vowel contrasts to learn new words and communicate. For example, after practising ship and sheep, they can apply the same contrast to bit/beat and live/leave and then use these words in sentences or short communicative tasks. Provide repeated practice, delayed review and personalised feedback, because vowel category formation is gradual.

Limitations

This study used a small production sample (N=7) and only conducted descriptive acoustic analysis; thus, the results are exploratory rather than generalisable. Acoustic analysis was limited to F1 and F2, omitting vowel duration, F3 and dynamic formant trajectories, ignoring potential temporal differences between vowels such as /i:/ and /ɪ/, using existing published native speaker reference data instead of a matched concurrent control group, and failing to perform formant normalisation, which limits direct comparison between learners and native speakers. Interview data relied on conscious self-reports rather than real-time perceptual processing measures, lacked independent coding validation or inter-coder reliability checks, production and interview participant data were not linked, thus precluding individual-level analysis of the relationship between pronunciation patterns, perceptual strategies and metacognitive awareness, and the interview did not include targeted /ʌ/–/ɑ:/ minimal-pair stimuli, making conclusions about the low spontaneous awareness of this contrast inconclusive. Standardised English proficiency was not used as a selection criterion and was not measured; because proficiency may affect the accuracy of pronunciation, this omission is acknowledged as a limitation. Future research will use larger samples, add duration, F3 and dynamic formant measures, include a matched native control group with formant normalisation, add formal perceptual tasks and inter-coder reliability verification, and integrate production, perception and interview data for individual participants to clarify vowel differentiation mechanisms and developmental patterns.

Conclusion

This paper studies how Cantonese L1 learners produce and perceive three English vowel contrasts: /i:/–/ɪ/, /æ/–/ɛ/, and /ʌ/–/ɑ:/. The study combines analysis of acoustic production and interview-based strategy profiling to explore how learners' vowel production patterns relate to their own processing strategies.

Acoustic analysis showed significant spectral crowding in all three contrasts, and the overlap was most severe for /i:/–/ɪ/. The results also point to a possible pattern of vertical compression and indicate that L1 influence may shape both individual vowel categories and the general organisation of the developing L2 vowel space. According to the interview data, the learners used several strategies, such as contextual guessing, L1-based comparison, duration cues, orthographic knowledge, and limited articulatory self-monitoring.

In short, the problems learners had with English tense–lax vowels were not due to a lack of awareness. Learners had some explicit knowledge of the contrasts, but this knowledge was not always in line with acoustic differentiation. Therefore, it can be assumed that L2 vowel learning involves the interplay of perception, production, L1 assimilation and conscious strategy use. Based on the above results, pronunciation teaching for Cantonese L1 learners should move beyond the traditional minimal-pair approach. High-variability input, acoustic or visual feedback, and explicit articulatory guidance can help learners increase inter-category spacing and reduce the demand for contextual or spelling-based strategies.

Data Availability

The de-identified, token-level F1 and F2 measurements underlying the acoustic analyses and Figures 2 and 3 are provided as a Supplementary file (Vowel_F1F2_tokens.csv). Further data is available from the authors on reasonable request.

-
- Acknowledgements:** The authors would like to thank the faculty members and student participants from the undergraduate program in Guangdong Province for their time and participation, which significantly contributed to the study. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.
- Funding Statement:** No Funding
- Conflict of Interest Statement:** The authors declare that there is no conflict of interest regarding the publication of this paper.
- Ethics Statement:** This minimal-risk linguistic study collected interview data and word pronunciation recordings from volunteer participants. Formal full ethical review was not required under the institutional guidelines for low-risk educational/linguistic research. Informed consent was obtained before audio recording. All participants were assured of voluntary participation, anonymity and data confidentiality. Recordings were de-identified, securely stored, and used exclusively for academic research. The research complies with accepted academic integrity and ethical standards for human subject's research.
- Author Contribution Statement:** All authors contributed significantly to the development of this manuscript. Ye Yizhou was responsible for the conceptualization, methodology, and overall supervision of the study, and handled data collection, analysis, and interpretation of results. Latisha Asmaak Shafie and Razlina Razali contributed to the literature review, drafting, and critical revision of the manuscript. All authors read and approved the final version of the manuscript prior to submission.
-

