Laboratory Phonology
Journal of the Association for Laboratory Phonology

Lee-Kim, S.-I., & Chou, Y.-C. 2022. Unmerging the sibilant merger among speakers of Taiwan Mandarin. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 13(1): 10, pp. 1–36. DOI: https://doi.org/10.16995/labphon.6446

Open Library of Humanities

# Unmerging the sibilant merger among speakers of Taiwan Mandarin

**Sang-Im Lee-Kim,** Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, TW, sangimleekim@nycu.edu.tw

**Yun-Chieh Chou,** Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, TW, iris61030@yahoo.com.tw

This study presents empirical evidence from read versus interactive speech to shed light on the nature of the alveolar-retroflex sibilant merger by young speakers of Taiwan Mandarin (TM). TM speakers often merge the two sibilants through deretroflexion of the retroflex category. The results of the reading task showed that the variation is on a full continuum, from a complete merger to clear contrasts, and the merger is more prevalent among male speakers, demonstrating the impact of the social stigma associated with the merger. However, the results of the interactive task demonstrated that speakers who merged the contrast produced the retroflex sounds as distinct from their alveolar counterparts, revealing hidden structures in the mental lexicon. The mismatch between the abstract phonological knowledge and actual implementation in production suggests that the exposure to phonological systems of other speakers, especially those who make clear distinctions, has led to the incorporation of discrete categories into the phonological knowledge of the merged speakers. These findings suggest that large individual variation in the early stages of sound change may provide evidence for possible categories in a given language for language learners; however, their implementation may be further modulated by social as well as other phonetic factors.

# 1. Introduction

## 1.1. Unmerging of phonemic mergers

According to Garde's principle (Garde, 1962; Labov, 1994), a merger, once completed, cannot be reversed by linguistic means as traces of phonetic differences have been eliminated and learners have no means of accessing the earlier forms during language acquisition. Yet a number of sociolinguistic studies of vowel mergers have generally concluded that phonemic mergers are far from straightforward and cannot be defined in a simplistic and homogeneous fashion (Clark, Watson, & Maguire, 2013; Harris, 1985; Labov, 1994). Specifically, the vast majority of the studies of contemporary phonological systems involves cases of near-mergers or suspended contrasts, a pattern that involves a separation between production and perception (Labov, 1994, pp. 349–70). For example, the vowels in SAUCE versus SOURCE in *r*-less NYC speech were thought to be identical or indistinguishable in perception. Labov, Yaeger, and Steiner's (1972) instrumental study of the data revealed, however, that the vowel in SOURCE was consistently produced as more retracted and higher as though the /r/ had been retained as in different varieties of American English.

Near-merger patterns offer a unique opportunity to examine the nature of phonological representations and their connection with actual phonetic implementation in production. Most notably, small but consistent acoustic differences in production strongly suggest that speakers have some knowledge of the non-overlapping phonological categories stored in their mental lexicon. In a cleverly designed study, Hay, Drager, and Thomas (2013) compared real and nonce words containing the vowels undergoing a merger across the sets /ɛl/ versus /æl/ (ELLEN-ALLAN) in New Zealand English (NZE) and /ɑ/ versus /ɔ/ (LOT-THOUGHT) in American English. The results of a reading task showed that the speakers who merged the contrasts did so to a greater degree for real words than for nonce words. Framed in an exemplar-based model (Johnson, 1997; Pierrehumbert, 2001), the results were taken to indicate that, in the mental lexicon of the merged speakers, the pairs of vowels are represented as belonging to separate phonemes, which were realized as distinct in the production of the nonce words. In contrast, real words are stored as whole forms with acoustic details in the word-level representations, and the implementation of those phonetically rich abstract forms leads to considerable overlap between the two categories.

Some degree of separation, if not full distinction, in the mental representation suggests that the (nearly) merged categories could potentially be unmerged in certain experimental conditions. Hay, Drager, and Warren (2009) explored this idea in a series of word list reading tasks in which NZE speakers were exposed to two different experimenters, one from New Zealand and the other from the US. NZE speakers often merge the vowel contrasts in the NEAR-SQUARE lexical set by approximating the diphthong /eə/ toward /iə/. When interacting with the US experimenter, however, they unmerged the two vowels to a greater degree, accommodating the speech of the experimenter who retained the vowel distinction. In another study, Babel, McAuliffe, and

Haber (2013) showed that NZE speakers may also merge or unmerge this contrast based on their perception of their interlocutor. Specifically, NZE speakers merge the contrast less when exposed to an Australian model talker with clear vowel distinctions who was described as holding positive views of New Zealand. When that same talker was described as having a negative view of New Zealand, however, the speakers did not imitate the vowel distinction made by the model talker. Taken together, the results of these studies suggest that social factors may also mediate the unmerging of merged categories.

How, then, can learners acquire discrete categories when they cannot perceive the differences? To begin to answer this question, we must consider the environment in which near-mergers emerge. In a given speech community, speech varies greatly among individuals who are exposed to different phonological systems, some of which may make a clear distinction between categories that they themselves merge. Among the sub-patterns of near-mergers (Harris, 1985; Labov, 1994; Clark et al., 2013),[1] 'merger by approximation' refers to a case in which one of the categories is brought closer to the phonetic space of another category, eliminating the phonetic distinctions between the categories. Clark et al. (2013) argued that this type of merger involves approximation of one category to another which is likely to involve a gradual change, so the merger is likely to be incomplete and variable throughout the course of the linguistic change. Even though some speakers eliminate the contrasts entirely in their own speech, they are still exposed to clear distinctions, which may be incorporated into their phonological knowledge.

The impact of community-level variation is clearly demonstrated in the disassociation between perception and production patterns by the speakers who merge the contrast. That is, speakers may merge the phonemic contrasts in their own speech, but they are not necessarily incapable of distinguishing the relevant categories in perception. For instance, Hay, Warren, and Drager (2006) showed that NZE speakers of the NEAR-SQUARE vowel merger were fairly accurate at identifying intended targets when forced to choose between minimal pair words (e.g., beer versus bear) given the stimuli produced by speakers who made the contrast. Hay et al. (2013) extended this to the perception of nonce words as well; they observed that NZE speakers were able to identify the /ɛl/ versus /æl/ contrasts in nonce word pairs (e.g., lellit versus lallit) nearly as well as those in real word pairs (e.g., Ellen versus Allan). Along these lines, Wade (2017) found that speakers in Youngstown, OH, who merged the POOL, PULL, POLE word sets were adequately sensitive to the secondary cues to the tense-lax vowel distinctions, namely vowel duration. Although they may have collapsed the tense-lax vowel contrasts along the spectral

[1] Other types identified in the literature include 'merger by transfer' which refers to a case where individual lexical items are mapped onto another item in the lexicon and 'merger by expansion' which refers to a sudden collapse in phonetic space between two or more categories. Each has argued to have characteristic features and are likely to go through different developmental paths. This paper focuses on 'merger by approximation,' as the sibilant merger in TM falls under this sub-pattern.

dimension, they relied more on vowel duration cues than speakers from Burlington, VT, who maintained the contrast when identifying the tense vowels, particularly those in the POOL lexical set. The weak link between perception and production has generally been interpreted to mean that speakers who merge a contrast still encounter distinct forms from other speakers in their speech community, and this experience helps them maintain separate exemplar clouds of the two categories (Hay et al., 2013; Hay et al., 2006; Nycz, 2013).

Building upon the previous research, the present study aims to contribute another piece of evidence for variable realizations of the phonemic merger conditioned by different experimental tasks. While earlier works have focused mainly on vowel mergers (e.g. Labov, 1994) or tonal mergers (Fung & Lee, 2019; Mok, Zuo, & Wong, 2013; Yu, 2007), the present study investigates the sibilant merger in TM. The merging of sibilants in TM is known to be variable, and its conditioning social factors have been extensively documented in traditional descriptive studies. In the first experiment, we use a reliable instrumental method to first establish the extent of merging among individual TM speakers and thereby assess how far the phonetic variation has progressed among the younger members of the speech community. In Experiment 2, we then examine whether the speakers who merged the contrast in Experiment 1 would be able to unmerge the category, using a novel application of the methodology that has been used to elicit contrastive hyperarticulation in speech production. In the following, the current status of the sibilant merger in TM is introduced in Section 1.2, and the details of the methodology and its core findings are reviewed in Section 1.3.

## 1.2. Variation in sibilant merging in Taiwan Mandarin

Standard Mandarin makes a phonemic distinction between alveolars /s ts ts$^{h}$/ and retroflexes /ʂ tʂ tʂ$^{h}$/. In the Taiwanese variety of Mandarin, however, speakers often merge the two categories primarily through deretroflexion of the retroflex category (Ing, 1984; Kubler, 1985). Some scholars have pointed out that the variation is not categorical but exists on a continuum (Chung, 2006; Lin, 2007). Chung (2006, p. 200), for example, argued that Taiwan Mandarin (TM) speakers' retroflex production demonstrates considerable variation ranging "from highly retracted, through the palato-alveolar area, [tʃ], [tʃ$^{h}$], [ʃ], all the way to dentals that are indistinguishable from the dental/apical z- [ts̪], c- [ts̪$^{h}$], s- [s̪] series." The impressionistic descriptions have been partially verified by instrumental studies. Chiu, Wei, Noguchi, and Yamane (2019), for example, investigated the production of TM sibilants by seven speakers using ultrasound imaging and identified three different groups: overlap, non-overlap, and context-dependent overlap. The tongue curves of the two overlap speakers were indistinguishable from the tongue tip to root for alveolar and retroflex sibilants. For the two non-overlap speakers, the tongue curves significantly diverged at the tongue tip with the alveolar sibilants exhibiting a significantly more concave tongue body gesture than their retroflex counterparts. For the remaining three speakers, sibilant merging was conditioned by vowel context. The articulatory differences were faithfully reflected in the acoustic properties of the frication noise operationalized by the Center of Gravity (CoG) values. Notably, even in Chiu et al.'s (2019) small sample of speakers, the varying degrees of sibilant merging was captured.

It is worth noting that even the TM speakers who maintain the contrast between sibilants diverge from the reported norms of retroflex articulation in Mainland Mandarin. For example, Chang (2012) compared sibilants produced in contrastive focus by Mainland Mandarin speakers with those produced by TM speakers who made clear phonemic distinctions. The results showed that the spectral distance between the two sibilants in Mainland Mandarin was larger ($\Delta M$ = 3,955 Hz: $M$(alveolar) = 9,116 Hz – $M$(retroflex) = 5,161 Hz) than that in TM ($\Delta M$ = 2,534 Hz: $M$(alveolar) = 8,758 Hz – $M$(retroflex) = 6,224 Hz), irrespective of the presence of contrastive focus. The differences were primarily driven by the higher spectral means of the TM retroflexes, indicating weaker retroflexion in TM. Anecdotally, the strong retroflexion in Mainland Mandarin is hardly ever observed in the speech of TM speakers. Chung (2006) argued that intermediate retroflexion has emerged as the socially neutral and acceptable norm for the retroflex category in Taiwan. The differences in sibilant production are well-known among the speakers from both regions and are considered one of the most salient dialect-distinguishing features (Chang, 2017).

