Laboratory Phonology
Journal of the Association for Laboratory Phonology

Lee-Kim, S.-I., & Tung, H.-Y. (2025). Linguistic experience and social factors in speech perception: The case of merged speakers of Mandarin sibilants. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 16(1), pp. 1–31.
DOI: https://doi.org/10.16995/labphon.15384

Open Library of Humanities

Linguistic experience and social factors in speech perception: The case of merged speakers of Mandarin sibilants

Sang-Im Lee-Kim*, HIPCS Hanyang University, Korea, sangimleekim@hanyang.ac.kr

Hsiang-Yu Tung, National Yang Ming Chiao Tung University, Taiwan, loveh89301@gmail.com

*Corresponding author.

This work explores the combined effects of social expectations and a speaker's production characteristics on the perception of alveolar versus retroflex sibilants that are variably merged in Taiwan Mandarin. The variation is socially structured in that the sibilant merger is regarded as a characteristic feature of speakers from southern Taiwan. The results of an AXB discrimination task showed that although merged speakers were outperformed by their distinct counterparts, they were able to discriminate the sibilants far beyond chance level. In an identification task with social guises, participants showed a pattern reflecting the implicit bias that a southern-labeled talker is less likely to produce retroflexes, and hence use the merged form, than a northern-labeled talker. Interestingly, merged participants were again shown to be less sensitive to frication noise cues, but they more readily switched between the social and acoustic cues than distinct participants. Together, these results indicate that frequent encounters with distinct forms in a speech community with large interspeaker variation might help merged speakers remain sensitive to phonological distinctions that they do not carry. Merged speakers might have been desensitized to the acoustic cues to some degree; however, they appear to use other cues to achieve coherent speech perception whenever possible.

1. Introduction

1.1. Perception-production misalignment in the context of phonemic mergers

In many changes-in-progress, categories are merged in production before they are merged in perception. A wealth of sociophonetic literature corroborates empirical evidence from English vowels. Hay et al. (2006b), for example, investigated the perception and production of the NEAR (/iə/)-SQUARE (/eə/) vowel merger in New Zealand English (NZE) whereby the SQUARE vowel approximates the phonetic space of the NEAR vowel. They showed that speakers with higher Pillai scores, a measure of category overlap ranging from 0 (complete overlap) to 1 (clear separation), were more accurate in vowel identification than those whose Pillai scores were lower. Despite the poorer performance of the "merged speakers", their misidentification rates were around 15% and 30% for NEAR and SQUARE, respectively, far beyond the chance level in a two-alternative choice (AFC) task, apparently indicating that the speakers did not entirely lose their perceptual sensitivity to the vowel distinction. Similarly, moderate performance in perceptual tasks by merged speakers is widely documented for other vowel mergers: e.g., the /el/ and /æl/ merger in NZE (Hay et al., 2013; Thomas & Hay, 2005), the PIN(/ɪ/)-PEN(/ɛ/) merger in American English (Austen, 2020), and multiple mergers /ʊl/-/ol/ and /ul/-/ʊl/ in some dialects of English (Wade, 2017). In all those cases, merged speakers are consistently outperformed by their distinct counterparts, reflecting some link between perception and production, but they remain sensitive to the phonological contrasts despite the lack of vowel distinction in their own speech.

The mismatch between perception and production is well attested for the domain of tone mergers as well. Cantonese has six lexical tones (three level tones and three contour tones), and multiple tone mergers (T2 (25) vs. T5 (23), T3 (33) vs. T6 (22), and T4 (21) vs. T6 (22)),[1] are reported to be in progress in Hong Kong Cantonese (Bauer et al., 2003; Fung & Lee, 2019b; Kei et al., 2002; Lin et al., 2021; Mok et al., 2013; Yiu, 2009; Zhang, 2019). The two rising tones, T2 (25, high-rising) and T5 (23, low-rising), among others, were shown to be particularly susceptible to merger by approximation, namely T5 produced as T2. However, merged speakers, as well as distinct speakers, performed at ceiling in an AX discrimination task for those tone pairs produced distinctively by a professional phonetician, although they were slower at discriminating the tones than distinct speakers. In another study, Lin et al. (2021) tested young Hong Kong Cantonese speakers for tone perception and shadowing and found that the merged speakers performed as well as the control distinct group (above 90% accuracy) in an AX tone discrimination task. As Mok et al. (2013) pointed out, simple same-different discrimination tasks might have participants tap into acoustic level processing, rather than a lexico-phonological level, facilitating ceiling performance. Nonetheless, it seems to hold that a merger in tone production does not necessarily

[1] The numbers in parentheses represent the relative pitch range of a speaker, with five being the highest and one being the lowest (Chao, 1930, 1968).

entail a merger in perception. The misalignment between the two domains of speech is a genuine linguistic pattern attested for both segmental and suprasegmental contrasts in the context of mergers-in-progress.

Previously, a merger in production with distinction in perception was argued to be rare (Labov, 2011),[2] but critics have suggested that the traditional methodologies used to assess phonemic merger were too conservative and underestimated merged speakers' actual perceptual abilities (Austen, 2020; Wade, 2017). Sociolinguistic research often uses so-called *commutation tests* wherein participants are given words from minimal pairs as auditory stimuli and asked to identify the word from multiple choices. Crucially, the cut-off point of this test is set near 100% accuracy; failure in this task is interpreted as the speaker being merged in perception. For example, Austen carried out a large-scale online phonetic investigation of the PIN-PEN vowel merger across the US. In an AFC identification experiment, participants were given stimulus items produced by a distinct-speaking woman with a standard accent. A participant was determined to be merged in perception if they had more than two items misidentified out of sixteen PIN-PEN tokens. In this way, many participants (51 out of 371) were classified as merged in perception. However, the author noted that most of them performed much better than chance, indicating that they retained some sensitivity to the vowel contrasts. Speakers in the cases reviewed earlier showing moderate perceptual accuracy would all have fallen into the perceptually merged group if the same criteria were applied. Therefore, the traditional methodology used to assess merger status is likely to underestimate merged speakers' knowledge of phonological contrasts.

Another issue with conventional methodologies lies in the elicitation of a speaker's explicit knowledge of phonological contrasts. Speakers are often asked to provide their self-judgments about whether the vowels in minimal pairs should be pronounced the same or differently (Di Paolo & Faber, 1990; Labov et al,, 2006; Labov et al., 1991). This could be potentially problematic because speakers might resort to differences in spelling, rather than their abstract knowledge of phonemic representations (Austen, 2020). Furthermore, vowel distinctions are often made through multiple acoustic cues, in which case tasks relying on speaker awareness are not appropriate for drawing conclusions about what they know subconsciously (Di Paolo & Faber, 1990; Wade, 2017). Specifically, Wade examined multiple pre-l vowel mergers, POLE-PULL and POOL-PULL, among speakers in Youngstown, Ohio, US. In addition to the differences in vowel quality, the two sets of vowels contrast in duration: the first vowel in each set manifests

[2] Notably, the stimuli for perception tests can come from two sources: one from merged speakers themselves and the other from distinct speakers. The former often leads to well-known cases of "near merger" (Labov, Yaeger, & Steiner, 1972); because the categories produced by merged speakers are approximated acoustically, the speakers themselves cannot differentiate them reliably in perception. However, the latter involves cases in which merged speakers are presented with speech materials produced by distinct speakers they might encounter frequently in their speech community. In this case, merged speakers tend to perform at above-chance levels. We are interested in this particular case because it reveals merged speakers' actual perceptual abilities.

an acoustically longer duration than the short /ʊ/ vowel. In an identification experiment, participants were asked to choose the vowels in a forced-choice format for stimuli whose rhyme durations were systematically manipulated along a continuum. The results showed that the speakers who were spectrally merged (lacking differences in vowel quality) could still utilize durational cues for vowel distinction at a level comparable to speakers from Burlington, a control group that maintained multiple vowel contrasts in production. This suggests the need for more sophisticated methods to elicit linguistic knowledge from merged speakers, particularly knowledge that is largely subconscious.

To this end, we conducted two perception tests (discrimination and identification) to investigate merged participants' perceptions of sounds in contrast. Specifically, we examined the perception of alveolar-retroflex sibilants in Taiwan Mandarin (TM) (e.g., /**s**an/ 三 'three' ~ /ʂan/ 山 'mountain') in connection with the speaker's production patterns. The sibilants in TM are often merged through deretroflexion, defined as the replacement of retroflexes with corresponding alveolars (e.g., /ʂan/ produced as [**s**an]) (Kubler, 1985; Lin, 1988; Wei, 1984). To date, the production of merged sibilants in TM has been studied extensively using both acoustic and articulatory methods (Chang, 2012; Chiu et al., 2019; Chuang & Fon, 2010; Chuang et al., 2019; Lee-Kim & Chou, 2022, 2024; Shih, 2012). However, the perceptual aspects of this pattern have rarely been explored, let alone how a participant's production pattern influences their perception of the sibilants. This pattern is likely to offer valuable insights into how the two aspects of speech sounds inform each other and are, to some extent, disconnected in the context of large interspeaker variations. Furthermore, our data from an understudied sound category are expected to expand the empirical coverage of sociophonetic research, which has largely focused on vowel mergers in English.

1.2. Linguistic experience and social factors in speech perception

Listeners incorporate social information about the talker, as well as the acoustic properties of the sounds, into their speech perception. A variety of factors is known to bias listeners toward hearing certain phonetic variants associated with a particular social group: e.g., region/dialect (D'Onofrio, 2015; Hay & Drager, 2010; Hay et al., 2006a; Niedzielski, 1999; Schertz et al., 2019), race (Staum Casasanto, 2010), and age (Hay et al., 2006b; Koops et al., 2008). Niedzielski was among the first in this line of research. That study tested the perceptions of Detroit listeners regarding the vowel /aʊ/, which is often produced as a more centralized variant [ɐʊ] through Canadian raising. In a matched-guise task, half the participants were told that the talker was from Canada, and the other half were told that the talker was from Detroit. Even though both groups were exposed to a set of acoustically identical vowels, listeners assigned to a Canadian-labeled talker heard centralized diphthongs significantly more than those assigned to a Detroit-labeled talker. This result clearly demonstrates that listener expectations about a talker's social background modulate the way they process the incoming speech signal.