References

- Baker, W., & Trofimovich, P. (2005). Interaction of native- and second-language vowel system(s) in early and late bilinguals. *Language and Speech*, 48(1), 1-27. <https://doi.org/10.1177/00238309050480010101>
- Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology*. Mouton de Gruyter.
- Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289-304). Kluwer.
- Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167-224). MIT Press.
- Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). John Benjamins. <https://doi.org/10.1075/llt.17.07bes>
- Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* (Version 6.3.10) [Computer software]. <http://www.praat.org/>
- Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 279-304). York Press.
- 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. <https://doi.org/10.3758/BF03206911>
- Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101(4), 2299-2310. <https://doi.org/10.1121/1.418276>
- Cao, G. W. (2024). Phonetic dissimilarity and L2 category formation in L2 accommodation. *Language and Speech*, 67(2), 301-345. <https://doi.org/10.1177/00238309231182967>
- Casillas, J. V. (2020). Phonetic category formation is perceptually driven during the early stages of adult L2 development. *Language and Speech*, 63(3), 550-581. <https://doi.org/10.1177/0023830919866225>
- Chan, A. Y. W. (2012). Cantonese English as a second language learners' perceived relations between "similar" L1 and L2 speech sounds: A test of the speech learning model. *Modern Language Journal*, 96(1), 1-19. <https://doi.org/10.1111/j.1540-4781.2012.01291.x>
- Chan, A. Y. W. (2013). The discrimination of English vowels by Cantonese ESL learners in Hong Kong: A test of the Perceptual Assimilation Model. *Open Journal of Modern Linguistics*, 3(3), 182-189. <https://doi.org/10.4236/ojml.2013.33025>
- Chan, A. Y. W., & Li, D. C. S. (2000). English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' English pronunciation problems. *Language, Culture and Curriculum*, 13(1), 67-85. <https://doi.org/10.1080/07908310008666590>

- Chang, Y.-S. (2023). Effects of production training with ultrasound biofeedback on production and perception of second-language English tense–lax vowel contrasts. *Journal of Speech, Language, and Hearing Research*, 66(5), 1479–1495. https://doi.org/10.1044/2023_JSLHR-22-00587
- Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, 7, Article 1345. <https://doi.org/10.3389/fpsyg.2016.01345>
- Deterding, D. (1997). The formants of monophthong vowels in Standard Southern British English pronunciation. *Journal of the International Phonetic Association*, 27(1-2), 47–55. <https://doi.org/10.1017/S0025100300005417>
- Escudero, P. (2005). Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization [Doctoral dissertation, Utrecht University]. LOT.
- Escudero, P. (2007). Second-language phonology: The role of perception. In M. C. Pennington (Ed.), *Phonology in context* (pp. 109–134). Palgrave Macmillan. https://doi.org/10.1057/9780230625396_5
- Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, 26(4), 551–585. <https://doi.org/10.1017/S0272263104040021>
- Feng, H., & Wang, L. (2024). Acoustic analysis of English tense and lax vowels: Comparing the production between Mandarin Chinese learners and native English speakers. *The Journal of the Acoustical Society of America*, 155(5), 3071–3089. <https://doi.org/10.1121/10.0025931>
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.
- Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press. <https://doi.org/10.1017/9781108886901.002>
- Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25(4), 437–470. <https://doi.org/10.1006/jpho.1997.0052>
- Georgiou, G. P. (2021). Effects of phonetic training on the discrimination of second language sounds by learners with naturalistic access to the second language. *Journal of Psycholinguistic Research*, 50(3), 707–721. <https://doi.org/10.1007/s10936-021-09774-3>
- Georgiou, G. P., & Savva, E. (2026). Exploring acoustic overlap in second language vowel productions. *International Journal of Applied Linguistics*, 36(2), 1293–1304. <https://doi.org/10.1111/ijal.12859>
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5), 3099–3111. <https://doi.org/10.1121/1.411872>
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <https://doi.org/10.1177/1049732305276687>