The conditioning factors of the variation in TM sibilants have been subject to extensive research among Taiwanese scholars who have proposed an interaction between linguistic and social factors. The merger has long been thought to arise from language contact with Taiwanese Southern Min (TSM**,** hereafter, also commonly referred to as 'Taiwanese' or 'Hokkien') which lacks the retroflex category altogether (Ing, 1984; Kubler, 1985), a feature typical of Southern varieties of Chinese (e.g., Cantonese, Shanghainese, and Hakka, Chen, 1999). Though not the official language of Taiwan, TSM is a major substrate language which is spoken by 70% of the population and of which most TM speakers have at least some passive knowledge (Huang, 1993; Sandel, 2003). Given the asymmetrical sibilant inventories of the two languages in contact, the predominant pattern of the sibilant merger in TM is considered deretroflexion, the approximation of the retroflex sibilants toward the alveolar area (Kubler, 1985; Lin, 1988; Wei, 1984).

However, the direct connection between TSM and the sibilant merger seems to have weakened in recent years. In a large-scale phonetic study, Chuang, Sun, Fon, and Baayen (2019) tested 331 young ethnically-Min TM speakers. Their results showed that speakers living in Southern Taiwan were more likely to speak TSM fluently than those living in Northern Taiwan, consistent with the linguistic landscape of the island (Ang, 2010). However, TSM proficiency did not predict the degree of sibilant merging; fluent speakers of TSM did not necessarily merge the two categories more frequently. Rather, speakers who lived in rural areas were more likely to merge the sibilants than those residing in urban areas, indicating that the merger no longer originates from direct language contact with the local substratum languages and is becoming an independent feature of TM. Building upon this line of research, the present study aims to establish the patterns of the sibilant merger in terms of linguistic and social factors and to investigate the extent to which the merger has spread among the speech community in Taiwan. In particular, if TM speakers who are not in contact with TSM are found to consistently merge sibilants, one could argue that this once-substrate feature has begun to emerge as a general feature of TM.

Sibilant merging in TM is somewhat stigmatized due to its association with the substratum language, leading to linguistic stratification. Since the retreat of the Nationalist party to Taiwan in 1949, Standard Chinese was forcefully imposed on the local population, and the social prestige of TM has been deeply solidified among the speech community (Feifel, 1994; Huang, 1993; Sandel, 2003). Sociolinguistic interviews with young college students revealed that female speakers who care for refinement (‘*qizhi (氣質)’* in Mandarin) and higher social prestige tend to avoid using TSM (Su, 2008). According to one of the male interviewees, “Women really do not speak Taiwanese as much. Maybe they find it lacking in *qizhi*” (Su, 2008, p. 345), reflecting the deep-seated stereotype of TSM and its negative cultural connotation. Notably, a conventional belief among TM speakers is that TM females produce retroflexes better than male speakers, which has been confirmed by researchers (Jeng, 2006; Tse, 1998). Deretroflexion, as one of the representative features of TSM, is thus found much less frequently in female speech than in male speech.

The goal of the current study is to frame the TM sibilant merger in terms of the relevant linguistic and social factors. Unlike phonemic mergers that are generally neutral in terms of social standing, the sibilant merger in TM is highly sensitive to social factors as discussed above. Merging is thus expected to be highly variable among individuals. Moreover, while some speakers may not contrast the sibilants themselves, the distinct categories produced by other speakers may have been incorporated into their phonological knowledge. It is therefore possible that the merged categories could be unmerged by such speakers in a particular experimental setting. To that end, we capitalize on an experimental paradigm that is known to elicit ‘contrastive hyperarticulation,’ an enhancement of phonetic cues to phonemic contrasts due to the existence of lexical competitors, as reviewed below. We investigate the conditions in which the sibilant contrast is fully realized, reflecting the underlying representations stored in the speakers’ mental lexicon.

### 1.3. Contrastive hyperarticulation in speech production

Minimal pair competitors, a special form of phonological neighbor, are known to trigger a significant enhancement of phonetic cues associated with the relevant phonological contrasts. Wedel, Nelson, and Sharp (2018), for example, used *The Buckeye Corpus of Conversational Speech* (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005) to investigate the ways in which phonetic cues are hyperarticulated as a function of the existence of lexical competitors. The results showed that English voiceless stops with lexical competitors tended to be produced with *longer* VOT (e.g., ***p**at/**b**at*) than those lacking competitors (e.g., ***p**ant/****b**ant*). In contrast, English voiced stops with lexical competitors were produced with *shorter* VOT (e.g., ***b**at/**p**at*) compared to those without (e.g., ***b**adge/****p**adge*). The opposite direction of the VOT changes of voiceless and voiced stops resulted in the enhancement of the distance between the two contrasting categories. Similar effects were found for vowel contrasts: Lax vowels tended to be centralized while tense vowels tended to be peripheralized in the vowel space (e.g., [ɪ] in *sh**i**p* and [i] in *sh**ee**p*).

The effect of lexical competitors on speech production has been shown to be robust, particularly in interactive tasks performed with a partner (Baese-Berk & Goldrick, 2009; Buz, Tanenhaus, & Jaeger, 2016; Kirov & Wilson, 2012; Schertz, 2013; Seyfarth, Buz, & Jaeger, 2016). In their seminal work using a cooperative interactive paradigm, Baese-Berk and Goldrick (2009) examined the implementation of the stop voicing contrast in English.[2] In this study, a speaker (participant) and a listener (experimenter) sat face-to-face, each with a computer screen in front of them (see **Figure 4**). The speaker produced a target word among three candidates out loud for the listener who would then indicate the target by clicking it on his/her screen. Crucially, each target word appeared in three different conditions: Context, No Context, and No Competitor. In the Context condition, a target word appeared along with two other words, one of which was its lexical competitor and the other a filler (e.g., ***c**od* [target], *god* [competitor], *yell* [filler]). In the No Context condition, the same target word appeared with two fillers (e.g., ***c**od, lamp*, *yell*). The No Competitor condition contained a target lacking a legal lexical competitor along with two fillers (e.g., ***c**op, lamp, yell*). The results showed elongated VOT for the voiceless stops in both the Context and No Context conditions compared with those in the No Competitor condition. Furthermore, the VOT was enhanced to a greater degree when the targets were presented overtly with their competitors in the Context condition.

Mandarin sibilants have also been examined in an interactive task with a partner, though with some methodological differences. Chang and Shih (2015) utilized prosodic focus to elicit the contrast enhancement between Mandarin alveolar and retroflex sibilants in a map task. Notably, they pre-screened their participants and included only those who made clear contrasts, excluding speakers who merged the contrast. Their stimuli included a pair of non-word location names, with the target containing one of the two sibilants (e.g., **扎**狗海岸 '/**tʂa**koʊ/ beach') and

[2] In the same study, they obtained similar results in a single-word reading task without the simultaneous presentation of minimal pair competitors (Experiment 1). The interaction with an interlocutor is, therefore, not essential for contrastive hyperarticulation, based on which the authors argued for a production-internal mechanism as the source of the observed effect. However, the present study is not concerned with whether contrastive hyperarticulation arises from a production- or perception-oriented mechanism, or whether the effect stems from competition with minimal pair competitors or broad lexical neighbors (Fricke, Baese-Berk, & Goldrick, 2016; Kirov & Wilson, 2012). Rather, the paradigm is adopted here to address the novel question of whether speakers who merge a category have stored distinct categories in their mental lexicon.

the corresponding control item without any sibilants (e.g., **莽**狗海岸 ‘/**maŋ**koʊ/ beach’). The experimenter would ask a question about a location on the map using a control item, and the participant-speaker would correct the direction with the target containing the sibilant. Their production would thus be under contrastive focus, and the sibilant place contrast was predicted to be enhanced. However, neither the TM speakers nor the Mainland Mandarin speakers enhanced the contrast in this condition, which led the authors to conclude that the coronal sibilants were not subject to cue-enhancing hyperarticulation. This result may suggest that, unlike lexical competition, prosodic focus alone may not be sufficient to drive significant contrast enhancement for the sibilants. The stimuli used in this study were geographic nonce words and their lexical competitors were not legitimate Mandarin words (e.g., ***tsa**koʊ). Hence, it remains to be determined whether contrast enhancement could be obtained via lexical competition for the Mandarin sibilants.

To that end, the present study capitalizes on the experimental paradigm modeled after Baese-Berk and Goldrick (2009), which has been shown to elicit an enhancement of phonemic contrasts in the presence of lexical competitors. The employment of Chinese characters as stimuli is expected to enable phonological effects to be isolated while minimizing a direct orthographic interference/facilitation. Unlike previous interactive studies in which the Roman alphabet was used to present the stimuli, the study at hand uses logographic and phonologically opaque Chinese characters, which are expected to provide stronger evidence for contrastive hyperarticulation. Clear contrasts, if any, cannot be attributed to the visual cues available in the alphabetic encodings of the contrasting sounds.

Additionally, the current study differed from previous studies on contrastive hyperarticulation in one crucial regard. Previously reported cases of contrastive hyperarticulation have focused primarily on phonological contrasts that are robustly represented by the speakers of the language (e.g., stop voicing and vowel tenseness). In these cases, the presence of minimal pair competitors, explicit or implicit, consistently leads to the hyperarticulation of the phonetic cues that enhance the contrast. TM sibilant contrasts, however, are subject to large speaker variation. Some speakers merge the contrast, while others retain it. Still, others may fall between these two extremes. How, then, would speakers with different degrees of merging cope with an experimental task designed to elicit contrastive hyperarticulation using lexical competitors?

Sociolinguistic studies, in fact, have shown that eliciting contrasts via minimal pairs may not necessarily improve the contrast made by speakers who typically merge it (Johnson & Nycz, 2015; Labov, Karan, & Miller, 1991; Labov et al., 1972; Nycz, 2013). For example, Nycz (2013) used a variety of tasks, including naturalistic conversation, word lists, and minimal pair reading, to test Canadian speakers of the COT-CAUGHT merger who had been exposed to the NYC dialect, which makes a robust distinction. Surprisingly, the speakers merged the vowels in the minimal pair context but carried small but consistent vowel distinctions in their conversational speech. A similar trend was observed in the speech of individuals moving in the opposite direction, from split to merged: Advanced forms (i.e., merged contrasts) were observed more frequently in conversation than in the minimal pair context (Johnson & Nycz, 2015). The authors conjectured that different tasks may reveal the multi-faceted nature of a speaker's linguistic knowledge. While minimal pair contexts encourage speakers to express the phonetic norms of their original dialect variety, conversation reveals their adaptation toward the characteristics of the sounds predominant in their new speech community.

Building upon these previous works, the present study examines how the variable sibilant merger in TM is realized across different experimental tasks, which would help us understand the interplay between abstract representations, phonetic implementation, and the conditioning social factors. In Experiment 1, we first establish the extent and the range of the merger between individuals, thereby assessing the extent to which this pattern has spread among young TM speakers. Experiment 2 examines whether the speakers who merged the contrast in Experiment 1 could be induced to make clear contrasts using an interactive task designed to elicit contrastive hyperarticulation.

## 2. Experiment 1: Sibilant production in read speech

The first experiment was designed to characterize the sibilant production of individual TM speakers. As a merger-in-progress, it is expected that the speakers' levels of mergedness would vary.