Capitalizing on the dynamic sociolinguistic variation in TM sibilants, we examined whether the model talker's regional affiliation would have a systematic influence on sibilant perception. Crucially, the TM sibilant merger is characterized as a phonetic feature representative of southern speech (TM speakers from the southern part of Taiwan) and associated with other social factors such as age, gender, education, and socioeconomic status (Chung, 2006; Kubler, 1985; Su, 2008; Tse, 1998). Our identification experiment manipulated socially motivated guises by describing the model talker as being from either the south or the north. A straightforward prediction is that, given identical acoustic inputs, listeners would bias toward hearing more retroflexes when they were guided to believe they heard the voice of a northern speaker, who would be more likely than one from the south to carry the sibilant distinction. In contrast, they would hear non-retroflexed sibilants, i.e., alveolars, more often when they believed the talker was from the south. Results in line with that expectation would show that the perceived social identity of the talker plays an independent role in speech categorization in the absence of acoustic differences.

Here, we distinguish between the effects arising from social information and those from perceptual compensation. Previous studies, such as those by Strand (1999) and Strand and Johnson (1996), have demonstrated that gender differences influence the perception of English sibilants /s/ versus /ʃ/. When faced with ambiguous noise signals, listeners are more likely to give /s/ responses when shown a photo of a male speaker than a female. This phenomenon is a case of perceptual compensation, where listeners adjust their perception based on the production characteristics they associate with the speaker. That is, listeners assume that frication noises produced by females are generally higher in frequency than those by males, leading them to seldom give /s/ (high frequency noise) responses for a female voice unless the noise spectrum is so high in frequency that it cannot be interpreted as /ʃ/ (low frequency noise). This parallels the so-called 'sushi' bias, where listeners hear more /s/ before /u/ and more /ʃ/ before /i/; because they expect the frication noise to be lower in frequency in the context of the rounded vowel /u/, listeners adjust their perception to hear the spectrally higher noise /s/ more often before the lower vowel /u/, and vice versa (Kang, Johnson, & Finley, 2016; Mitterer, 2006; Smits, 2001; Whalen, 1989).

This low-level cognitive processing, based on perceptual compensation, is fundamentally different from social priming effects, which arise from the episodic accumulation of exemplars throughout a speaker's lifetime. Essentially, the associations between linguistic and social features, such as southern speakers with deretroflexion and northern speakers with clear retroflexes, are arbitrary. When exposed to southern social cues, the corresponding phonetic traits—deretroflexed sibilants—become activated, biasing listeners toward hearing more alveolars. Therefore, the processing based on episodic traces predicts a pattern opposite to that of perceptual compensation. Specifically, TM listeners are expected to hear more retroflexes when the talker is described as coming from the north than the other way around, demonstrating a clear mapping between the talker's dialectal origin and the phonetic features of their speech.

The TM sibilant merger adds another interesting aspect to consider in our search for the factors affecting speech perception: the participants themselves vary in their production characteristics, merged versus distinct, for the sibilant contrasts. The initial evidence shows that a speaker's production pattern modulates the extent to which socio-indexical cues are used in speech perception (Hay et al., 2006b; Schertz et al., 2019). In NZE, the NEAR-SQUARE vowel merger is attested more frequently for speakers at the intersection of the lower-class and young generation. Hay et al.'s (2006a) study was formatted as a social guise experiment in which distinct vowels were presented along with different photos stratified by age and social class. The results revealed that participant's perceptual accuracy was higher when the sound and the photo matched: a photo of a talker who looked old and high-status, i.e., congruent with distinct vowels, yielded more accurate perceptions. However, the results were more nuanced in that that finding was limited to speakers who merged the vowels in their production, whereas those who conveyed distinct vowels in their own production did not reliably use social cues in their perceptions. Those authors speculated that this might reflect asymmetries in the speakers' linguistic experience; merged speakers are regularly exposed to (high-status) distinct speech through the media, for example, whereas the reverse might not be true.

Inspired by previous studies, this work addresses both listener and talker characteristics in speech perception by involving socially structured interspeaker variation and excluding potential asymmetry in the amount of linguistic experience. Specifically, the identification experiment in our study adopted a between-subject design: one group was exposed to a southern-labeled talker and the other to a northern-labeled talker, and the participants were cross-balanced for social valence, including gender and region.

1.3. The sibilant merger and its social meaning in Taiwan

TM is the official and dominant language of Taiwan. As a variety of Mandarin, TM contrasts two coronal sibilants for place: alveolars (/s ts ts^h/) versus retroflexes (/ʂ tʂ $tʂ^h$/), like Standard Mandarin spoken in mainland China (Lin, 2007). However, TM is well known for the variable merging of the two sets of sibilants through deretroflexion, which is conditioned by a variety of social factors: age, gender, region, education, and socioeconomic status (Chung, 2006; Su, 2008). It is generally assumed that old men from the underprivileged social class in the south are more likely than others to merge the two categories. The association between the sibilant merger and the somewhat negative social connotation is deeply rooted in the sociolinguistic history of the country.

Historically, two major waves of migration occurred from the mainland to Taiwan, the first of which brought people from the Southeast coastal areas of China during the 17th and 18th centuries. The native language of those people was Southern Min, a Chinese language that differs considerably from Mandarin in its sound system and is mutually unintelligible with it. Most relevant to the current study, Min varieties lack retroflexes in the system altogether, and they

have only one set of coronal sibilants, the alveolars. Min-speaking Chinese people soon became the representative ethnic group of the island, and their language, Taiwanese Southern Min (TSM), emerged as the dominant language at that time. Demographic reports note that around 70% of the contemporary Taiwanese population speaks TSM, as well as TM, and those who do not speak TSM still have at least some passive knowledge of it (Feifel, 1994; Huang, 1993a; Sandel, 2003). The second wave of immigration took place in 1949 when the Nationalist party retreated to Taiwan after being defeated in the Chinese Civil War. These migrants brought Northern Chinese, TM, a Mandarin variety in which retroflexes contrast with alveolar counterparts. The Nationalist party forcefully imposed its language onto Min-speaking people, and TSM was banned from being spoken in all public sectors until martial law was finally lifted in 1987 (Feifel, 1994; Huang, 1993a; Tse, 2000). Due to the initial settlement of the government and its subsequent residency in the north of Taiwan, the language landscape in Taiwan presents a unique vertical division between the north, occupied primarily by TM-speaking people, and the south, which is occupied by the TSM-speaking population (see a language map here: https://rb.gy/73g3bx).

For several decades, TM was promoted as a prestigious language while TSM was demoted as a stigmatized one, and the social connotations of language and prestige remain deeply solidified within the speech community. Many elderly people whose native language is TSM speak little or heavily accented TM, which is referred to as Taiwan Guoyu, a language that is ranked low in its perceived linguistic hierarchy (Khoo, 2019). The situation began to change in the late 1980s when the ban on TSM in public sectors was officially removed. A movement to promote TSM was launched around that time, and its status has since risen; the language is now used in both public and private sectors and is argued to be a symbol of regional identity (Cheng, 1978, 1989; Huang, 1993b).

The complex sociolinguistic situation in Taiwan has led to mixed connotations of the sibilant merger. On the one hand, deretroflexion has been attributed to intense language contact between TM and TSM; due to the influence of TSM, a speaker might not be able to produce “proper” retroflexes. However, recent instrumental studies have reported a gradual decoupling between TSM fluency and the sibilant merger (Chuang et al., 2019; Lee-Kim & Chou, 2022). Speakers with high fluency in TSM do not necessarily merge the sibilants more than those who speak little TSM. The merged phonetic variant is widespread in the Taiwanese speech community and is, by and large, independent of its original source based on language contact. On the other hand, despite the declining stigma associated with deretroflexion, those instrumental studies have confirmed the role of gender: women are more likely to retain the distinction than men. This indicates that the perception of TM as the “standard” language is still profoundly rooted among Taiwanese people. The socially desired phonetic norms appear to persist through the avoidance of deretroflexion in speech by women, who might care about social refinement more than men (Lee-Kim & Chou, 2024; Su, 2008). Social awareness of phonetic variation in sibilant production

is fairly high; when speakers were asked about salient features of TM that distinguish it from mainland Mandarin, "retroflexion" was named most frequently by both Taiwanese and mainland speakers (Chang 2017). As such, the sibilant merger is often subject to overt comments by TM speakers (Chung, 2006).

The dynamic language contact situation on the island, coupled with socio-political attitudes, provides a unique opportunity to empirically test the broad question of how social information about a talker is integrated with the production patterns of listeners to enable coherent speech perception. Here, we report the results of two perception experiments. The first experiment was an AXB discrimination task designed to compare the perceptual sensitivity of two speaker groups, merged versus distinct. A straightforward prediction about this group comparison is that the distinct group will outperform the merged one, assuming a reasonable link between perception and production. Such a finding would corroborate the consistent findings of previous sociolinguistic research into different types of mergers (Wade, 2017; Hay et al., 2006b; Thomas & Hay, 2005). Capitalizing on the association between the region and the expected likelihood of the sibilant merger, the second experiment examined whether a model talker's social background would have a systematic influence on sibilant identification. This experiment was, therefore, designed to include socially motivated guises: the model talker was described as being from either a northern city where TSM is not commonly spoken or a southern city where TSM is widely used. The two experimental groups were comparable in their linguistic experience with TSM, and we tested whether their use of the social cues was modulated solely by their own production characteristics. The results are expected to shed light on how phonetics, social cues, and a listener's merging status jointly influence speech perception.

2. Experiment 1: Sibilant discrimination

The sibilant discrimination experiment was conducted approximately half a year after the sibilant identification task. We present the results of the discrimination task first to establish the perceptual performances of merged participants in comparison with distinct participants. We then present the results of the identification task in the next section (Section 3) to show the interaction between social and acoustic cues in speech categorization.

2.1. Participants

The entire pool of participants for the two perception experiments consisted of seventy-five TM-speaking college/graduate students (42 women, 33 men; aged 20–29, $M = 22$) at National Yang Ming Chiao Tung University and National Tsinghua University located in the north of Taiwan. Two groups were identified with respect to their merger status: thirty-three merged and forty-two distinct speakers (**Table 1**). The specifics of the speaker groups are discussed later in this section. Participant TSM fluency was balanced; half of the speakers were fluent in TSM,

and the other half were TM-dominant.[3] Among those speakers, thirty-five people were invited to participate in the discrimination task: sixteen participants from the merged group (4 women, 12 men) and nineteen from the distinct group (16 women, 3 men). All but two distinct-speaking women had participated in the between-subject identification task with social guises, which required a large number of participants.

	F	M	Total
merged	13	20	33
distinct	29	13	42
total	42	33	75

Table 1: Numbers of participants by merger status and gender in the two perception experiments.