- Ji, J., & Jiang, A. (2022). The influence of dialect on the perception and production of lax-tense vowel distinction in English learning. *International Journal of Language and Linguistics*, 10(2), 103-110. <https://doi.org/10.11648/j.ijll.20221002.16>
- Law, W. L., Dmitrieva, O., & Francis, A. L. (2021). Language attitudes modulate phonetic interactions between languages in bilingual speakers in diglossic settings. *Linguistic Approaches to Bilingualism*, 11(3), 289-322. <https://doi.org/10.1075/lab.18085.law>
- Lee, W.-S. (2016). Articulatory-acoustical relationship in Cantonese vowels. *Language and Linguistics*, 17(4), 477-500. <https://doi.org/10.1177/1606822X16637058>
- Llompарт, M. (2021). Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon. *Bilingualism: Language and Cognition*, 24(3), 481-496. <https://doi.org/10.1017/S1366728920000656>
- Mok, P. P. K., Fung, H. S. H., Cao, G. W. L., & Leung, C. W. (2023). How similar are the formants in the speech of bilingual speakers? *International Journal of Speech, Language and the Law*, 30(1), 52-86. <https://doi.org/10.1558/ijsl.22946>
- Morrison, G. S. (2008). L1-Spanish speakers' acquisition of the English /i/-/ɪ/ contrast: Duration-based perception is not the initial developmental stage. *Language and Speech*, 51(4), 285-315. <https://doi.org/10.1177/0023830908099067>
- Nagle, C. L. (2018). Examining the temporal structure of the perception-production link in second language acquisition: A longitudinal study. *Language Learning*, 68(1), 234-270. <https://doi.org/10.1111/lang.12275>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259. <https://doi.org/10.1037/0033-295X.84.3.231>
- Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, 69(3), 652-708. <https://doi.org/10.1111/lang.12345>
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158. <https://doi.org/10.1093/applin/11.2.129>
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge University Press.
- Simonchyk, A., & Darcy, I. (2023). Development of production skills in the absence of precise phonolexical representations. *Second Language Research*, 39(3), 623-648. <https://doi.org/10.1177/02676583211066291>
- Song, J., & Eckman, F. (2024). The relationship between L2 learners' production and perception of English vowels: The role of native-speaker acoustic patterns in production. *Second Language Research*, 41(2), 309-338. <https://doi.org/10.1177/02676583241240868>
- Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, 36(3), 326-344. <https://doi.org/10.1093/applin/amu076>
- Wong, J. W. S. (2012). Training the perception and production of English /e/ and /æ/ of Cantonese ESL learners: A comparison of low vs. high variability phonetic training. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST 2012)* (pp. 37-40).

- Wong, J. W. S. (2013). The effects of perceptual and/or productive training on the perception and production of English vowels /ɪ/ and /i:/ by Cantonese ESL learners. In INTERSPEECH 2013 (pp. 2113-2117). <https://doi.org/10.21437/Interspeech.2013-501>
- Wong, J. W. S. (2014). The effects of high and low variability phonetic training on the perception and production of English vowels /e/-/æ/ by Cantonese ESL learners with high and low L2 proficiency levels. In INTERSPEECH 2014 (pp. 524-528). <https://doi.org/10.21437/Interspeech.2014-129>
- Zee, E. (1999). Chinese (Hong Kong Cantonese). In Handbook of the International Phonetic Association (pp. 58-60). Cambridge University Press.
- Zee, E. (2003). Frequency analysis of the vowels in Cantonese from 50 male and 50 female speakers. In M. J. Solé, D. Recasens, & J. Romero (Eds.), Proceedings of the 15th International Congress of Phonetic Sciences (pp. 1117-1120). Causal Productions.

Appendix A: Target Word List for the Acoustic Production Task

Vowel contrast	Target vowel	Target words
/ʌ/-/ɑ:/	/ʌ/	coming, us, but, some
/ʌ/-/ɑ:/	/ɑ:/	aren't, barked, hard, party
/i:/-/ɪ/	/i:/	beat, street, please
/i:/-/ɪ/	/ɪ/	fill, milk, windy, strict, velvet
/æ/-/ɛ/	/æ/	bag, tag, cab, cap
/æ/-/ɛ/	/ɛ/	weather, velvet, vest, west

Appendix B: Interview protocol

The interviews were conducted in Mandarin Chinese. The following guiding questions were used:

- Which English vowel contrasts do you find most difficult to hear or pronounce?
- How do you usually distinguish similar words such as street and strict?
- How do you usually distinguish words such as bag and beg?
- Do you rely on spelling when deciding how to pronounce English vowels?
- Do you rely on word meaning or sentence context when you cannot hear the vowel clearly?
- Do you compare English vowels with Cantonese or Mandarin sounds?
- Do you pay attention to mouth opening, tongue position, lip shape, or vowel length?
- Have you received any formal pronunciation or phonetic training?
- Do you correct yourself when you notice a vowel pronunciation problem?
- Are there any English vowels that you often guess rather than clearly distinguish?