### 2.1. Participants

Sixty native TM speakers (32 female, 28 male; aged 20–29) participated in the production study. The participants were divided into two subgroups with respect to their TSM proficiency: 31 TSM-fluent versus 29 TSM-weak speakers. Prior to data collection, participants were pre-screened for their proficiency in TSM as well as TM, Hakka, and English. In a language background questionnaire, participants rated their confidence of listening and speaking for each language on a 7-point Likert scale ('1' being not at all confident, '7' being highly confident). Those who rated their TSM proficiency as greater than 5 (TSM-fluent, hereafter) or lower than 3 (TSM-weak, hereafter) were invited to participate in the study. In addition to language proficiency, the questionnaire gathered details about birthplace, location of residency, family language background, and daily language use. As shown in **Table 1**, most TSM-fluent speakers were born and raised in Southern or Central Taiwan, and TSM-weak speakers were predominantly from Northern Taiwan, reflecting the general linguistic landscape in Taiwan (Ang, 1997; Ang, 2010).

| | TSM-fluent (*N* = 31) | TSM-weak (*N* = 29) |
|---|---|---|
| Age<br>Gender | *M* = 22.4 (2.4)<br>13 M; 18 F | *M* = 21.9 (1.9)<br>15 M; 14 F |
| Birthplace & Residency<br>(N: North; C: Central; S: South) | 4 N; 11 C; 16 S | 22 N; 4 C; 3 S |
| TM listening | 6.61 (0.62) | 6.09 (1.12) |
| TM speaking | 6.61 (0.56) | 5.86 (1.11) |
| TSM listening | 6.10 (1.22) | 2.07 (0.98) |
| TSM speaking | 5.71 (1.30) | 1.59 (0.58) |
| English listening | 4.97 (1.11) | 4.93 (1.27) |
| English speaking | 4.61 (1.17) | 4.32 (1.39) |
| Hakka listening | 1.40 (1.07) | 1.71 (1.51) |
| Hakka speaking | 1.19 (0.54) | 1.39 (0.99) |

**Table 1:** Mean (M) language proficiency ratings and standard deviation (in brackets) on a seven-point scale (1 = low confidence, 7 = high confidence) self-rated by the participants.

Notably, all speakers identified themselves as most fluent in TM. For TSM-fluent speakers, their fluency in TSM was ranked lower than that in TM, the difference between which reached statistical significance (listening: $t(44) = 2.102$, $p = .041$; speaking: $t(41) = 3.564$, $p = .001$). While TSM-weak speakers indicated low fluency in TSM, the mean TSM listening of 2.07 indicates that they had at least some passive knowledge of this language, consistent with descriptions in the literature (Huang, 1993; Sandel, 2003). Both groups self-rated their level of English proficiency as intermediate; there was no significant difference between groups (listening: $t(54) = .125$, $p = .901$; speaking: $t(53) = .865$, $p = .391$). It is worth noting that TSM-fluent speakers expressed higher confidence in TSM than in English, while TSM-weak speakers ranked their confidence in TSM lower than English.

Participants' TSM fluency was verified during their lab visit through a short conversation in TSM with the experimenters who were fluent in TSM. None of the participants reported any hearing or speech disorders. Participants received small monetary compensation for their time.

## 2.2. Stimuli

The stimuli consisted of four disyllabic words containing the sibilants /s ʂ tsʰ tʂʰ/ in the word-initial position followed by the vowel /a/ carried by Tone 1 (X$^{55}$).[3] The non-target second syllable of the stimuli was fixed with a labial initial followed by the vowel /a/. Stimuli items varied in their morphosyntactic compositions; for example, some were nouns (e.g., 沙發 ‘couch’), while others were phrasal (e.g., 撒滿 *sprinkle-full* ‘fully sprinkled’). Since frication noise is sensitive to the neighboring phonological environment, especially lip rounding, the phonological conditions were balanced in the selection of the stimuli words. None of the target items had minimal pair competitors distinguished by the initial sibilants (e.g., 沙發 /ʂa.fa/, */sa.fa/). In addition to the target items, 32 filler items were included. The experimental stimuli for Experiment 1 are listed in **Table 2**.

| | **alveolars** | **retroflexes** |
|---|---|---|
| fricatives | /**sa**$^{55}$ man$^{214/}$<br>撒滿<br>‘fully sprinkled’ | /**ʂa**$^{55}$ fa$^{55/}$<br>沙發<br>‘couch’ |
| aspirated affricates | /**tsʰ**a$^{55}$ pan$^{214}$/<br>擦板<br>‘wiping a board’ | /**tʂʰ**a$^{55}$ pan$^{55/}$<br>插班<br>‘transfer classes’ |

**Table 2:** Stimuli of the production study.

## 2.3. Procedure

Participants were recorded individually in a sound-attenuated booth in the Experimental Phonology lab at National Yang Ming Chiao Tung University in Taiwan. They were asked to read aloud each stimulus item on a computer screen in a frame sentence (/wo$^{21(4)}$ ʂuə$^{55}$____ tʂə$^{51}$kə$^{0}$ tsi$^{51}$/ “I say _ this word”). Because target words were carried by a frame sentence, they were under narrow focus, and participants’ production was clear and formal. Participants were familiarized

[3] Some of the words initially chosen as target items were treated as fillers in the analysis stage and were not analyzed further. For one, a pair of words with the unaspirated affricates (/tsa$^{55}$pa$^{214}$/ 紮吧 ‘tuck in!’ versus /tʂa$^{55}$man$^{214}$/ 扎滿 ‘injection’) was problematic: specifically, large speaker variation was found for /tsa$^{55}$/ 紮 ‘tuck.’ Unlike the dictionary transcription, many DISTINCT speakers produced it as /tʂa$^{55}$/. The spectral characteristics of this item patterned together with other retroflexes /ʂa tʂʰa/, indicating that this word was represented in the mental lexicon as having a retroflex. Second, the stimuli list also included corresponding sibilants followed by the vowel /u/. However, the rounded vowel sometimes leads to a bimodal spectral distribution (see Figure 4 in Lee-Kim, 2011, especially for the retroflex sibilants, which is not ideal for the spectral moment analysis (Forrest, Weismer, Milenkovic, & Dougall, 1988).

with the stimuli items prior to the recordings, and no prosodic disfluency or abnormalities were observed during the experiment. Participants clicked a computer keyboard to proceed to the next trial at their own pace. A randomized reading list was repeated five times (36*5 = 180 trials). All stimuli were presented in Chinese characters, which is logographic and essentially non-alphabetical. The recording was made using AKG C520L condenser microphone with Zoom H4n digital recorder at a sampling rate of 44,100 Hz.

The recordings were first annotated for the frication noise in Praat (Boersma & Weenink, 2020). Since the sibilants differed in the manner of articulation (i.e., aspirated affricates, fricatives), the boundaries of the noise signal had to be carefully labelled. For the aspirated affricates /tsʰ tʂʰ/, the burst and aspiration were segmented out, leaving only the frication noise for acoustic analysis. The frication noise was marked from the onset of high-frequency noise to the onset of aspiration or that of the following vowel. The separation between frication and aspiration was not always well defined; in some cases, the noise interval was fully fricated without aspiration, and in other cases the aerodynamic change was gradual making it difficult to place a clear-cut boundary. While generally following the criteria implemented in Chang and Shih (2015), we ensured that the noise portion with the highest energy concentration of the frication was located in the middle of the two boundaries for all cases. The middle interval of the frication noise was used for the acoustic analysis.

The segmented sibilants were then saved as individual sound files and submitted to a multitaper spectral analysis implemented in Matlab (Blacklock, 2004; Blacklock & Shadle, 2003). This particular spectral analysis ensures reliable and accurate spectral estimates and has been adopted in previous works (e.g., Koenig, Shadle, Preston, & Mooshammer, 2013; Lee-Kim, Kawahara, & Lee, 2014; Żygis, Pape, & Jesus, 2012). Good spectral estimates were particularly important for the present study because the TM speakers were likely to be on a continuum of intermediate retroflexion, and the spectral differences between the contrasting sibilants were likely to be gradient. For the spectral analysis, frequencies below 1,000 Hz were filtered out to eliminate the effects of voicing from surrounding vowels. A 25 ms window was applied to three intervals during frication (i.e., beginning, middle, and end). Only the data drawn from the middle portion of the frication noise are reported in the results, as it was expected to show the most genuine acoustic properties of the sounds under investigation without much interference from surrounding segments.

Among the four spectral moments (Forrest et al., 1988) drawn from the multitaper spectral analysis, the first moment (M1) representing the mean of the spectral energy distribution was used to summarize the noise property ('spectral mean,' hereafter). Following Chang and Shih (2015), we further computed 'spectral distance' for each speaker by subtracting the spectral mean of the retroflex sibilants from that of the alveolar sibilants ($\Delta M = M$(alveolar) – $M$(retroflex)). A spectral distance of zero would indicate the two categories had been merged, whereas a large spectral distance would indicate a speaker made a clear distinction between the two categories.

## 2.4. Results

**Figures 1** and **2** (top) summarize the distribution of spectral means of the two sibilants in boxplots with alveolars in light grey and retroflexes in dark grey for male and female speakers, respectively.[4] In these figures, TSM-fluent speakers are represented with solid lines, and TSM-weak speakers with dotted lines. In addition, the bottom panels of the figures plot the spectral distance (Δ*M*) computed for individual speakers. Speakers were aligned according to this value from the lowest to the highest. As evident from the figures, the spectral distance was on a full continuum ranging from zero, marked by dotted horizontal lines, to approximately 4 kHz, for both male and female speakers.

This finding confirms previous impressionistic descriptions about TM sibilants (Chung, 2006; Lin, 2007)—sibilant merging by TM speakers is on a continuum, ranging from fully merged to completely distinct categories, making it difficult to draw a boundary between speakers. The pattern can be more precisely summarized as such that TM speakers differ in terms of *the degree of merging*. The results also confirm the previous impressionistic and instrumental studies which found that TM sibilant contrasts are acoustically less salient than those found in Mainland Mandarin. The maximum spectral distance was around 4 kHz for the data based on 60 young TM speakers in this study, while those based on Mainland Mandarin have been reported to be greater than 4 kHz (Chang & Shih, 2015; Lee-Kim, 2011).

**Figure 3** illustrates example multitaper spectra of the sibilant fricatives of two male speakers from the two ends of the continuum, one apparent merged speaker (C21, top) and the other speaker retaining full contrasts (ZXK, bottom). The two sets of spectra of the merged speaker present a complete overlap in the frequency-amplitude dimension, making it impossible to distinguish the spectra of the alveolars from those of the retroflex sibilants. In contrast, the spectra of Speaker ZXK clearly show a bimodal distribution with the alveolar fricatives having energy concentration at higher frequencies and the retroflex fricatives at much lower frequencies. It was often the case that the retroflex spectra demonstrated sharp peaks at the frequency region 3–5kHz for those who made a clear distinction between the two sibilants.

The spectral means of the frication noise were analyzed further using two statistical methods. First, we performed a Bayes factor test using the Bayesian First Aid package (Bååth, 2014) in R v.4.0.3 (R Development Core Team, 2020), which enables quantified evidence for the lack of differences across experimental conditions (Rouder, Speckman, Sun, Morey, & Iverson, 2009). In this analysis, a MERGED speaker was defined as one who demonstrated no significant difference in spectral distribution between the underlying alveolars and retroflexes. A DISTINCT speaker was defined as one with a significant difference. The results of the Bayesian test showed that 12 male and 5 female speakers had completely merged the sibilants, while the rest of the speakers

[4] All data, R codes, and figures are freely available at https://osf.io/4r97f/.