Prior to the discrimination task, the participants completed a wordlist reading task designed to establish their merger status. The wordlist consisted of four disyllabic words with word-initial sibilants (alveolars /s ts^h/ vs. retroflexes /ʂ $tʂ^h$/) followed by the vowel /a/ (e.g., /**s**a^{55}man^{214}/ 撒滿 'fully sprinkled' vs. /ʂa^{55} fa^{55}/ 沙發 'couch'). Filler items were also included. The randomized wordlist was presented in Chinese characters and repeated five times, and the speech signals were recorded at a sampling rate of 44.1 kHz. The frication noise was labeled manually in Praat (Boersma & Weenink, 2020) and a multitaper spectral analysis was applied to it in Matlab (Blacklock, 2004; Blacklock & Shadle, 2003) to obtain spectral mean (M1) and peak frequency values. The average frequency values of the retroflexes were then subtracted from those of the alveolars for each individual speaker to obtain the spectral mean and peak distance. This perception study is a part of a larger research project, and more details about the acoustic analysis of the baseline production task are found in Lee-Kim and Chou (2022, 2024).

Figure 1 illustrates the distributions of the by-participant spectral mean and peak distance, which are highly correlated ($r = .950$, $p < .0001$). Merged participants were defined as those whose spectral mean and peak distance were lower than 1500 Hz and 2000 Hz, respectively. These cutoff values were slightly higher than those used in previous studies (Lee-Kim & Chou, 2022, 2024). These values were chosen because the distributions of the two categories were not significantly distinct for most participants who fell in that range when two-sample *t*-tests were performed for each individual's data. Furthermore, the number of sibilant tokens used for the statistical analysis was obviously very small (2 sibilants for each place repeated 5 times), which could easily yield distinct distributions. The slight loosening of the threshold in the selection of the merged group was, therefore, deemed to be reasonable.

[3] TSM fluency is on a continuum, and we recruited speakers who were at the two ends of the spectrum at the stage of prescreening. Speakers who indicated that their TSM fluency was higher than 5 or lower than 3 on a 7-point scale were invited to participate in the experiment. TSM-fluency generally reflects a person's sociolinguistic background; TSM-fluent speakers are likely to have experience with TSM-speaking family and neighbors living in the south.

Additionally, a linear regression model was fitted to the spectral distance data in R (R Development Core Team, 2023) to establish conditioning social factors of the sibilant merger. Those results revealed that gender had a significant effect but TSM fluency did not: men were more likely to merge the sibilants than women (p = .0055), but TSM fluency (p = .8486) and its interaction with gender were not significant predictors of the merger (p = .7234). This is consistent with previous findings (Chuang et al., 2019; Lee-Kim & Chou, 2022) that the sibilant merger is not necessarily triggered by TSM fluency—the pattern has become widespread throughout the speech community in Taiwan.

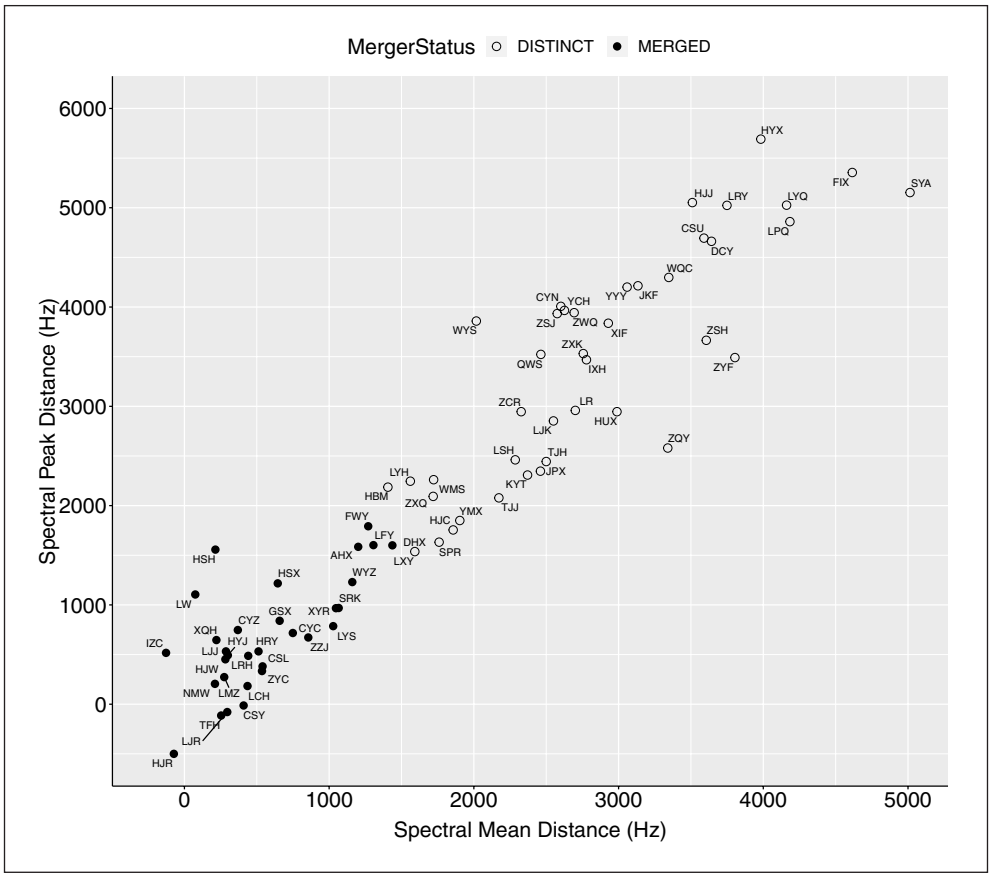


Figure 1: Distribution of spectral peak distance against spectral mean distance for each participant. Merged participants are represented by filled circles, and distinct participants by empty circles.

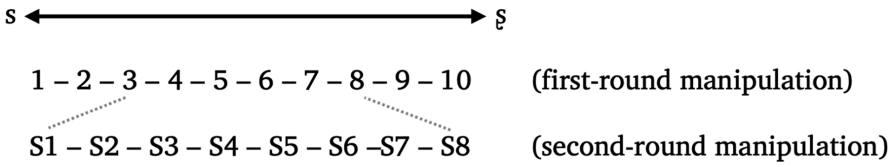
2.2. Stimuli

The acoustic stimuli for the perception experiments were created by digitally manipulating naturally produced Mandarin syllables. All stimulus items were recorded and created once, and subsets of the sounds were used for the discrimination and identification tasks.

The acoustic recordings of the stimuli were made by a TM-speaking man, aged 37, who conveyed a clear alveolar-retroflex distinction. He produced monosyllabic words containing initial alveolars /s ts tsʰ/ and retroflexes /ʂ tʂ tʂʰ/ followed by /a/ and /u/ carried by the high-level Tone 1 (X^{55}). All syllables were legitimate Mandarin CV words (e.g., /sa^{55}/ 撒 'spread' and /$ʂa^{55}$/ 殺 'kill'). In addition to the target sibilants, 42 monosyllables containing non-sibilant sounds were also recorded. Those include a variety of Mandarin syllables with different vowels (/i^{55}/, /y^{55}/) and differing tones (/tu^{35}/, /$tʰu^{214}$/, /i^{51}/). The stimulus tokens were produced in isolation and repeated three times. The recordings were made using a Marantz recorder connected to an AKG C520 head-mounted microphone at a sampling rate of 44.1 kHz. The recording took place in a sound-attenuated booth.

Among the three repetitions, one token of each condition was selected based on clarity. The selected tokens containing initial sibilants were digitally manipulated in Praat (Boersma & Weenink, 2020). Two portions of frication noise, /s/ and /ʂ/, were excised from the original speech signal, equated for duration,[4] and mixed with each other with different amplitude proportions to create 8-step continua (S1-S2-S3-S4-S5-S6-S7-S8) using Praat scripts. S1 and S8 are tokens with 100% amplitude from the alveolar and retroflex, respectively, and those between them result from a combination of the two sibilants; for example, S2 contains 86% amplitude from S1 and 14% amplitude from S8. The digital manipulation was conducted twice: 10-step continua were created first, and the frication noise from steps 3 and 8 were used as inputs for the second-round manipulation, as illustrated in (1).[5] The sibilants from steps 3 and 8 were chosen because they were considered to be good perceptual exemplars of the alveolars and retroflexes of TM, as judged by both authors. The stimulus talker had highly distinct speech, and the use of intermediate sibilants from this specific talker were used to ensure the sound of natural but still distinct Taiwanese sibilants. The same procedure was applied to the affricate sibilants. The vowels were later spliced from the original alveolar tokens.[6]

(1) Illustration of the manipulation of fricative sibilants used for the stimuli



[4] The durations of noise were similar between contrasting sibilants of the same manner. In the context of the /a/ vowel, for example, the durations of /s/ and /ʂ/ in the original tokens were 153 ms and 146 ms respectively, /ts/ and /tʂ/'s were 62 ms and 51 ms, and /tsʰ/ and /tʂʰ/'s were 145 ms and 110 ms.

[5] It is not necessary to mix the noise signals in two steps, but we could ensure the naturalness of the sounds more easily in this way. All sound files are available at OSF (https://osf.io/7b6fd/).

[6] The vowel formants following the two coronal sibilants do not differ significantly from each other (Lee-Kim, 2014, Ch3), and the use of the vowel following either sibilant would be valid for the purpose of this experiment.

Figure 2 presents the two sets of spectra for the /sa/-/ʂa/ and /su/-/ʂu/ continua. As shown in the figures, the spectral peaks at higher frequencies (8–9 kHz) were attenuated in amplitude, whereas those at lower frequencies (1–3 kHz) were amplified as the signals changed from most /s/-like to most /ʂ/-like. The acoustic difference between the sibilants from one end to the other was larger when preceding /u/ than when preceding /a/, which likely had perceptual consequences. This was neither intended nor expected; however, it was noted in the previous literature that alveolar-retroflex sibilant contrast can be enhanced in the rounded vowel context (Rau & Li, 1994). Note that this pattern is opposite of the reduced acoustic distance for /s/-/ʃ/ contrasts in English (Jongman et al., 2000; Soli, 1981; Yu, 2019). The results of the perception tasks with respect to the fine-grained acoustic properties of the sibilants will be discussed later.