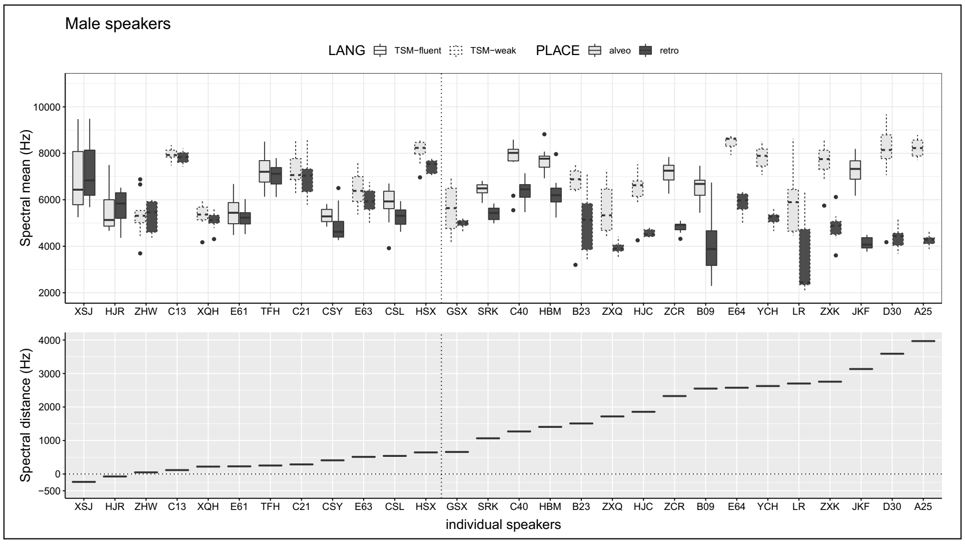

**Figure 1:** (top) Boxplots of spectral means of the alveolar (light grey) and retroflex (dark grey) sibilants produced by TM male speakers with differing TSM fluency. TSM-fluent speakers are represented in solid lines, and TSM-weak speakers in dotted lines. (bottom) Mean spectral differences between the alveolar and retroflex sibilants for individual speakers. The horizontal dotted line marks spectral distance of 0. The vertical dotted line divides speakers of the complete merger and those with clear contrasts.

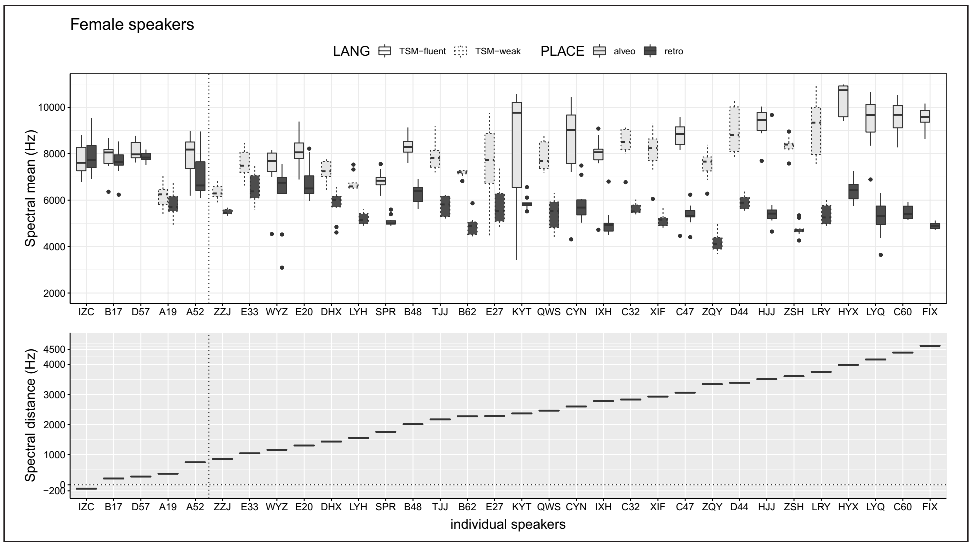

**Figure 2:** (top) Boxplots of spectral means of the alveolar (light grey) and retroflex (dark grey) sibilants produced by TM female speakers with differing TSM fluency. TSM-fluent speakers are represented in solid lines, and TSM-weak speakers in dotted lines. (bottom) Mean spectral differences between the alveolar and retroflex sibilants for individual speakers. The horizontal dotted line marks spectral distance of 0. The vertical dotted line divides speakers of the complete merger and those with clear contrasts.

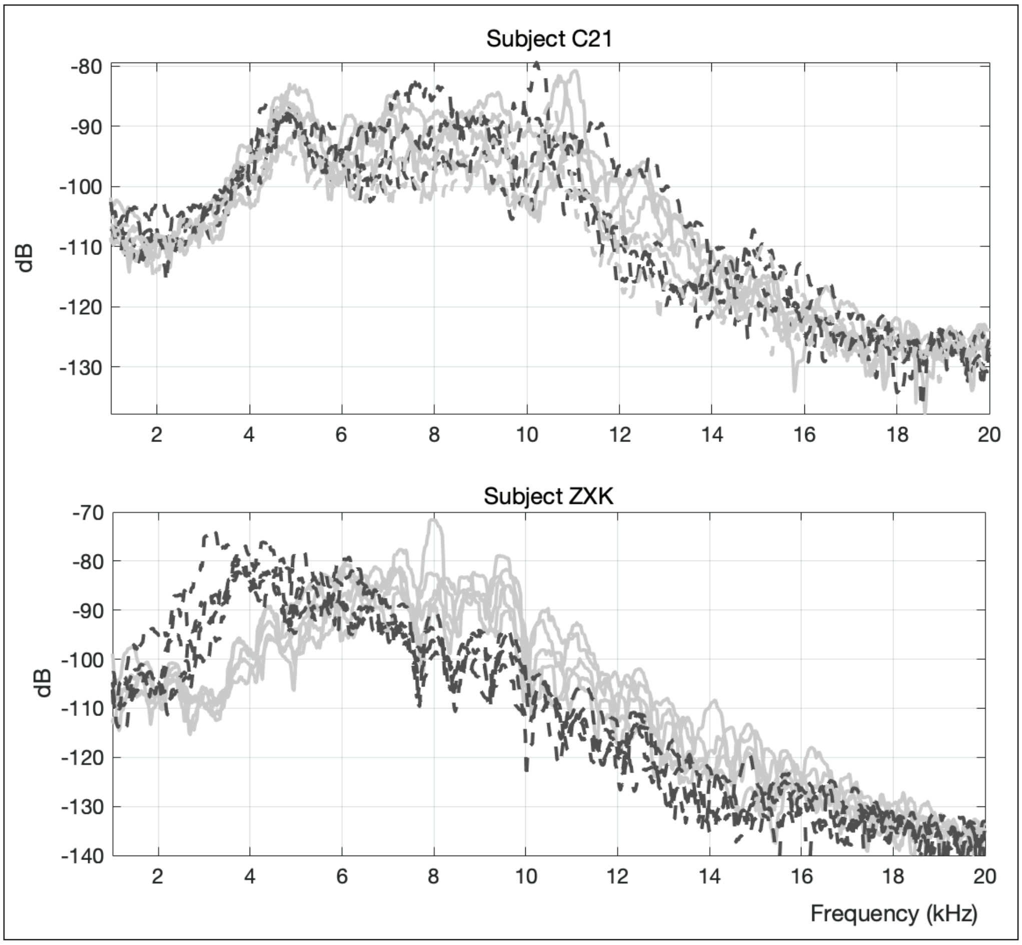

**Figure 3:** Example multitaper spectra of the sibilant fricatives /s/ (light grey) and /ʂ/ (dark grey) taken mid-phase of the frication noise from two male speakers C21 (MERGED) and ZXK (DISTINCT).

maintained statistically significant differences between the two categories, no matter how small the differences were.[5] The dotted vertical lines in **Figures 1** and **2** represent the dividing line between MERGED and DISTINCT speakers. Generally, the spectral distance for MERGED speakers was less than 1000 Hz.

[5] The categorical distinction between MERGED and DISTINCT speakers was obtained through statistical analyses. Among the DISTINCT speakers, in particular, there was a large variation ranging from speakers barely implementing spectral differences of 1,000 Hz to speakers implementing spectral differences of 4,000 Hz. The speakers who maintained small but reliable spectral differences between the two categories were likely to be the so-called 'near-mergers' (Labov, 1994). The spectral difference of 1,000 Hz in the high-frequency region is ostensibly small, e.g., 6,000 Hz (29.7 in ERB) and 7,000 Hz (30.8 in ERB), and is highly likely to be imperceptible, but, nonetheless, those speakers maintained a consistent and reliable spectral distance in their production of the two categories. Although a perception study is warranted to determine if these were cases of near-merger or full contrasts, it is beyond the scope of the present study. Here we simply aim to establish a conservative but reasonable boundary between the two groups.

Next, we examined the effects of the social and linguistic factors, gender and TSM-fluency, on spectral distance (Δ*M*). A linear regression model was performed in R v.4.0.3 (R Development Core Team, 2020) with the independent variable SPEC.DISTANCE (Hz) against the predictors including GENDER (2 levels contrast-coded: female = –1 versus male = 1), TSM (2 levels contrast-coded: TSM-weak = –1 versus TSM-fluent = 1), and their interaction. The results are summarized in **Table 3**. The model fit revealed that GENDER was a significant predictor ($\beta$ = –471.4, $p$ = .0002); specifically, female speakers (*M* = 2,285 Hz) maintained a significantly greater spectral distance between the two categories than male speakers (*M* = 1,381 Hz). TSM-fluency, however, did not usefully inform the spectral distance (*M*(TSM-fluent) = 1,783 Hz; *M*(TSM-weak) = 1,939 Hz) ($\beta$ = –135.7, $p$ = .2746), nor did its interaction with GENDER ($\beta$ = –133.8, $p$ = .2810).

| | *β* | *SE* | *T* | *P* |
|---|---|---|---|---|
| **(Intercept)** | **1813.9** | **123.5** | **14.687** | **<.0001** |
| **GENDER** | **–471.4** | **123.5** | **–3.817** | **.0002** |
| TSM | –135.7 | 123.5 | –1.098 | .2746 |
| GENDER*TSM | –133.8 | 123.5 | –1.083 | .2810 |

**Table 3:** Summary of the linear regression model predicting spectral distance (Hz). Formula: SPEC.DISTANCE ~ GENDER*TSM. Significant results are represented in bold text.

These results align well with the figures based on Bayesian tests showing that MERGED speakers were not limited to TSM-fluent bilinguals; in particular, male MERGED speakers were equally represented across the two groups (6 TSM-fluent, 6 TSM-weak). Apparently, sibilant merging is no longer a consequence of direct contact with TSM. TSM-weak speakers could barely speak the language (TSM-speaking score: 1.59/7, **Table 1**) and seldom heard it (TSM-listening score: 2.07/7) in their everyday lives. Despite past studies positing the association between TSM and the merger, the results of the present study add to the mounting evidence being compiled for the autonomy of the sibilant merger from TSM usage (Chuang et al., 2019).

However, the gender-dependent asymmetry was shown to hold true as described in the literature (Jeng, 2006; Tse, 1998) such that male speakers maintained smaller spectral distances between the two categories than female speakers as reflected in the results of the regression analysis (**Table 3**). This is not entirely surprising given that men have been shown to acquire socially stigmatized forms more readily than women (Eckert, 1989; Labov, 1990). This echoes the gender-ideologies in Taiwan reported by sociological studies—women are more frequently subject to the evaluation of their *qizhi (refinement)* (Su, 2008). As such, if a representative feature from TSM were to spread to non-native speakers, it would be adopted first by male speakers who are presumably under less pressure to avoid stigmatized features in their speech. The results are thus indicative of the complex interplay between gender and social prestige.

A closer look at the results, in fact, reveals another gender-dependent difference in the absolute spectral means of the sibilants. Some female speakers with clear contrasts, in particular, demonstrated spectral means as high as 10 kHz for the alveolar category, which indicates a more anterior location of articulation, namely dentalization. This might have to do with the stigmatization of strong retroflexion, which is a well-known characteristic of Mainland Mandarin spoken in China (Chang, 2017; Chang & Shih, 2015). Chung (2006, p. 210) has argued that "Full retroflexing is now the marked form; intermediate forms are the default, covert prestige forms for all groups of speakers" in contemporary TM. To avoid strong retroflexion while carrying clear contrasts, female speakers seem to have chosen to front the alveolar category. A similar case has been reported in the literature. Shih (2012) showed that middle-aged TSM-dominant female speakers tend to produce more anterior alveolar sibilants to compensate for the lack of the retroflexes in their speech. In both cases, dentalization seems to be favored by female speakers as a means to enhance the phonological contrast while appearing sufficiently 'refined.'

## 3. Experiment 2: Sibilant production in an interactive task

Having established the spectral characteristics of individual TM speakers' sibilant production, we investigated whether the merged categories could be unmerged in a particular experimental condition. To that end, an interactive task was administrated to the MERGED and DISTINCT speakers who fell on the extreme ends of the continuum. The latter group was included as a reference which was expected to show contrastive hyperarticulation in the presence of lexical competitors as shown in previous studies.

### 3.1. Participants

Twenty TM speakers ($M$(age) = 23.5) who had participated in the first experiment were invited to participate in the interactive task after as short as a few days or as long as 1.5 years of completing the reading task. The participants for this task were chosen from the two extreme ends of the continuum established in Experiment 1. Again, speakers who completely merged the sibilants are referred to as MERGED speakers, and those who maintained a sufficiently large spectral distance between the sibilants are referred to as DISTINCT speakers. There was no definitive objective criterion constituting 'sufficient' spectral distance; however, the spectral distance greater than approximately 2 kHz was deemed a reasonably large value for TM speakers, given the range of the spectral distance from 0 to around 4 kHz.[6] Crucially, the establishment of the two groups was motivated to assess similarities and differences in the way the sibilants are produced by speakers with varying degrees merging during the interactive task. **Table 4** summarizes the itemized numbers of the participants according to the known factors influencing the sibilant merger. TSM-fluency was balanced between the two groups, while the gender factor was slightly skewed as there were inherently more male MERGED speakers.

| | **MERGED** | **DISTINCT** | **Total** |
|---|---|---|---|
| TSM-fluent | 6 (2 F/4 M) | 5 (4 F/1 M) | 11 |
| TSM-weak | 4 (1 F/3 M) | 5 (3 F/2 M) | 9 |
| Total | 10 (3 F/7 M) | 10 (7 F/3 M) | 20 |

**Table 4:** The distribution of the participants by mergedness, TSM-fluency, and gender.

## 3.2. Stimuli

Eighteen minimal pairs contrasting in the initial sibilants (山腳 /**ʂ**an$^{55}$ tɕiaʊ$^{214}$/ 'hillside' versus 三角 /**s**an$^{55}$ tɕiaʊ$^{214}$/ 'triangle') were compiled for Experiment 2. The items were balanced for manner of articulation: six pairs each of fricative sibilants, unaspirated affricates, and aspirated affricates. The sibilants were followed by either unrounded homorganic approximants (e.g., [sɹ̪ ʂɻ], represented as /si ʂi/ below) (Lee-Kim, 2014) or the rhymes /an/ or /aŋ/. The retroflex items were the target words that the participants produced during the interactive task, and the corresponding alveolar items served as lexical competitors. The word frequencies were obtained through the Academia Sinica Balanced Corpus of Modern Chinese (http://asbc.iis.sinica.edu.tw/) and are summarized in Appendix B. A *t*-test run in R v.4.0.3 (R Development Core Team, 2020) confirmed null differences in the word frequency between the retroflex-targets ($M$ = 3.15) and the alveolar-competitors ($M$ = 3.00) ($t(33) = 0.4234$, $p = 0.6747$).

[6] While there seems to be no single, uncontroversial method for further dividing the DISTINCT speakers, Chang and Shih (2015) provided some insight for the present case. In their study comparing TM speakers with Mainland Mandarin speakers on sibilant production, perceptual judgments were first employed to screen out some TM speakers who did not convey clear distinctions between the two sibilants. Two out of ten speakers were excluded from the experiment for not contrasting the retroflexes more than 60% of the time. The remaining eight participants implemented a spectral distance of approximately 2 kHz (Figure 2 in Chang & Shih, 2015). Although this value was not established through a well-controlled perception study, it seems to be a reasonable boundary for marking relative perceptibility. Further, based on our own perceptual impressions, we confirmed that those classified as DISTINCT speakers made clear category distinctions. It is hoped that future studies explore this topic with a focus on perceptual consequences of spectral distance.

These eighteen retroflex target items were presented in two experimental conditions: Context and No Context. In the Context condition, target words were presented with a corresponding competitor (e.g., target: /ʂan$^{55}$ tɕiaʊ$^{214}$/ 山腳 'hillside' versus competitor: /san$^{55}$ tɕiaʊ$^{214}$/ 三角 'triangle'). A filler item beginning with a non-coronal sibilant initial was also presented along with the target and competitor words in this condition. In the No Context condition, target words were accompanied by two filler items but no lexical competitor. The eighteen minimal pairs were randomly divided into two halves balanced for manner of articulation. Two sets of the stimuli list were constructed based on this. For one set, one half was presented along with their competitors in the Context condition, and the other half was presented without their competitors in the No Context condition. For another set, the experimental conditions were counterbalanced across the two halves. One of the two sets was randomly assigned to the participants for the experiment.

In addition, another nine words beginning with retroflex sibilants were included as target words. Unlike the target words with minimal pair competitors in the Context condition, the potential competitors of these words were not existing Mandarin words (e.g., /ʂan$^{214}$ tuə$^{214}$/ 閃躲 'blink' versus */san$^{214}$ tuə$^{214}$/). The target words of this type were presented in the No Competitor condition where they appeared along with two filler items and were presented to all participants. The structure of the experimental design and relevant examples are presented in **Table 5**, and a full list of the target stimuli is presented in Appendix A.

| **Condition** | | **Target** | **Competitor/Filler** | **Filler** |
|---|---|---|---|---|
| with competitor | Context | /ʂan$^{55}$ tɕiaʊ$^{214}$/<br>山腳<br>'hillside' | /san$^{55}$ tɕiaʊ$^{214}$/<br>三角<br>'triangular' | /liɛ$^{51}$ wu$^{51}$/<br>獵物<br>'prey' |
| | No Context | /ʂan$^{55}$ tɕiaʊ$^{214}$/<br>山腳<br>'hillside' | /piŋ$^{55}$ xə$^{35}$ /<br>冰河<br>'iceberg' | /liɛ$^{51}$ wu$^{51}$/<br>獵物<br>'prey' |
| without competitor | No Competitor | /ʂan$^{214}$ tuə$^{214}$/<br>閃躲<br>'dodge' | /tɕy$^{55}$ kəŋ$^{55}$/<br>鞠躬<br>'bow' | /liəʊ$^{35}$ jɛn$^{35}$/<br>留言<br>'leaving a message' |

**Table 5:** Examples of the target retroflex sibilants.

In addition to the target items, eighteen control items beginning with the alveolar sibilants were added to the stimuli list, balanced by manner of articulation, i.e., fricative sibilants, unaspirated affricates, and aspirated affricates. The sibilants were followed by either the unrounded vowel /a/ or the homorganic approximants. These words lacked corresponding minimal pair competitors (e.g., 死命 /si$^{214}$ miŋ$^{51}$/ 'doom', */ʂi$^{214}$ miŋ$^{51}$/), similar to the target words in the No Competitor condition. These control items provided the baseline productions for the alveolar category to assess the spectral distance between alveolar and retroflex sibilants drawn from the interactive task. The log frequencies of the stimuli words in different experimental conditions are summarized in Appendix B. A regression model was fit in R v.4.0.3 (R Development Core Team, 2020) with the log frequency as the dependent variable and the experimental condition as the predictor with the Context items as the baseline. The results confirmed that the word frequency of the Context items ($M = 3.15$) was not significantly different from that of the No Competitor items ($M = 3.42$) ($\beta = 0.2700$, $t = 0.745$, $p = 0.460$), nor from that of the Control items ($M = 3.62$) ($\beta = 0.4706$, $t = 1.590$, $p = 0.119$).

## 3.3. Procedure

During the experiment, the participant and the experimenter sat face-to-face at a table, each with a separate laptop computer. The experimenter (listener) told the participant (speaker) that they would be playing a language game together.[7] Upon seeing three words on the screen, the participant would read the highlighted word out loud (**Figure 4**). Both parties would see the exact same three words, but the target word would not be highlighted on the experimenter's screen. The participant's task was to read out the target words highlighted for the experimenter who would click on the word on her screen. Instructions for the procedure were given to the participants with some examples, and five practice trials were completed prior to starting the experiment. The participants wore an AKG C520L head-mounted microphone connected to a Zoom H4n recorder, and their production was recorded throughout the experiments. The experiment was carried out in the Production and Perception lab at National Yang Ming Chiao Tung University. It took approximately twenty minutes for participants to complete each block, and they were given a five-minute break between blocks.

A total of 150 stimuli (9 retroflexes [Context] + 9 retroflexes [No Context] + 9 retroflexes [No Competitor] + 18 alveolars [Control] + 105 fillers) was repeated three times across three blocks. Note that there were nearly three times as many filler items than target and control

[7] The experimenter, the second author of this article, is a TSM-weak young female speaker carrying clear contrasts between the two categories. Although we are aware of potential speech accommodation with the experimenter (e.g., Hay, Drager, & Warren, 2009), it is not clear to what extent the unmerging of the sibilants could be attributed to the approximation to the experimenter. Baese-Berk and Goldrick (2009) observed contrastive hyperarticulation of English voiceless stops in a single-word reading task, as well as in an interactive task, suggesting that the interaction with an interlocutor is not essential for contrastive hyperarticulation. Furthermore, as part of a bigger research project, the recordings of the read speech in Experiment 1 were made by other experimenters who, again, carried fairly clear distinctions of the contrast. It should be noted that many TM speakers, nonetheless, showed a high degree of sibilant merging in that study.

items to mask the purpose of the experiment (Baese-Berk & Goldrick, 2009). In each block, the target words appeared in the first, middle, and last position on the screen randomly. All trials were randomized within each block. The stimuli were presented in Chinese characters to the participants using Microsoft PowerPoint. A total of 7,000 ms was allotted for each trial. A fixation cross appeared on the screen for 1,000 ms, followed by all three words presented horizontally for 2,000 ms; a purple square box appeared around one of the three words on the participants' screen (**Figure 4**). Participants were given 4,000 ms to name the target. They were told that the response time would be limited so they needed to produce the target as quickly as possible. As soon as the participants produced the words, the experimenter clicked on the corresponding word on her screen. The experimenter's performance was not recorded, but the interactive nature of the task was expected to encourage the participants to produce the target words in a careful manner.