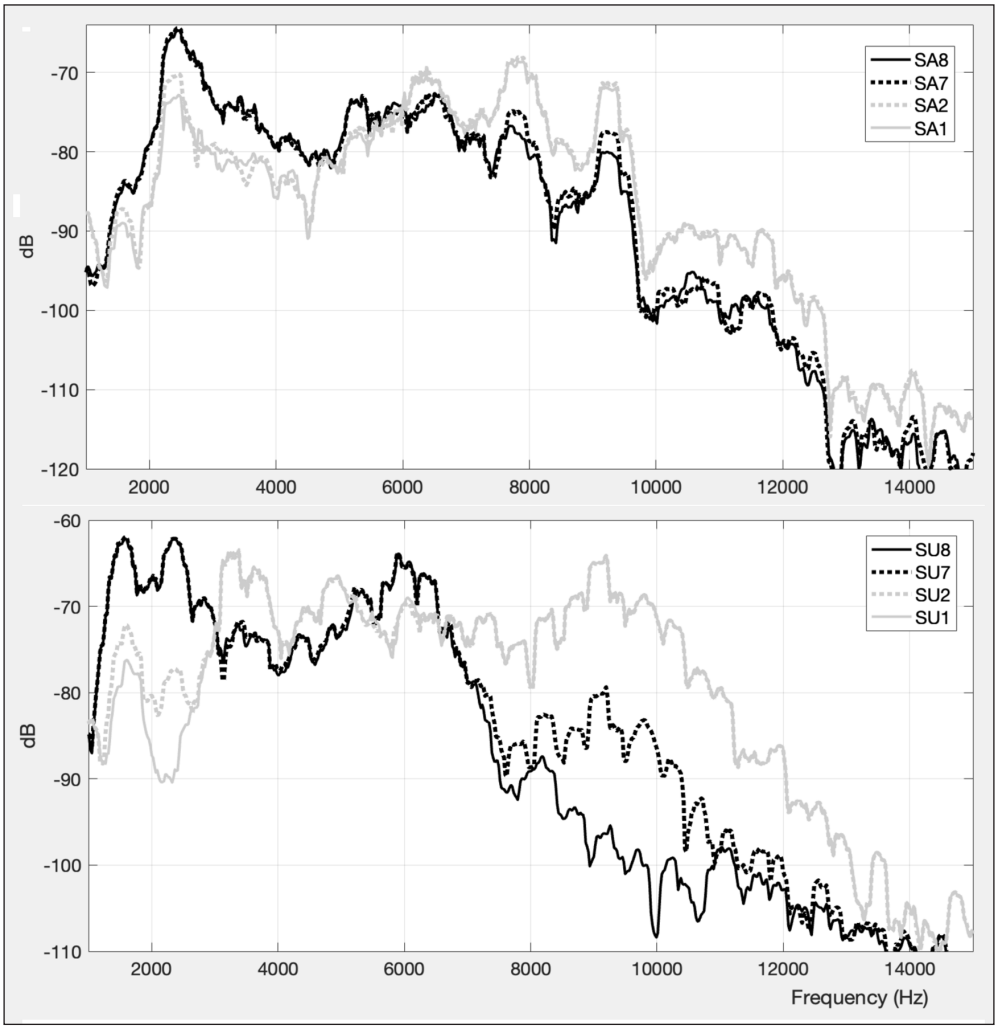


Figure 2: /s/-/ʂ/ spectral continua in the /a/ (top) and /u/ (bottom) vowel context drawn from the multitaper spectral analysis. SA1 and SU2 are the tokens with the most alveolar-like signals, and SA7 and SU8 are those with the most retroflex-like signals.

For the discrimination task, the S1 and S2 tokens from the 8-step sibilant continua were selected as exemplars of alveolars, and the S7 and S8 tokens were chosen as exemplars of retroflexes. The two tokens from each end were similar but not identical acoustically, enabling us to test listener knowledge of sibilant categories, rather than their sensitivity to the purely acoustic nature of the specific sounds used in the experiment.

2.3. Procedure

The experiment was conducted individually in a sound-attenuated booth using AKG-K240 MKII headphones connected to a laptop computer.

The discrimination task was designed in an AXB paradigm run in E-Prime 3.0 (Psychology Software Tools, 2016). For each trial, a sequence, S1(alveolar)-S2(alveolar)-S8(retroflex) for example, was played, with the interstimulus interval set at 500 ms. Participants were instructed to judge whether the middle stimulus was more like the first one or like the third one, so in the example just given, the correct answer would be the first one. They were asked to press "1" for the first sound and "2" for the last one using the number keys on the keyboard. For each AXB trial, four combinations were created by manipulating the presentation order: AAB, ABB, BBA, and BAA. In addition, the position of the two tokens of the same category, e.g., S1 and S2 or S7 and S8, was counterbalanced, resulting in 4 different possible pairs. As a result, 96 tokens (4 pairs $\times$ 4 orders $\times$ 3 manners $\times$ 2 vowels) composed one block, and they were repeated three times (96 $\times$ 3 blocks $=$ 288 trials) as three blocks. Within each block, the trial order was automatically randomized each time by E-Prime.

Before the start of the experiment, participants were given verbal instructions about the procedure, and written instructions in Chinese appeared on the computer screen. For each trial, a cross-fixation mark appeared in the center of the computer screen while the three sounds were played. Each subsequent trial began after the presentation of a blank screen for 1 second. Before the first test block, participants completed three practice trials to familiarize themselves with the experimental procedure. The participants were told that the task was timed, and they were encouraged to respond as quickly and accurately as possible. If the participant did not respond within 3 seconds, the experiment was designed to automatically move on to the next trial. The entire experiment took about 20 minutes to finish for most participants.

Discarding a small number of missing responses (73 tokens), we compiled 10,007 tokens (288 trials $\times$ 35 speakers) from the sibilant discrimination task. The accuracy data were converted to *d'* scores, a measure of perceptual sensitivity that takes response biases into account (Macmillan & Creelman, 2005). The response time (RT) data were scaled using z-score conversion, and data points exceeding 2.5 standard deviations from the mean were trimmed, which removed 395 tokens (3.94% of all data points collected) from further analysis. The RT values were then log-transformed.

2.4. Results

Figure 3 shows boxplots of the mean sensitivity (*d'*) scores and log RTs for the two vowel contexts across the two participant groups. Overall, the merged group ($M = 2.03$) showed lower sensitivity than the distinct group ($M = 2.88$), and the /u/ vowel context ($M = 3.07$) produced higher sensitivity than the /a/ vowel context ($M = 1.92$). The groups performed similarly in terms of RT, and overall, responses were faster with /u/ ($M = 389$ ms) than with /a/ ($M = 433$ ms).

The discrimination data were analyzed in mixed-effects regression models using the *lmer* function in the *lme4* package (Bates et al., 2015) in R (R Development Core Team, 2023).[7] Two separate models were fitted, one predicting sensitivity (*d'*) scores and the other predicting RTs. The models included fixed effects for MergerStatus (2 levels: merged $= -1$ vs. distinct $= 1$) and Vowel (2 levels: /a/ $= -1$ vs. /u/ $= 1$). The two categorical variables were sum-coded. An interaction term between MergerStatus and Vowel was also included in the model. Initially, Block (3 levels: block 1, block 2, block 3) was included in the model along with its interactions with other factors to address whether listeners' performance changes significantly over time. The model revealed no significant effect of Block and its interactions with other factors (all p values $>.05$). A model comparison using a log-likelihood test showed a lack of significant differences between the two models ($\chi^2 = 11.033$, $p = .1998$), based on which a simpler model was adopted as a final model.

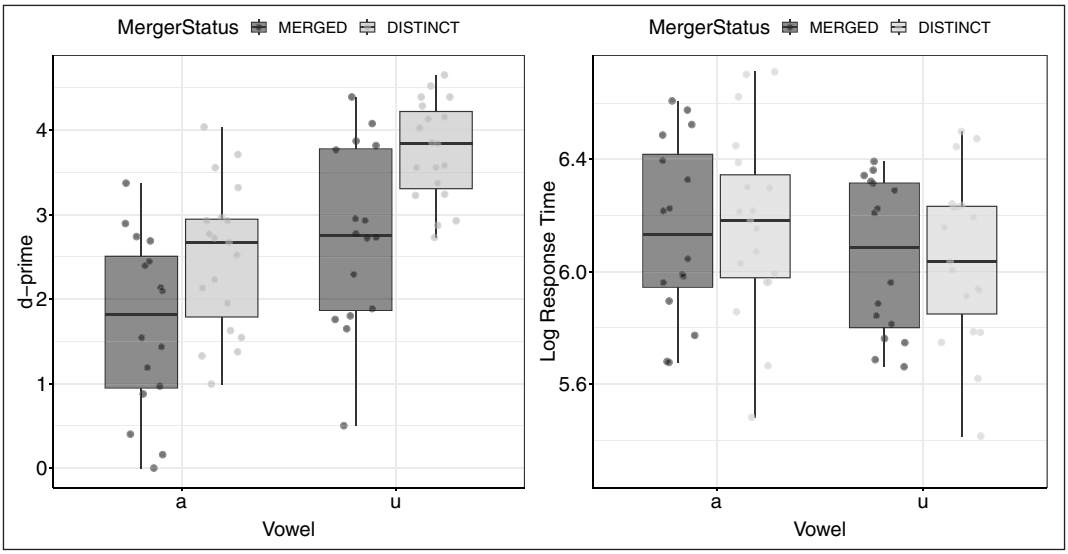


Figure 3: Boxplots of by-subject mean sensitivity (*d'*) scores (left) and log response times (right) by vowel and the participants' merger status.

[7] All data and R codes are available at OSF (https://osf.io/7b6fd/).

Random structures were extended as maximally as possible. Specifically, SUBJECT was included as a random intercept for both models, and MANNER was included only for the *d'* model because the RT model did not converge. We do not predict a systematic difference in the perception of sibilants of different manners of articulation, which is why MANNER (levels: fricatives, unaspirated affricates, aspirated affricates) was treated as a random intercept. In addition, VOWEL was included as a random slope for SUBJECT in the *d'* model. **Tables 2** and **3** summarize the two sets of statistical model fits.

The regression model predicting *d'* scores revealed significant main effects of MERGERSTATUS and VOWEL with no interaction between them. First, the merged participants (mean accuracy = 81%) were significantly less accurate at sibilant discrimination than the distinct participants ($M = 89\%, p < .01$). This result indicates a link between perception and production, i.e., the merged participants' representations of the contrasts are presumably less robust than those of the distinct participants. Nonetheless, the merged participants performed well above chance level (mean accuracies of the merged group: 74% in /a/ and 85% in /u/). Consistent with past works, speakers of the sibilant merger did not entirely lose their perceptual sensitivity to the phonological contrast. This study thus contributes a new piece of evidence from a sibilant category merger to the literature on vowel mergers (Wade, 2017; Austen, 2020; Hay et al., 2013; Hay et al., 2006b; Thomas & Hay, 2005) and tone mergers (Fung & Lee, 2019a; Mok et al., 2013).