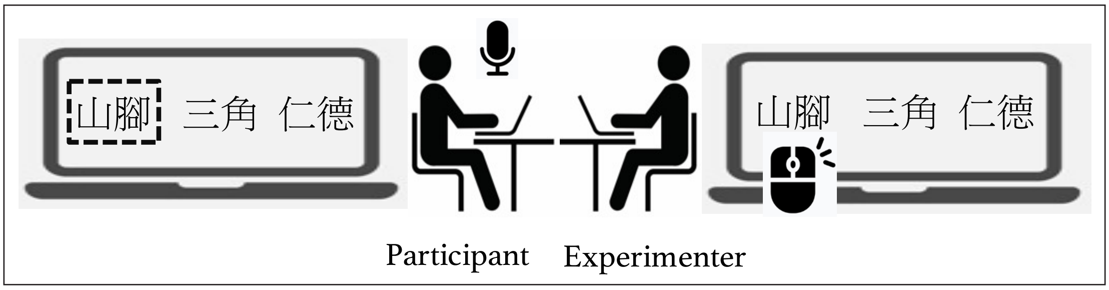

**Figure 4:** The experimental setting for the interactive task.

After removing 19 poorly-recorded tokens, a total of 2,816 tokens (99.3%) was collected for analysis. The sound files were first labeled automatically using Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), and the boundaries were corrected manually. As in Experiment 1, the first spectral moment (M1) of the frication noise was computed using the multitaper spectral analysis in Matlab (Blacklock, 2004; Blacklock & Shadle, 2003).

To examine the effect of phonological neighbors on the production of the target sibilants, a mixed-effects regression model was fitted using the *lmer* function in the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) in R v.4.0.3 (R Development Core Team, 2020). The dependent variable was SPEC.MEAN (Hz). Fixed effects included GROUP (2 levels: DISTINCT versus MERGED) and CONDITION (4 levels: Control versus Context versus No Context versus No Competitor). GROUP(MERGED) and CONDITION(NoCont) were set as the baseline for analyses to directly compare these conditions with the other conditions for the target items. Apart from the main effects, a two-way interaction between CONDITION and GROUP was also included in the model to examine whether the two groups performed differently for the experimental conditions. In addition, the model included SUBJECT as a random intercept.

## 3.4. Results

**Figure 5** plots the distribution of spectral means (Hz) of the sibilants as a function of different conditions in Experiment 2 for the two groups, MERGED (left) versus DISTINCT (right). The results of the read speech for those participants are summarized below as well for comparison. Notably, the MERGED group made clear distinctions between the categories in the interactive setting, even though they had merged the two categories completely in the reading task.

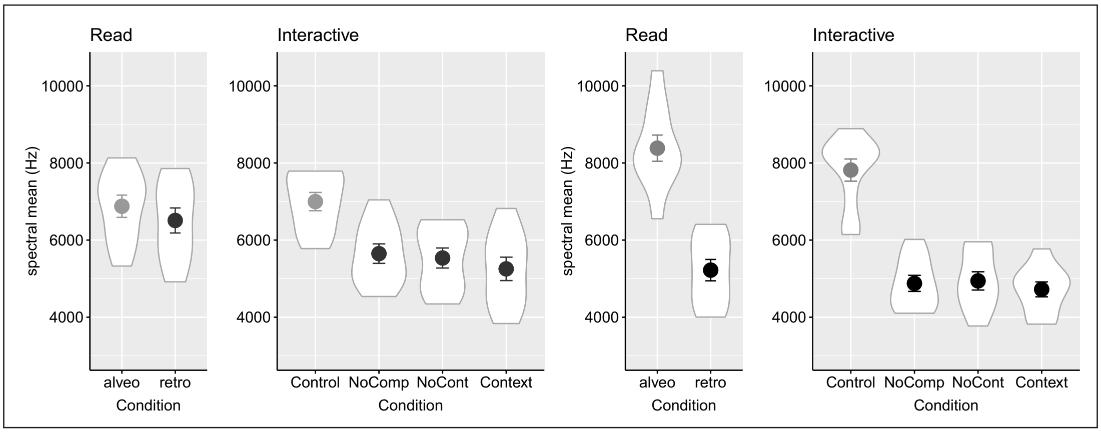

**Figure 5:** Spectral means of the alveolars (light grey) and retroflex (dark grey) sibilants by MERGED (left) and DISTINCT (right) speakers in the read versus interactive tasks, respectively. Error bars represent one standard error.

**Table 6** summarizes the results of the regression model fit. For the MERGED speakers, the results revealed that the spectral mean of the retroflex sibilants in the baseline No Context condition ($M$ = 5,534 Hz) was significantly lower than that of the control alveolar sibilants ($M$ = 6,996 Hz) ($p$ =.0001). The differences between the two categories are substantial ($\Delta M$ = 1,461 Hz) especially given that they had completely merged them in the reading task, suggesting that these MERGED speakers indeed have clear mental representations of the two discrete categories which can be realized with fully distinct articulation in a particular experimental setting. Although the spectral means in this particular condition appear to be slightly higher than those produced by DISTINCt speakers in the same condition ($M$ = 4,947 Hz), the difference was marginal, as shown in the main effect of GROUP(DISTINCT) ($p$ = .0681). This result, once again, indicates that while MERGED speakers may habitually deretroflex the retroflex category, the merger does not arise from articulatory limitations. That is, merged speakers can produce the retroflexes when necessary and to a similar degree as the DISTINCT speakers.

The spectral means of the retroflexes produced in the No Context ($M$ = 5,534 Hz) and No Competitor conditions ($M$ = 5,649 Hz) were comparable, and the small difference between them did not reach statistical significance ($p$ = .1684). However, the spectral means of the retroflexes in the Context condition ($M$ = 5,247 Hz) were significantly lower than those in the baseline No Context condition ($p < .0005$), demonstrating the impact of the minimal pair competitors during online processing. The absolute mean difference between the two conditions was ostensibly small, i.e., 286 Hz, which is unlikely to be audible; however, the strong effect of this variable in the mixed-effects modeling suggests that speakers made small but consistent differences in response to the presence of the lexical competitors during sibilant production. A three-way distinction among the experimental conditions was not obtained in the present study, unlike the findings in Baese-Berk and Goldrick (2009), an issue that will be addressed in the discussion below.

In addition to the main effects, the model also revealed a significant two-way interaction for CONDITION(alveolar):GROUP(DISTINCT) ($p < .0001$). This can be attributed to the spectral mean distance for the DISTINCT group (2,874 Hz, $M$(Alveolar) = 7,815 Hz – $M$(NoCont) = 4,947 Hz) being nearly twice that of the MERGED group (1,461 Hz, $M$(Alveolar) = 6,996 Hz – $M$(NoCont) = 5,534 Hz). As evident in the figure, this interaction arises jointly from both greater dentalization of the alveolar sibilants and greater retroflexion of the retroflex sibilants implemented by the speakers in the DISTINCT group. This shows that DISTINCT speakers not only conveyed clear retroflexes but also dentalized the alveolar sibilants, leading to a better contrast for the phonemic distinctions.

Other interactions were shown to be insignificant. In particular, the lack of the significant CONDITION(Context):GROUP(DISTINCT) interaction ($p = .6150$) suggests that the effect of the lexical competitors was comparable across the two groups. In order to verify whether this effect was independently present for each of the DISTINCT speakers, a separate analysis with the DISTINCT group as the baseline was performed with the identical model structure. The results of the model fit showed that the predictor CONDITION(No Competitor) was not significant ($\beta = –68.29$, $t = –0.829$, $p = .4070$) but CONDITION(Context) was significant ($\beta = –227.54$, $t = –2.751$, $p = .0060$), which corresponds with the results of the model with the MERGED group as the baseline. Together, the results demonstrate that both groups were sensitive to the presence of the minimal pair competitors, which led to the enhancement of the phonetic distance between the phonologically contrastive sounds.

As the effect of explicit phonological neighbors was ostensibly small, we also examined individual data to ensure the context effects were independently present for each participant. **Figure 6** presents some representative cases of the merged speakers.[8] Not surprisingly, large individual variation is clearly evident in the figure; some speakers showed smaller within-category variation (e.g., CSY), while others showed much larger dispersion across all conditions (e.g., XSJ). The degree of unmerging differed across the MERGED speakers as well; those presented

[8] The figures of each and every individual speaker from both MERGED and DISTINCT groups can be accessed at https://osf.io/4r97f/.

| **Predictor** | ***β*** | ***SE*** | ***t*** | ***p*** |
|---|---|---|---|---|
| **(Intercept)** | **5534.96** | **216.18** | **25.603** | **<.0001** |
| **CONDITION(control)** | **1459.20** | **71.49** | **20.411** | **<.0001** |
| CONDITION(No Competitor) | 113.72 | 82.54 | 1.378 | 0.1684 |
| **CONDITION(Context)** | **–286.39** | **82.78** | **–3.460** | **0.0005** |
| GROUP(DISTINCT) | –589.22 | 305.75 | –1.927 | 0.0681 |
| **CONDITION(alveo):GROUP(DISTINCT)** | **1427.19** | **101.31** | **14.087** | **<.0001** |
| CONDITION(No Competitor):-GROUP(DISTINCT) | –182.01 | 116.59 | –1.561 | 0.1186 |
| CONDITION(Context):GROUP(DISTINCT) | 58.86 | 117.01 | 0.503 | 0.6150 |

**Table 6:** Summary of the mixed-effects regression model. Formula: SPEC.MEAN ~ CONDITION * GROUP + (1|SUBJECT). Significant results are shown in bold.
The baselines values are set as follows: Group = MERGED and Condition = No Context.

in the top row show a larger degree of unmerging than those in the bottom row. Regardless of idiosyncratic variation, however, spectral means were consistently lower in the Context condition, albeit to a small degree, for most speakers. Notably, one speaker, E63, did not show any difference between the first three comparisons: the alveolar controls and the retroflexes in the No Competitor and No Context conditions. However, this speaker still demonstrated lower spectral means for the retroflexes in the Context condition. The speaker predominantly carried deretroflexed sibilants but made an effort to produce retroflexes when lexical disambiguation was necessary. Despite individual variation, most speakers showed a small but consistent effect of phonological competitors on the production of the sibilants, which accounts for the strong effect of CONDITION in the model fit.

# 4. General discussion

## 4.1. Unmerging of the sibilant merger via contrastive hyperarticulation

The results of the interactive task revealed that the MERGED speakers increased the spectral distance considerably between the two sibilant categories. In fact, they were able to reverse the deretroflexion and produce reasonably good retroflexes, although the degree of retroflexion was slightly smaller than that of the DISTINCT speakers. This finding provides evidence in favor of discrete representations of the two categories in the mental lexicon of the speakers who merged the sibilants. These speakers, however, seldom produced retroflexes in their speech, as shown

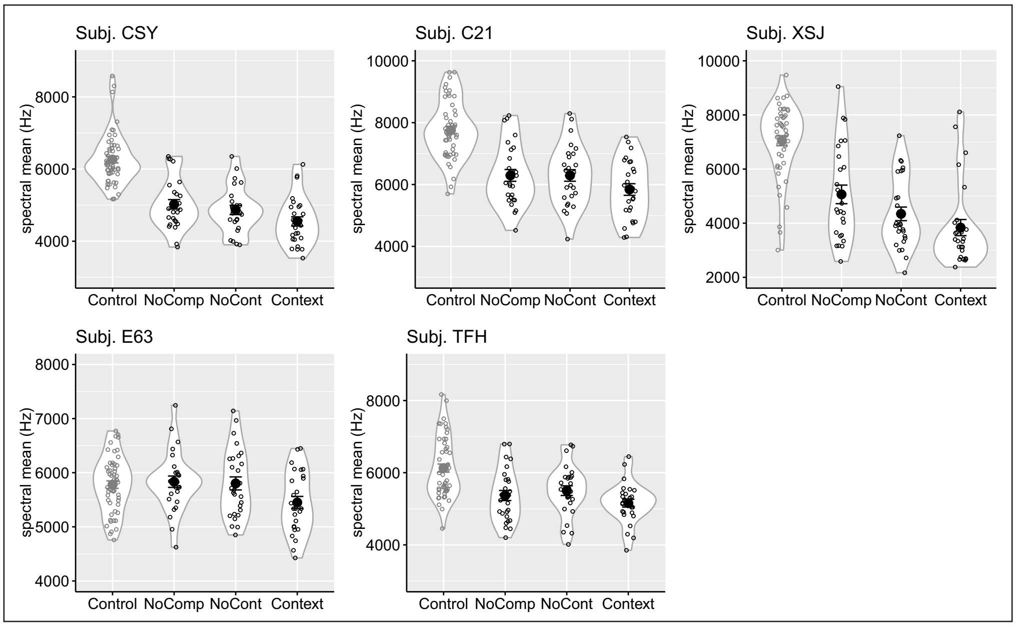

**Figure 6:** The distribution of spectral means by condition in the interactive task for a subset of the MERGED speakers. The filled circles represent the means, and error bars represent one standard error of the spectral means.

in their performance in Experiment 1. Recall that the reading task was designed to ensure clear speech, especially given the particular structure of the frame sentence, in which the target words were produced under narrow focus. Yet these speakers completely merged the sibilants in this formal register. When the experimental task encouraged interaction with another speaker, the MERGED speakers, however, revealed that the distinct categories were indeed stored in their mental lexicon.

The results of this study are, on one hand, in line with previous studies of contrastive hyperarticulation (Baese-Berk & Goldrick, 2009; Wedel et al., 2018). Both MERGED and DISTINCT speakers responded to the experimental condition as predicted—all the TM speakers enhanced the spectral distance in the Context condition significantly more than in the No Context condition. That is, retroflexion was stronger when the target retroflexes were produced in the presence of the alveolar counterparts, compared to when no such competitors were presented. This result adds to the growing body of literature concerning the cognitive mechanisms of speech production (Baese-Berk & Goldrick, 2009; Fricke et al., 2016; Kirov & Wilson, 2012). Based on VOT enhancement in the presence of lexical competitors, Baese-Berk and Goldrick (2009) argued that word-specific phonetic variation is driven primarily by *online* processing in which a target is triggered by the activation of an 'explicit' competitor (the Context condition). The results of the present study contradict an alternative account which postulates that speakers' production systems are permanently restructured to hyperarticulate words in dense neighborhoods. Such a model would predict no differences between the Context and No Context conditions, which was again not the case in the current study. Additionally, this particular experiment, with its use of logographic Chinese characters, highlights that the general competitor effects are not driven by visual cues available in alphabetic orthographies (e.g., **p**at versus **b**at), rather its origin is rooted in more abstract phonological knowledge.

On the other hand, the results of the present study were not entirely consistent with previous studies. Baese-Berk and Goldrick (2009) found a three-way distinction in VOT for English aspirated stops between the three conditions: shortest VOT in the No Competitor condition, intermediate in the No Context condition, and longest VOT in the Context condition. However, only a two-way distinction was observed in the present study due to the lack of significant differences between the No Competitor and No Context conditions for both groups of TM speakers. Also note that Wedel et al. (2018) reported differences in VOT enhancement between words with and without competitors produced in natural conversation. Despite some differences, the conditions in Wedel et al. (2018) roughly correspond to the No Context and No Competitor conditions in the present study, respectively. Yet English speakers enhanced VOT to a greater degree when target words had minimal pair competitors.

This discrepancy may be attributed to the overall enhancement of the phonetic space specific to the sibilant place contrasts, i.e., spectral means (Hz), during the interactive task. In particular, the retroflexes in the No Competitor condition were fully retroflexed, as well as other conditions, the source of which, apparently, was not the lexical competition. Rather, it is likely that the social interaction with the experimenter may have encouraged speakers to avoid the deretroflexed forms, irrespective of the lexical properties of the words. Recall that deretroflexion in TM is fairly stigmatized and used to be associated with lower levels of education and economic status (Chung, 2006; Hsiau, 1997; Su, 2008). This was partially reflected in the results of the first experiment; namely, female speakers were found to merge the sibilants less than the male speakers. Given that the magnitude of lexically-driven hyperarticulation was ostensibly small, the socially-driven enhancement of phonetic space seems to have neutralized some of the word-specific properties. In previous studies of English stop VOT, no stigma is associated with voicing/devoicing, and therefore the three-way distinction in VOT between the three conditions could be driven solely by the lexical status of the words. With the introduction of social pressures to complicate matters, the effects of lexical factors may only manifest to a limited degree.

Lastly, we address the seemingly conflicting results of our work and those of previous sociophonetic studies (Johnson & Nycz, 2015; Nycz, 2013). As discussed earlier, native speakers of Canadian English merged the vowels in the minimal pair tasks but unmerged them in spontaneous conversation to approximate the phonetic characteristics of the dialect spoken in their new speech community. Conversely, the TM speakers in our study unmerged the sibilants in the task utilizing minimal pairs. There are, however, many fundamental differences between the previous studies and the current one. For one, in terms of the nature of the task, our study utilized minimal pair competitors in an interactive linguistic setting. This methodology was motivated to see if the speakers who merged the contrast would alter their production for the sake of the interlocutor. This differs from the above-mentioned studies in which the minimal pair task was meant to induce the speakers' knowledge of the 'correct' phonetic norms by explicitly asking whether they were aware of the phonetic differences between the pairs of the words. In the present case, lexical disambiguation implicitly capitalized on the speaker's desire to facilitate communication, without the intention to tap into the speaker's phonetic norms.

The present case also differs from the English vowel mergers in that TM speakers are astutely aware of the phonetic characteristics of the contrasting sounds and the lingering social stigmatization attached to the deretroflexed forms. In Nycz's (2013) judgment task, the speakers' knowledge of the vowel contrast was marginal, at best, and the merged forms did not carry a strong stigma. In an experimental task designed to specifically induce the explicit knowledge of the contrasts, TM speakers are likely to give the 'proper' pronunciations consistent with the prescriptive grammar explicitly taught in school, trying their best to unmerge the two categories. This critically differs from the unmerging of the low back vowels by Canadian English speakers residing in NYC, which served to express their accommodation of the phonetic characteristics of the sounds in their new speech community. Bearing in mind the similarities and differences between these studies, the interactive task presented here could be utilized to inform the structure of abstract linguistic knowledge and could be a useful addition to existing tools for sociolinguistic research.

### 4.2. Representations and implementation of the sibilants in the merger-in-progress

The results of the interactive task provided evidence in favor of the clear mental representations of the contrastive categories being stored in the mental lexicon of the MERGED speakers. What is the source of the apparent mismatch between the phonological knowledge of the merged speakers and the actual implementations of the sound categories in speech production? Why did these speakers merge the sounds if they were represented separately in their minds? More fundamentally, where does this linguistic knowledge, i.e., sound contrasts, originate? Will the variable merger progress into a more stable sound change throughout the community?

Experiment 1 provided some diagnostic means regarding the status of this particular merger. Taiwanese scholars have maintained the deeply-rooted belief that the sibilant merger arose from language contact with local substratum languages that do not have the retroflex category (Ing, 1984; Kubler, 1985). However, the present study has shown that sibilant merging has become, by and large, independent of TSM fluency as TSM-weak TM speakers merged the category as frequently as TSM-fluent bilingual speakers. This indicates that the merger is widespread among the younger generation in Taiwan, and weak retroflexion is becoming a characteristic phonetic feature of TM, departing from Mainland Mandarin. This echoes recent sociolinguistic studies that have argued that the stigma associated with TSM is dramatically declining among young TM speakers and a supra-ethnic and cross-linguistic Taiwanese identity is being formed (Hsiau, 1997; Huang, 2019; Tse, 2000). In this dynamic socio-historical context, the social meaning associated with deretroflexion may be changing—no longer is it necessarily a derogatory feature of a substratum language; it is instead becoming a feature of a unique phonetic variant that enables speakers to express a positive orientation toward the Taiwanese language. A future study is warranted to directly address the connection between social attitudes and linguistic performance of TM speakers.

Yet, despite its recent rise, TSM is a non-standard language that cannot claim as much prestige as TM in Taiwan (Sandel, 2003; Tse, 2000). The lingering stigma or negative social pressure seems to have deterred female speakers from incorporating the once-substrate feature into their speech. Formal school education and conservative cultural practices in Taiwan still impose the importance of the standard norms. While sound change led by men is less common and often limited to relatively isolated patterns (Labov, 1990; Labov, 1994), with the mixed signals co-existing in the society, it is not surprising that female speakers lag behind for this particular change.

The mixed social connotations attached to the merger seem to have brought about a fully variable and gradient merger pattern as demonstrated in the results of the reading task. An immediate consequence of this full continuum is that speakers of the linguistic community are exposed to drastically different grammars; speakers who merge sibilants encounter unmerged patterns, and speakers who contrast the category encounter merged patterns. In particular, the exposure to different sound systems may have prevented the complete merging of the categories in the mental representations, which may have been amplified by some lingering stigma (Clark et al., 2013). As proposed by the exemplar-based theories of speech perception and production, phonological representations may not be fully discrete and categorical (Johnson, 1997; Pierrehumbert, 2001). Rather, they form clouds of exemplars, which is constantly updated as speakers encounter various forms of the same category from other speakers. With many distinct speakers in the speech community, the exemplar cloud of the retroflex category would expand with the addition of somewhat outlying fully retroflexed sounds, which could increase the dispersion between the two categories. Of course, the reverse process may happen to the speakers who contrast the category, creating a fuzzier boundary between the categories. Based on the results of the present study, however, it is not clear whether those who merge sibilants have less clear category boundaries than those who do not. While it is a likely scenario, we believe it could be more adequately addressed by future comprehensive empirical research covering production, perception, and psycholinguistic processing.

Apparently, the sibilant merger in contemporary TM is variable and could be best characterized as a merger-in-progress. The outlook—whether this pattern will develop into a full-grown sound change—is unclear. At the onset of sound change, considerable phonetic variation is observed among the individuals in a speech community. Sociolinguists have argued that social factors, as well as grammar-internal factors, modulate whether certain variations eventually develop into systematic sound changes. Labov (1963, 1990) famously identified social motivations in sound change such that a particular phonetic variable may gain some social meaning and trigger imitation by other speakers. Sound changes result from the regularization of the advanced forms that spread among the speech community.