Next, the model revealed that the vowel context had a significant effect on sibilant discrimination. Overall, participants performed significantly better with the /u/ vowel ($M = 90\%$) than with the /a/ vowel ($M = 79\%, p < .0001$). Furthermore, the MERGERSTATUS-VOWEL interaction turned out to be non-significant ($p = .2821$), indicating that the vowel effect was comparable across the two groups of participants. As reflected in **Figure 4** (left), distinct, as well as merged participants, showed an increased sensitivity to the /s/-/ʂ/ contrast with the /u/ vowel than with the /a/ vowel. The robust vowel effect regardless of merger status can be attributed to the salient acoustic cues in the frication noise before the /u/ vowel. Recall that the spectral differences in the frication noise were enhanced with the rounded vowel (**Figure 3**), which seems to have produced significant perceptual benefits.

The results of the RT model confirm that the /u/ vowel facilitated sibilant discrimination ($p < .0001$); the participants were faster at discriminating sibilants with the /u/ vowel ($M = 440$ ms) than with the /a/ vowel ($M = 497$ ms). Unlike the *d'* model, however, the RT model revealed no significant effect of MERGERSTATUS ($p = .6072$) or its interaction with VOWEL, confirming once again that the acoustic saliency of retroflexes is enhanced in the rounded vowel context. Along with the results of the *d'* scores, this perception experiment demonstrates that the merged participants are as fast as the distinct- participants in discriminating the sibilants, but they are more prone to inaccurate discrimination.

Predictors	Coefficient	*SE*	*t*	*p*
(Intercept)	**2.4573**	**0.1893**	**12.9804**	**<.0001**
MergerStatus	**0.4270**	**0.1303**	**3.2773**	**.0025**
Vowel	**0.5670**	**0.0673**	**8.4275**	**<.0001**
Vowel:MergerStatus	0.0736	0.0673	1.0935	.2821

Table 2: Summary of the mixed-effects regression model predicting discrimination sensitivity. Formula: *d'* ~ MergerStatus*Vowel + (1 + Vowel|Subject) + (1|Manner). Significant results are in bold.

Predictors	Coefficient	*SE*	*t*	*p*
(Intercept)	**5.8083**	**0.0509**	**114.1812**	**<.0001**
MergerStatus	0.0264	0.0509	0.5190	.6072
Vowel	**−0.0504**	**0.0094**	**−5.3723**	**<.0001**
Vowel:MergerStatus	0.0028	0.0094	0.2945	.7703

Table 3: Summary of the mixed-effects regression model predicting response time. Formula: logRT ~ MergerStatus*Vowel + (1 + Vowel|Subject). Significant results are represented in bold.

3. Experiment 2: Sibilant identification with social guises

Having established the perceptual performances of merged participants in comparison with distinct participants, here we present the results of the identification task designed to show the interaction between social and acoustic cues in speech categorization.

3.1. Participants

All speakers except for two distinct women (40 women, 33 men) reported in **Table 1** and **Figure 2** participated in the identification experiment.

3.2. Stimuli

All sibilants digitally manipulated on the 8-step continuum (S1-S2-S3-S4-S5-S6-S7-S8) in (1), as well as all filler items, were used in the identification experiment. Each block contained 90 tokens (48 targets (3 manners × 2 vowels × 8 steps) + 42 fillers), and all tokens were repeated three times across three blocks, resulting in 270 trials for each participant.

3.3. Procedure

The identification experiment was conducted individually in a sound-attenuated booth. Participants were randomly assigned to one of the social conditions. In the Taipei condition,

they were told that the talker was originally from Taipei (台-北, Taiwan-north), the capital city of Taiwan located in the North, and does not speak TSM. In the Tainan condition, they were told that the talker was from Tainan (台-南, Taiwan-south), a city in the Southern part of Taiwan, and has a good command of TSM. In this experiment, 19 merged and 22 distinct participants were assigned to the Tainan condition, and 13 merged and 19 distinct participants were assigned to the Taipei condition.

To strengthen the connection between the talker and his social background, participants were instructed to watch a YouTube video of a singer performing a popular song in their respective languages prior to the experiment. The singer and his songs were described as the talker's favorite music. The singer *EggPlantEgg* (茄子蛋), used in the Tainan condition, is well known among young Taiwanese for singing songs in TSM. TSM songs were once thought to be old-fashioned and a cultural product of old TSM speakers; however, along with the recent changes in language ideology, TSM has recently been promoted as a symbol of regional identity (Cheng, 1978, 1989; Huang, 1993b). Participants in the Taipei condition watched a video clip of the singer *Wu Qing-feng* (吳青峰) who does not speak TSM, so his songs are exclusively in TM. In addition, landmark photos of the respective cities were presented on the computer screen between trials during the experiment. **Table 4** shows screenshots from the music videos and the photos used in the experiment. The socio-indexical information was provided implicitly in our experiment; however, we exploited resources shared widely among young people in Taiwan because matched guise experiments require effective and engaging social guises.

The participants took part in a 2-AFC identification task administered in E-Prime 3.0. As in the discrimination task, a fixation mark appeared on the screen while the sound was played. Upon completion of the sound file, two numbered words in the *Zhuyin* phonetic alphabet appeared vertically on the screen (e.g., /a/: 1. ㄙㄚ [sa] vs. 2. ㄕㄚ [ʂa]; /u/: 1. ㄙㄨ [su] vs. 2. ㄕㄨ [ʂu]). Participants were instructed to press "1" for the alveolars and "2" for the retroflexes. The order between the two sibilants was fixed throughout the experiment. For filler items, the task involved the differentiation of vowels (e.g., ㄧㄝ [iɛ] vs. ㄩㄝ [yɛ]) or stop aspirations (e.g., ㄉㄨˊ [tu^{35}] vs. ㄊㄨˊ [t^{h}u^{35}]). After the participant responded, a photo of the relevant city appeared on the screen for 500 ms, and then the experiment moved on to the next trial. The participants finished three practice trials before the test session to ensure that they were familiar with the task. They were told that the task was timed and encouraged to respond as quickly and accurately as possible. Being administered in an AFC task, they were asked to make their best guesses and respond to every trial even when they were uncertain. Other than that, no particular explicit instructions especially on social cues were provided to the participants. The E-Prime script was designed to move on automatically if the participant did not respond within 3 seconds. Overall, it took approximately 30 minutes for participants to complete the whole experiment, including watching the video clip for five minutes.

	Tainan condition	Taipei condition
Before experiment		
During experiment		

Table 4: (top) Screenshots from the YouTube video clips of singers singing in TSM (left) and TM (right), and (bottom) representative landmark buildings, Anping Castle in Tainan (left) and Taipei 101 in Taipei (right).

There were no missing responses, so 19,710 data points (270 trials × 73 speakers) were compiled from the sibilant identification task. Excluding the 9,198 filler items, the data analysis was performed using the 10,512 data points drawn from the target items containing initial sibilants.

3.4. Results

Figure 4 plots the mean retroflex choices as a function of the sibilant steps (from the most alveolar-like to the most retroflex-like stimuli) and the two social guise conditions. The results are divided by the listeners' merger status and vowel context.

To analyze the results of the binary responses, a mixed-effects logistic regression model was fitted to the identification data using the *glmer* function in the *lme4* package (Bates & Maechler, 2015) in R (R Development Core Team, 2023). The dependent variable was the listeners' judgment of the initial sibilants as retroflex (coded as "1") or alveolar (coded as "0"). The model included fixed effects for MergerStatus (2 levels: merged = −1 vs. distinct = 1), Vowel (2 levels: /a/ = −1 vs. /u/ = 1), SocCond (2 levels: Tainan = −1 vs. Taipei = 1), and SibStep (8 levels: −3.5, −2.5, −1.5, −0.5, 0.5, 1.5, 2.5, 3.5, reflecting the eight-step sibilant continuum). All categorical variables were sum-coded.

Along with those main effects, two interaction terms were also included in the statistical model. First, a three-way interaction among SibStep, Vowel, and MergerStatus was included

in the model. The predictor SibStep was used to assess the effects of acoustic cues on sibilant identification. A participant's reliance on acoustic cues would manifest in the response slopes in **Figure 4**, with steeper slopes indicating greater reliance on noise properties during sibilant identification. The interaction between SibStep and Vowel, therefore, helps to elucidate whether the effects of acoustic cues were further modified by the vowel context. The three-way interaction SibStep-Vowel-MergerStatus examines whether the two groups of listeners differed in their sensitivity to acoustic cues, potentially conditioned by vowel context. The second three-way interaction included in the statistical model was among SocCond, Vowel, and MergerStatus. The variable SocCond encodes the effect of social guises on sibilant identification. The magnitude of this effect is assessed by the divergence of the two curves in **Figure 4**, with greater divergence of the curves indicating that social information had a greater effect on sibilant identification. Its interaction with Vowel and MergerStatus, therefore, indicates whether the effects of social cues are modulated jointly by the vowel context and listener merger status. In addition to the main effects, the model included random intercepts for Subject and Manner. **Table 5** summarizes the results of the statistical model fit.

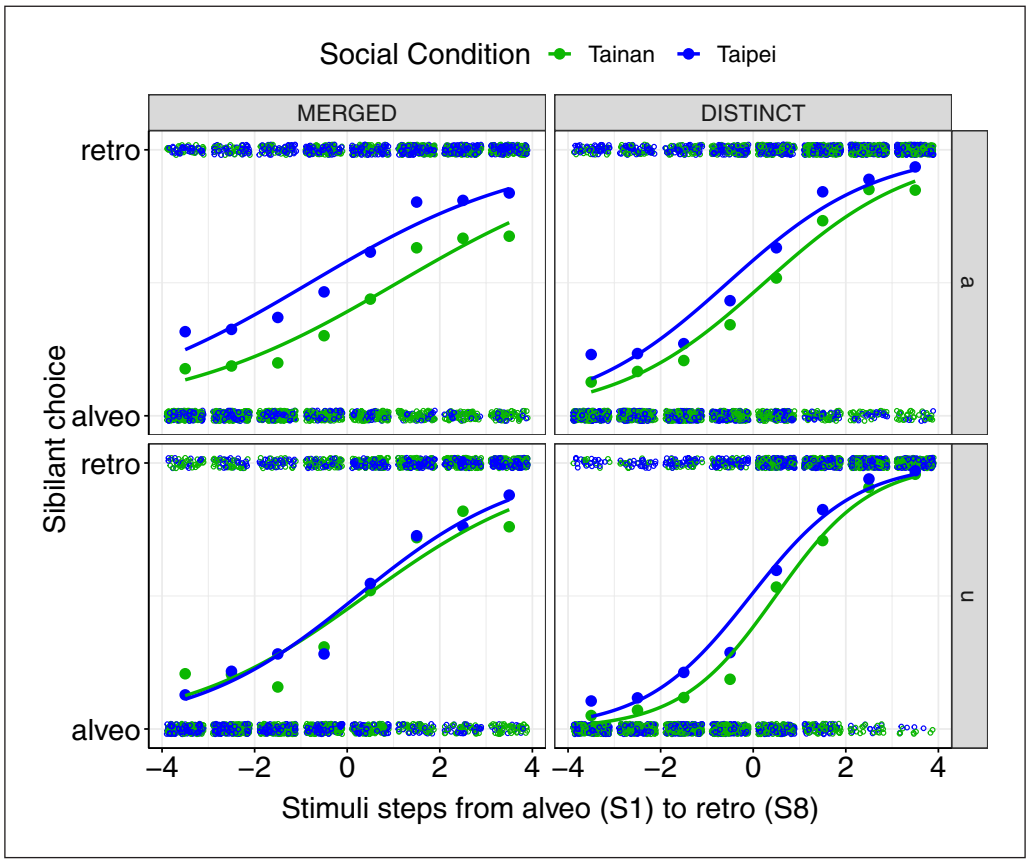


Figure 4: By-subject mean retroflex choices (filled circles) and logit curves for the identification of sibilants against SibStep and social conditions, Tainan (green) and Taipei (blue). Empty circles represent actual binary responses. Merged (left) versus distinct (right) participants and /a/ (top) versus /u/ (bottom) vowel contexts.

Predictors	Coefficient	*SE*	*z*	*p*
(Intercept)	**−0.0863**	**0.1458**	**−0.5919**	**.5539**
SibStep	**0.6775**	**0.0139**	**48.6524**	**<.0001**
Vowel	**−0.1085**	**0.0253**	**−4.2976**	**<.0001**
SocCond	**0.2514**	**0.086**	**2.923**	**.0035**
MergerStatus	0.0285	0.086	0.3319	.7400
SibStep:MergerStatus	**0.1688**	**0.0138**	**12.2667**	**<.0001**
SibStep:Vowel	**0.1139**	**0.0132**	**8.6168**	**<.0001**
Vowel:MergerStatus	**−0.0514**	**0.0253**	**−2.0351**	**.0418**
SocCond:Vowel	**−0.0907**	**0.0254**	**−3.5783**	**.0003**
SocCond:MergerStatus	0.0205	0.086	0.2383	.8117
SibStep:Vowel:MergerStatus	**0.0514**	**0.0132**	**3.8855**	**.0001**
SocCond:Vowel:MergerStatus	**0.0949**	**0.0254**	**3.7432**	**.0002**

Table 5: Summary of the mixed-effects logistic regression model predicting sibilant place perception (retroflex: 1, alveolar: 0). Formula: Retroflex.Choice ~ SibStep*Vowel* MergerStatus + SocCond*Vowel*MergerStatus + (1|Subject) + (1|Manner), family = binomial(link = "logit"). Significant results are represented in bold.

The predictor SibStep encodes differences in noise properties that were manipulated to range from the most alveolar-like sound to the most retroflex-like sound on an 8-step scale. The model fit revealed that listeners made significantly more retroflex responses as the frication shifted toward retroflex-like noise signals ($p < .0001$). As evident in **Figure 4**, the veridical responses to the acoustic cues are reflected as positive slopes across all experimental conditions. This is not surprising, given that the noise properties would be one of the most integral segment-internal cues to the phonological contrasts in question. The significant effect of this variable confirms, once again, that the merged participants can actively incorporate acoustic cues into sibilant identification, indicating that they did not lose sensitivity to the phonetic cues essential for the phonological contrast.

Nonetheless, the interaction between SibStep and MergerStatus ($p < .0001$) turned out to be significant, suggesting that the merged participants were less sensitive to the fine-grained acoustic details of the stimuli than the distinct participants. **Figure 4** shows that the slopes of the curves differ across the two groups of speakers; overall, the curves from the distinct group exhibit steeper slopes (mean slope = 0.8161) than those of the merged group (0.4664).[8] This can

[8] Here, the slopes are coefficient values of the predictor SibStep drawn from simple logistic regression models predicting response biases based on subsets of the group of interest.

be hinted at partly by the larger extent of the stretching curves toward the two endpoints, close to 100% retroflex or alveolar responses, by the distinct group. Notably, this mirrors the findings of the sibilant discrimination task, in which the merged group was outperformed by the distinct group.

The model fit also reveals a significant interaction between SibStep and Vowel ($p < .0001$), indicating that the reliance on the noise properties differed across vowel contexts. It was established independently through the discrimination task that the /s/-/ʂ/ contrasts are more salient with /u/ than with /a/, leading to more accurate discrimination of the sibilants in the former than in the latter, regardless of an individual's merger status. Indeed, in sibilant identification, TM speakers relied more on the acoustic cues when the sibilants were preceded by the /u/ vowel than when preceded by the /a/ vowel. This is shown by steeper slopes in the former (mean slope = 0.8362) than the latter (0.6505) in **Figure 4**. Furthermore, the three-way interaction SibStep:Vowel:MergerStatus was significant ($p = .0001$), which suggests that the reliance on the noise cues changed across vowels to a greater degree for the distinct group than for the merged one. This finding further indicates that the distinct participants were more sensitive to the fine-grained acoustic details in the noise signal than their merged counterparts.

Turning to the effect of the social condition, the model fit confirmed the main effect of SocCond on sibilant identification ($p = .0035$). Overall, speakers perceived more retroflexes in the Taipei condition ($M = 53\%$) than in the Tainan condition ($M = 45\%$), as shown by the blue curves (Taipei condition) being above the green curves (Tainan condition) in most panels in **Figure 4**. This result reflects the implicit bias held by TM speakers that a speaker from Tainan is less likely than one from Taipei to produce "proper" retroflexes. Here, it is worth noting that participant biases about the connection between regional affiliation and phonetic variants are reflected unambiguously in their perception, even though TSM fluency has lost its predictive power for the sibilant merger in production (see Section 2.1).

Interestingly, the model fit revealed a significant two-way interaction between SocCond and Vowel ($p = .0003$), which further interacted with MergerStatus ($p = .0002$). These interactions are reflected in the relative distance between the two curves across the experimental conditions. In **Figure 4**, the two curves approximate one another in the /u/ context, indicating that the effect of social information is greatly attenuated in the rounded vowel condition. Notably, this pattern is clearer in the merged group. This finding suggests that the acoustic and social cues are in a trade-off relationship for sibilant identification: when acoustic cues are sufficiently salient, as in the /u/ vowel context, listeners do not necessarily rely much on the social cues. To put it differently, the effects of extralinguistic social cues emerge most clearly when the acoustic cues are ambiguous, as in the /a/ context. Furthermore, the significant three-way interaction suggests that the merged participants switched between the two types of cues to resolve ambiguity in the speech signal more readily than the distinct participants.

To demonstrate the tradeoff between acoustic and social cues, data from two individual speakers from the merged group are presented in **Figure 5**. These two speakers, HRY and ZHW, are men who merged the sibilants in the baseline production task. In the /a/ vowel context, HRY, who was assigned to the Tainan condition, perceived alveolars most of the time, whereas ZHW (in the Taipei condition) heard mostly retroflexes and simply ignored the changes in the acoustic properties of the frication noise. Even those two speakers, who appeared to hold rather extreme implicit bias, however, exhibited relatively good sensitivity to acoustic cues when the sibilants were preceded by the /u/ vowel, as evident in the steeper slopes in the /u/ condition (bottom). This result demonstrates the dynamic interplay between the acoustic and social cues at the individual level.

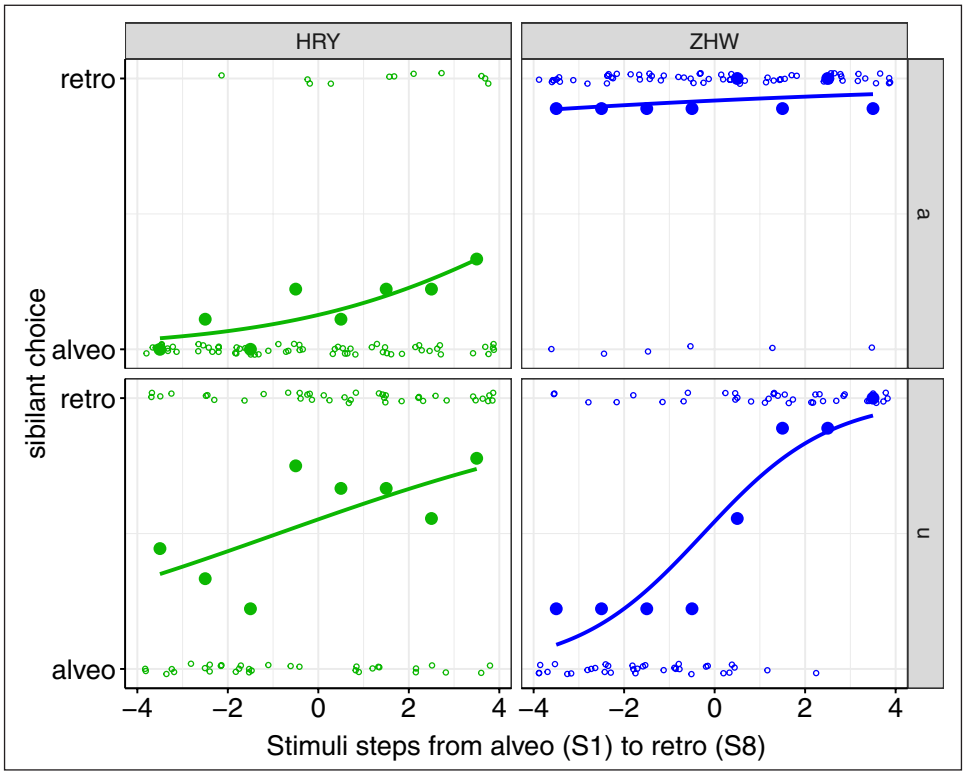


Figure 5: Data for two men from the merged group. HRY was assigned to the Tainan (green) condition, and ZHW was assigned to the Taipei condition (blue).

4. Discussion

4.1. Linguistic experience and the maintenance of marginal phonological contrasts

In this study, we investigated the perception of alveolar-retroflex sibilants in TM in conjunction with the speech production patterns of the participants. The results of the discrimination task show that merged participants were able to discriminate the sibilants far beyond chance level

(80% mean accuracy), but they did so less accurately than distinct participants (88% accuracy) (**Figure 3**). Similarly, the merged group was less sensitive to acoustic cues in the frication noise than the distinct group, as shown by the flatter slopes in the sibilant identification task (**Figure 4**). These results demonstrate that the pattern—merger in production versus distinction in perception—is indeed widely attested for phonemic mergers and should therefore be considered as a genuine linguistic pattern in the context of changes-in-progress.

The greater perceptual sensitivity to phonological contrasts observed among participants who convey the distinction aligns with patterns documented for English vowel mergers (Wade, 2017; Austen, 2020; Hay et al., 2006b; Thomas & Hay, 2005). This perceptual asymmetry, driven by an individual's production patterns, suggests less robust phonemic category representations among merged speakers. The two categories in question may be approximated, to some extent, in their mental representation, leading to the frequent merging observed in production. Assuming a perception-production inner loop, in which a speaker's production outputs are continually monitored by their perceptual system (Levelt, 1983, 1993), merged outputs likely reinforce their abstract representations. This reinforcement, in turn, could lead to poorer behavioral performance in perceptual tasks.

Nonetheless, one of the key findings of this study is that the merged participants still retained moderate sensitivity to the sibilant contrasts, leading to an apparent perception-production misalignment. The question that naturally arises then is how those speakers retain sensitivity to contrasts that they do not distinguish in production. One possible reason is that merged speakers are exposed to distinct speech in a speech community in which large variation exists among the speakers (Hay et al., 2013; Hay et al., 2006b; Thomas & Hay, 2005; Warren & Hay, 2006). This is apparently the case for the TM sibilants, as shown in instrumental studies (Chuang et al., 2019; Lee-Kim & Chou, 2022) as well as impressionistic descriptions (Chung, 2006). Frequent encounters with exemplars of distinct tokens could keep the categories from being completely merged in the mental representations of merged speakers.

Other accounts for the lack of a perception-production link in mergers-in-progress do not stand in the current case. One possibility is that the contrasts maintained in a non-merging context could facilitate the perception of distinct categories (Austen, 2020). For example, /ɪ/ and /ɛ/ are merged before nasals in the PIN-PEN merger, but the vowel contrast is robust in non-pre-nasal contexts even among habitually merged speakers (e.g., *bit* vs. *bet*). Merged speakers might, therefore, extend their sensitivity to acoustic cues available in non-merging contexts to the merging context. However, the TM sibilant merger is unconditional; the sibilant fricatives appear only in the syllable initial position (Lin, 1988), and once merged, no other phonological contexts are available for category distinctions. Another account based on spelling (Di Paolo & Faber, 1995; Herold, 1990) does not work for the Mandarin sibilant merger either, because the Chinese writing system is principally logographic and does not provide transparent

phonetic cues. The merged participants' moderate perceptual ability in this study was thus most likely to arise from their lifetime exposure to distinct forms carried by other speakers in the speech community.

Additionally, the high social awareness of the sibilant merger might have boosted the merged participants' sensitivity to the distinct categories. The sibilant merger is subject to overt comments among TM speakers (Chung 2006), which might have facilitated the maintenance of separate exemplars for the two categories. Notably, speakers of the sibilant merger were shown to be able to unmerge the sibilants when social primes were evident. In Lee-Kim and Chou (2022), merged speakers enhanced their sibilant contrasts by recovering the retroflex categories in a laboratory setting in which they interacted with a high-status experimenter. Therefore, TM speakers of the sibilant merger appear not to be truly merged, instead retaining at least some knowledge of the phonological contrast in their mental lexicon.

4.2. Social information and expectation-driven speech perception

This study has extended previous reports on the role of social information in speech perception by testing Mandarin speakers of the sibilant merger. The results of the identification task accompanied by social guises demonstrated that TM speakers overall gave significantly more retroflex responses in the Taipei-label condition (associated with the sibilant distinction) than the Tainan-label condition (associated with the sibilant merger). This result demonstrates that TM speakers are sensitive to social cues presumably rooted in an implicit bias—a talker from the south cannot produce "proper" retroflexes—that discourage them from hearing retroflexes in the southern talker condition.

The underlying motivation for this observed pattern is fundamentally different from one based on perceptual compensation. A perceptual compensation account, in fact, makes the opposite prediction; given an ambiguous noise signal, participants would be more likely to give retroflex responses for a talker from the south than for one from the north. Participants would retain the implicit bias that speakers from the south are less likely to produce retroflexes but would *compensate* for this lack of full retroflexion in a southern talker's speech during perception. In this scenario, participants would interpret ambiguous frication noise as the best possible retroflex for a southern talker, perceptually correcting the sound that falls short of full retroflexion into an acceptable retroflex token. In contrast, an ambiguous noise signal might not suffice to indicate a typical retroflex for a northern talker. A northern talker, if intending to produce retroflexes, would likely have done so with greater precision. This mechanism parallels the /s/-/ʃ/ perception conditioned by a talker's gender in English, wherein listeners are more likely to give /s/ responses for a male talker than for a female talker (Strand, 1999; Strand & Johnson, 1996). A noise signal is interpreted as sufficiently high in frequency due to the typically lower frequency values associated with a male talker's sibilants.

The production data clearly show that the implicit bias is rooted in extralinguistic knowledge drawn from social biases. Recall that the production experiment used to determine each participant's merging status revealed no significant effects of TSM-fluency on the likelihood of the sibilant merger (Section 2.1). The acoustic analysis of the sibilants produced by the participants found that the merger has become disconnected from its origin, TSM influence, and widespread in the speech community of Taiwan. Nonetheless, general TM speakers still hold the old connection between the sibilant merger and southern Taiwan in their minds and use it robustly in their perceptions. In other words, the implicit bias, which is deeply solidified among the members of the speech community, has persisted far beyond the point at which the original conditioning factor, Mandarin-Min language contact, has lost its predictive power on the distribution of the phonetic variant, and it continues to exert strong effects on the way listeners process incoming speech signals. This could be one reason why patterns of speech perception can diverge from those of production in the context of changes-in-progress.

The significant vowel context effects in both of the perceptual tasks in this study show how socio-indexical cues interact with acoustic cues to jointly influence speech categorization. The relationship between the two types of cues is a trade-off: greater reliance on one cue reduces the reliance on the other. Specifically, spectral differences between the two end points of the sibilant continua were larger with /u/ than with /a/ (**Figure 2**), indicating that the acoustic cues for the /s/-/ʂ/ contrasts are more salient in the former than in the latter. This asymmetry in acoustic saliency has consequences for perceptual performance: both the merged and distinct participants performed better and faster with the /u/ than with the /a/ in the discrimination task (**Figure 3**). In the identification task, the merged participants, in particular, showed reduced sensitivity to social cues, as shown by the two curves that approximate each other and become steeper in the /u/ context (**Figure 4**). This might run counter to the general observation for English /s/ and /ʃ/ contrasts, whose spectral differences typically become reduced with rounded vowels than with unrounded vowels (Jongman et al., 2000; Soli, 1981; Yu, 2019). It is possible that retroflexion is enhanced with lip rounding for the following vowel (Rau & Li, 1994), which needs to be independently verified in future studies.

Notably, the trade-off between social and acoustic cues was spelled out more clearly for the merged participants than the distinct participants. This is similar to Hay et al. (2006b), in which NZE speakers of the SQUARE-NEAR merger were shown to use social cues (i.e., age) more readily than the distinct group. Those authors speculated that speakers of the merger (low social status) were more likely to have had greater exposure to the distinct variety than the other way around. However, our result was drawn from groups with the same amount of linguistic experience with TSM, which we ensured by controlling for speaker regional/language background at the stage of the experimental design and also by independently verifying that the participants' language background had no significant effect on the likelihood that they would use the sibilant merger.

Alternatively, the merged participants' weaker perceptual sensitivity to the contrasts, verified in the discrimination task, could encourage them to seek other available cues, e.g., social cues, more readily, which resulted in a robust trade-off between social and acoustic cues. This speculation certainly warrants further investigation.

5. Conclusion

This study investigated the interplay between social expectations and a speaker's merger status on the perception of alveolar versus retroflex sibilants in Taiwan Mandarin. Consistent with previous studies of mergers-in-progress, speakers of the sibilant merger were less sensitive to acoustic cues in frication noise than those who maintain the distinction; however, their performance in perceptual tasks demonstrated moderate to good sensitivity to sibilant distinction. Despite their habitual merger, speakers may find it beneficial to relax the perception-production link to achieve efficient communication with others exhibiting different production patterns. Furthermore, the merged participants were able to make use of social cues, specifically the model talker's dialectal background, more readily than the distinct participants, especially when the acoustic signals were ambiguous. Merged participants seem to have developed a strategy to utilize various available cues, social information in this case, as well as the primary acoustic cues, presumably to compensate for their lower sensitivity to the phonological distinction. This study contributes a novel set of data from understudied sound categories to the literature on the frequent perception-production mismatch in vowel mergers.

Acknowledgements

We would like to thank Yun-Chieh Chou, Bamboo Liang, Kuei-hong Lin, and Wei-cheng Weng for their assistance with data collection and processing. The paper was significantly improved by the insightful comments of Yu-an Lu and Li-Hsin Ning. The core findings of this study have been presented at the Keio-ICU Linguistics Colloquium, the GULP meeting (Glasgow University Laboratory of Phonetics), LSA 2021, NWAV-AP6, HIPCS Lab meeting, ICPhS 20, and the Korean Phonology Circle. We are grateful to editor Bettina Braun, reviewer Marzena Żygis, and an anonymous reviewer for their constructive feedback at various stages of the manuscript. The first author dedicates this paper to the memory of her father, Tae-yeong Lee, with heartfelt appreciation for his unwavering faith in and support of her academic journey.

Funding

This work was supported by the research funds of the Ministry of Science and Technology, Taiwan (MOST110-2410-H-A49-058-MY3) and of Hanyang University, Korea (HY-202400000003765).

Competing Interests

The authors have no competing interests to declare.

Author Statement

Sang-Im Lee-Kim: conceptualization, methodology, formal analysis, writing of the original draft, and review and editing. Hsing-Yu Tung: stimuli selection, subject screening and recruitment, data collection, data curation, formal analysis, and writing of the original draft.

References

Austen, M. (2020). Production and perception of the Pin-Pen merger. *Journal of Linguistic Geography*, *8*(2), 115–126. https://doi.org/10.1017/jlg.2020.9

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, R. S., Cheung, K. H., & Cheung, P. M. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change*, *15*, 211–225.

Blacklock, O. S. B. (2004). *Characteristics of variation in production of normal and disordered fricatives using reduced-variance spectral methods* [Unpublished doctoral dissertation]. University of Southampton.

Blacklock, O. S. B., & Shadle, C. H. (2003). Spectral moments and alternative methods of characterizing fricatives. *Journal of the Acoustical Society of America*, *113*, 2199. https://doi.org/10.1121/1.4780184

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer*. In (Version 6.1.30) www.praat.org

Chang, Y. H. S. (2012). *Variability in cross-dialectal production and perception of contrasting phonemes: The case of the alveolar-retroflex contrast in Beijing and Taiwan Mandarin* [Unpublished doctoral dissertation]. University of Illinois, Urbana.

Chao, Y. R. (1930). A system of tone-letters. *Le Maître Phonétique*, *45*, 24–27.

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. University of California Press.

Cheng, R. L. (1978). Taiwanese morphemes in search of Chinese characters. *Journal of Chinese linguistics*, *6*, 306–314.

Cheng, R. L. (1989). *Essays on written Taiwanese*. Zili Wanbao She Wenhua Chubanbu.

Chiu, C., Wei, P. C., Noguchi, M., & Yamane, N. (2019). Sibilant fricative merging in Taiwan Mandarin: An Investigation of tongue postures using ultrasound Imaging. *Language and Speech 63*(4), 877–897. https://doi.org/10.1177/0023830919896386

Chuang, Y. Y., & Fon, J. (2010). The effect of prosodic prominence on the realizations of voiceless dental and retroflex sibilants in Taiwan Mandarin spontaneous speech. *Proceedings of Speech Prosody 2010*, Paper 414. https://doi.org/10.21437/SpeechProsody.2010-168

Chuang, Y. Y., Sun, C.-C., Fon, J., & Baayen, R. H. (2019). Geographical variation of the merging between dental and retroflex sibilants in Taiwan Mandarin. *Proceedings of the 19th International Congress of Phonetic Sciences*, 472–476. https://doi.org/10.31234/osf.io/beapz

Chung, K. S. (2006). Hypercorrection in Taiwan Mandarin. *Journal of Asian Pacific Communication*, *16*(2), 197–214. https://doi.org/10.1075/japc.16.2.04chu

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley Girls and California vowels. *Journal of Sociolinguistics*, *19*(2), 241–256. https://doi.org/10.1111/josl.12115

Di Paolo, M., & Faber, A. (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, *2*(2), 155–204. https://doi.org/10.1017/S0954394500000326

Di Paolo, M., & Faber, A. (1995). The discriminability of nearly merged sounds. *Language Variation and Change*, *7*(1), 35–78. https://doi.org/10.1017/S0954394500000892

Feifel, K. E. (1994). *Language Attitudes in Taiwan: A Social Evaluation of Language in Social Change*. The Crane Publishing.

Fung, R. S. Y., & Lee, C. K. C. (2019a). Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *Journal of the Acoustical Society of America*, *146*(5), EL424–EL430. https://doi.org/10.1121/1.5133661

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892. https://doi.org/10.1515/ling.2010.027

Hay, J., Drager, K., & Thomas, B. (2013). Using nonsense words to investigate vowel merger. *English Language and Linguistics*, *17*(2), 241–269. https://doi.org/10.1017/S1360674313000026

Hay, J., Nolan, A., & Drager, K. (2006a). From fush to feesh: exemplar priming in speech perception. *The Linguistic Review*, *23*, 351–379. https://doi.org/10.1515/TLR.2006.014

Hay, J., Warren, P., & Drager, K. (2006b). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484. https://doi.org/10.1016/j.wocn.2005.10.001

Herold, R. (1990). *Mechanisms of merger: The implementation and distribution of the low back merger in eastern Pennsylvania* [Unpublished doctoral dissertation]. University of Pennsylvania.

Huang, S. (1993a). *Yuyan, Shehui yu Zuqun Yishi: Taiwan Yuyan Shehuixue de Yanjiu [Language, Society, and Ethnicity: A Study of the Sociology of Language in Taiwan]*. The Crane Publishing.

Huang, S. (1993b). *語言、社會與族群意識 [Language, society and ethnicity]*. The Crane Publishing.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263. https://doi.org/10.1121/1.1288413

Kang, S., Johnson, K., & Finley, G. (2016). Effects of native language on compensation for coarticulation. *Speech Communication*, *77*, 84–100. https://doi.org/10.1016/j.specom.2015.12.005

Kei, J., Smyth, V., So, L. K. H., Lau, C. C., & Capell, K. (2002). Assessing the accuracy of production of Cantonese lexical tones: A comparison between perceptual judgement and an instrumental measure. *Asia Pacific Journal of Speech, Language and Hearing*, *7*, 25–38. https://doi.org/10.1179/136132802805576535

Khoo, H. L. (2019). The language attitudes in post Guoyu movement era in Taiwan—A study of Taiwanese young people's attitudes towards five Mandarin varieties. *臺灣語文研究 [Journal of Taiwanese Language and Literature]*, *14*(2), 217–253.

Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *Penn Working Papers in Linguistics*, Article 12.

Kubler, C. C. (1985). The influence of Southern Min on the Mandarin of Taiwan. *Anthropological Linguistics*, *27*(2), 156–176.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Walter de Gruyter. https://doi.org/10.1515/9783110167467

Labov, W., Karan, M., & Miller, C. (1991). Near-mergers and the suspension of phonemic contrast. *Language Variation and Change*, *3*, 33–74. https://doi.org/10.1017/S0954394500000442

Labov, W., Yaeger, M., & Steiner, R. (1972). *A Quantitative Study of Sound Change in Progress*. The U. S. Regional Survey.

Lee-Kim, S. I. (2014). *Contrast neutralization and enhancement in phoneme inventories: Evidence from sibilant place contrast and typology* [Unpublished doctoral dissertation]. New York University.

Lee-Kim, S. I., & Chou, Y.-C. (2022). Unmerging the sibilant merger among speakers of Taiwan Mandarin. *Laboratory Phonology*, *13*(1), Article 10. https://doi.org/10.16995/labphon.6446

Lee-Kim, S. I., & Chou, Y.-C. (2024). Unmerging the sibilant merger via phonetic imitation: Phonetic, phonological, and social factors. *Journal of phonetics*, *103*, Article 101298. https://doi.org/10.1016/j.wocn.2024.101298

Levelt, W. J. (1993). *Speaking: From intention to articulation*. MIT press.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Lin, Y., Yao, Y., & Luo, J. (2021). Phonetic accommodation of tone: Reversing a tone merger-in-progress via imitation. *Journal of phonetics*, *87*, Article 101060. https://doi.org/10.1016/j.wocn.2021.101060

Lin, Y. H. (1988). Consonant variation in Taiwan Mandarin. *Linguistic Change and Contact: New Ways of Analyzing Variation (NWAV) XVI*, 200–208.

Lin, Y. H. (2007). *The Sounds of Chinese*. Cambridge University Press.

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics*, *68*(7), 1227–1240. https://doi.org/10.3758/BF03193723

Mok, P. P. K., Zuo, D., & Wong, P. W. Y. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change*, *25*, 341–370. https://doi.org/10.1017/S0954394513000161

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology*, *18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005

Psychology Software Tools, Inc. (2016). *E-Prime 3.0*. https://support.pstnet.com/

R Development Core Team. (2023). *R: A language and environment for statistical computing*. In (Version 4.3.0) R Foundation for Statistical Computing. http://www.r-project.org/

Rau, D., & Li, J. (1994). Phonological variation of (ts), (tsh), and (s) in Mandarin Chinese. *Proceedings of the 4th International Conference on Teaching Chinese as a foreign Language*, 345–366.

Sandel, T. L. (2003). Linguistic capital in Taiwan: The KMT's Mandarin language policy and its perceived impact on language practices of bilingual Mandarin and Tai-gi speakers. *Language in Society*, *32*, 523–551. https://doi.org/10.1017/S0047404503324030

Schertz, J., Kang, Y., & Han, S. (2019). Sources of variability in phonetic perception: The joint influence of listener and talker characteristics on perception of the Korean stop contrast. *Laboratory Phonology*, *10*(1), Article 13. https://doi.org/10.5334/labphon.67

Shih, Y. T. (2012). *Taiwanese-Guoyu bilingual children and adults' sibilant fricative production patterns* [Unpublished doctoral dissertation]. The Ohio State University.

Smits, R. (2001). Evidence for hierarchical organization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception & Performance*, *27*(5), 1145–1162. https://doi.org/10.1037//0096-1523.27.5.1145

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*(4), 976–984. https://doi.org/10.1121/1.387032

Staum Casasanto, L. (2010). What do listeners know about sociolinguistic variation? *Penn Working Papers in Linguistics, 15*.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86–100. https://doi.org/10.1177/0261927X99018001006

Strand, E. A., & Johnson, K. (1996). Gradient and Visual Speaker Normalization in the Perception of Fricatives. In G. Dafydd (Ed.), *Natural Language Processing and Speech Technology* (pp. 14–26). De Gruyter Mouton. https://doi.org/10.1515/9783110821895-003

Su, H. Y. (2008). What does it mean to be a girl with *qizhi*?: Refinement, gender and language ideologies in contemporary Taiwan. *Journal of Sociolinguistics*, *12*(3), 334–358. https://doi.org/10.1111/j.1467-9841.2008.00370.x

Thomas, B., & Hay, J. (2005). A pleasant malady: The ellen/allan merger in New Zealand English. *Te Reo*, *48*.

Tse, J. K. P. (1998). Do the young people of Taiwan really not distinguish between zh, ch, sh and z, c, s? *The World of Chinese Language*, *90*, 1–7.

Tse, J. K. P. (2000). Language and a rising new identity in Taiwan. *International Journal of the Sociology of Language*, *143*(1), 151–164. https://doi.org/10.1515/ijsl.2000.143.151

Wade, L. (2017). The role of duration in the perception of vowel merger. *Laboratory Phonology*, *8*(1), 1–34. https://doi.org/10.5334/labphon.54

Warren, P., & Hay, J. (2006). *Using sound change to explore the mental lexicon*. Australian Academic Press.

Wei, X. (1984). *Changes in the Mandarin language in Taiwan*. National Taiwan University.

Whalen, D. H. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception & Psychophysics*, *46*(3), 284–292. https://doi.org/10.3758/BF03208093

Yiu, C. Y. (2009). A preliminary study on the change of rising tones in Hong Kong Cantonese: An experimental study. *Language and Linguistics*, *10*, 269–291.

Yu, A. C. L. (2019). On the nature of the perception-production link: Individual variability in English sibilant-vowel coarticulation. *Laboratory Phonology*, *10*(1), 1–29. https://doi.org/10.5334/labphon.97

Zhang, J. (2019). Tone mergers in Cantonese: Evidence from Hong Kong, Macao, and Zhuhai. *Asia-Pacific Language Variation*, *5*(1), 28–49. https://doi.org/10.1075/aplv.18007.zha