In light of this, we should consider another merger-in-progress of an entirely different sound category in TM. TSM lacks certain nucleus and nasal coda sequences, e.g., /iŋ/ and /əŋ/, which are often replaced with /in/ and /ən/ by TSM-dominant speakers (Kubler, 1985; Tse, 1998). Just like the stigma associated with detroflexion, these TSM features were once stigmatized, conditioned by age, gender, and socioeconomic status (Kubler, 1985; Tse, 1992). However, the stigmatization of the nasal merger has declined dramatically over the years, and the nasal merger has emerged, along with other nasal merger patterns, as a common feature of TM, as verified by instrumental studies (Chiu & Lu, 2020; Fon, Hung, Huang, & Hsu, 2011; Hsu & Tse, 2007). Compared to this relatively mature sound change, the sibilant merger seems to still be in its early stages of development. Anecdotally, TM speakers often say that they cannot recover the correct *Zhuyin* symbols of /n/ and /ŋ/ when typing electronic documents and mostly guess one of the two, which is rarely the case for the /s/ and /ʂ/ contrast. The apparent asymmetry in the sound change with similar historical origins seems to lie in the relative saliency of the sounds involved. Compared to final nasal places, the sibilant place contrasts in the initial position are more salient acoustically and perceptually, which would have slowed down the progression of the merger among the speakers in the community. If the stigma continues to diminish and this merger gains some positive social meaning, e.g., Taiwanese identity, it would be feasible that the sibilant merger would continue to develop into a mature sound change.

## 5. Conclusion

Variation, as an essential aspect of speech sounds, provides a window into the linguistic architecture connecting abstract mental representations stored in speakers' mental lexicon and the way they are implemented in production. The formal reading task showed that the TM sibilant merger exists on a full continuum, from a complete merger to clear contrasts, and is more prevalent among male speakers, demonstrating the impact of the social stigma associated with the merger. Moreover, though rooted in historical contact with TSM, the sibilant merger is becoming independent of TSM. Regardless of whether speakers merged the sibilants or maintained the contrast in the reading task, the TM speakers all made a clear distinction between the two categories in the interactive task, indicating that they had non-overlapping discrete representations of the contrasting sounds. The apparent dichotomy between what they have stored in the mental lexicon and what they implement in production suggests the role of variation at the onset of sound change. Speakers are exposed to the phonological systems of others, especially those retaining the contrasts, which, along with the social stigma, may have prevented the complete merging of the categories in their mental representations. This case study provides some insight into the apparent paradox of near-mergers, that speakers cannot perceive certain distinctions even though they maintain small but consistent differences in their production. Variation framed in social and linguistic dynamics may provide substantial evidence for possible categories in a given language for language learners; however, their implementation may be further modulated by social factors as well as grammar-internal or phonetic factors.

## Additional files

The additional files for this article can be found as follows:

- **Appendix A.** Stimuli list for Experiment 2. DOI: https://doi.org/10.16995/labphon.6446.s1
- **Appendix B.** Log frequencies of the minimal pair words and of the stimuli by experimental condition. DOI: https://doi.org/10.16995/labphon.6446.s2

## Acknowledgements

We would like to thank Waan-rur Lu, Hsiang-Yu Tung, and Wei-Cheng Weng for help with data collection and processing, and Yu-an Lu for help with the forced aligner. We are grateful to Chenhao Chiu, Edward Flemming, Carlos Gussenhoven, and Kristine Yu for their valuable comments and suggestions at various stages of this project and Hannah Parsamehr for her careful editing of the manuscript. We also thank audiences at the NYCU Phon Brown Bag, National Conference on Linguistics, Keio-ICU Linguistics Colloquium series, and Academia Sinica Phonetics Laboratory. Our deep gratitude also goes to the editor Abby Walker and anonymous reviewers of Laboratory Phonology for their constructive comments throughout different stages of the manuscript, which led to substantial improvements to this work.

## Funding information

This work was supported by research grants MOST107-2410-H-009-016-MY3 and MOST110-2410-H-A49-058-MY3 from the Ministry of Science and Technology, Taiwan.

## Competing interests

The authors have no competing interests to declare.

## Author contributions

Sang-Im Lee-Kim is the principal investigator of the project "Phonological representations and social factors in sound changes in progress" (MOST110-2410-H-A49-058-MY3) and initiated and supervised this study. She was responsible for conceptualization, methodology, acoustic analysis, data curation, formal analysis, writing of the original draft, and review and editing.

Yun-Chieh Chou was an MA student at National Chiao Tung University and was a research assistant for the above project. She performed stimuli construction, data collection, labeling of the recorded data, formal analysis, and writing of the original draft. Part of the data collected for this project was used in her MA thesis.

## References

Ang, U. (1997). Taiwan gonggong changsuo shiyong yuyan diaocha [A survey on language use in public occasions]. In C. Tung (Ed.), *Taiwan yuyan fazhan xueshu yantaohui lunwenji [Proceedings of the Conference on Language Development in Taiwan]* (pp. 83–100). Quanmin Bookstore.

Ang, U. (2010). Taiwan diqu de yuyan fenbu [The language distribution in Taiwan area]. In X. Zhuang & X. Wang (Eds.), *Qiaokai Yuyan de Chuangkou: Huayu de Shiyong Xianxiang (Qingkuang yu Diaocha) [Knock Open the Window of Language: The Usage of Mandarin (Pattern and Investigation)]* (pp. 1–39). United Publishing House.

Bååth, R. (2014). Bayesian First Aid: A Package that Implements Bayesian Alternatives to the Classical *.test Functions in R. UseR! 2014 – the International R User Conference.

Babel, M., McAuliffe, M., & Haber, G. (2013). Can mergers-in-progress be unmerged in speech accommodation? *Frontiers in Psychology*, Article 653. DOI: https://doi.org/10.3389/fpsyg.2013.00653

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes*, *24*(4), 527–554. DOI: https://doi.org/10.1080/01690960802299378

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

Blacklock, O. S. B. (2004). *Characteristics of variation in production of normal and disordered fricatives using reduced-variance spectral methods.* 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. DOI: 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.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86. DOI: https://doi.org/10.1016/j.jml.2015.12.009

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.* University of Illinois, Urbana.

Chang, Y.-H. S. (2017). The influence of dialect information on the perception of the Mandarin alveolar-retroflex contrast. *Concentric: Studies in Linguistics*, *43*(1), 1–23.

Chang, Y. H. S., & Shih, C. (2015). Place contrast enhancement: The case of the alveolar and retroflex sibilant production in two dialects of Mandarin. *Journal of Phonetics*, *50*, 52–66. DOI: https://doi.org/10.1016/j.wocn.2015.02.001

Chen, P. (1999). *Modern Chinese: History and Sociolinguistics*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139164375

Chiu, C., & Lu, Y.-A. (2020). Articulatory evidence for the syllable-final nasal merging in Taiwan Mandarin. *Language and Speech*. DOI: https://doi.org/10.1177/0023830920948084

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*. DOI: https://doi.org/10.1177/0023830919896386

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 ICPhS XVIIII, Melbourne, Australia. DOI: 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. DOI: https://doi.org/10.1075/japc.16.2.04chu

Clark, L., Watson, K., & Maguire, W. (2013). Introduction: What are mergers and can they be reversed? *English Language and Linguistics*, *17*, 229–239. DOI: https://doi.org/10.1017/S1360674313000014

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change*, *1*, 245–268. DOI: https://doi.org/10.1017/S095439450000017X

Feifel, K.-E. (1994). *Language attitudes in Taiwan: A social evaluation of language in social change*. The Crane publishing.

Fon, J., Hung, J.-m., Huang, Y.-H., & Hsu, H.-j. (2011). Dialectal variations on syllable-final nasal mergers in Taiwan Mandarin. *Language and Linguistics*, *12*(2), 273–311.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, *84*, 115–123. DOI: https://doi.org/10.1121/1.396977

Fricke, M., Baese-Berk, M. M., & Goldrick, M. (2016). Dimensions of similarity in the mental lexicon. *Language, Cognition and Neuroscience*, *31*(5), 639–645. DOI: https://doi.org/10.1080/23273798.2015.1130234

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

Garde, P. (1962). Réflexions sur les différences phonétiques entre les langues slaves. *Word*, *17*, 34–62. DOI: https://doi.org/10.1080/00437956.1961.11659746

Harris, J. (1985). *Phonological variation and change: Studies in Hiberno-English*. Cambridge University Press.

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

Hay, J., Drager, K., & Warren, P. (2009). Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics*, *29*(2), 269–285. DOI: https://doi.org/10.1080/07268600902823128

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

Hsiau, A.-C. (1997). Language ideology in Taiwan: The KMT's language policy, the Tai-yu language movement, and ethnic politics. *Journal of Multilingual and Multicultural Development*, *18*(4), 302–315. DOI: https://doi.org/10.1080/01434639708666322

Hsu, H.-J., & Tse, J. K.-P. (2007). Syllable-final nasal mergers in Taiwan Mandarin—Leveled but puzzling. *Concentric: Studies in Linguistics*, *33*(1), 1–18.

Huang, K. (2019). Language ideologies of the transcription system Zhuyin fuhao: A symbol of Taiwanese identity. *Writing Systems Research*, *11*(2), 159–175. DOI: https://doi.org/10.1080/17586801.2020.1779903

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

Ing, R. O. (1984). Issues on the pronunciations of Mandarin. *The World of Chinese Language*, *35*, 6–16.

Jeng, J.-Y. (2006). The acoustic spectral characteristics of retroflex fricatives and affricates in Taiwan Mandarin. *Journal of Humanistic Studies*, *40*(1), 27–48.

Johnson, D. E., & Nycz, J. (2015). Partial mergers and near-distinctions: Stylistic layering in dialect acquisition. *University of Pennsylvania Working Papers in Linguistics, 21*(2), Article 13.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165).

Kirov, C., & Wilson, C. (2012). *The specificity of online variation in speech production.* 34th annual meeting of the cognitive science society, Sapporo, Japan.

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1175–1189. DOI: https://doi.org/10.1044/1092-4388(2012/12-0038

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

Labov, W. (1963). The social motivation of a sound change. *Word*, *19*(3), 273–309. DOI: https://doi.org/10.1080/00437956.1963.11659799

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, *2*, 205–254. DOI: https://doi.org/10.1017/S0954394500000338

Labov, W. (1994). *Principles of linguistic change, vol. 1: Internal factors*. Blackwell.

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

Labov, W., Yaeger, M., & Steiner, R. (1972). A quantitative study of sound change in progress. In *U.S. Regional Survey*. Philadelphia.

Lee-Kim, S.-I. (2011). Spectral analysis of Mandarin Chinese sibilant fricatives. ICPhS XVII, Hong Kong.

Lee-Kim, S.-I. (2014). Revisiting Mandarin 'apical vowels': An articulatory and acoustic study. *Journal of the International Phonetic Association*, *44*(3), 261–282. DOI: https://doi.org/10.1017/S0025100314000267

Lee-Kim, S. I., Kawahara, S., & Lee, S. J. (2014). The 'Whistled' fricative in Xitsonga: Its articulation and acoustics. *Phonetica*, *71*(1), 50–81. DOI: https://doi.org/10.1159/000362672

Lin, Y.-H. (1988). Consonant variation in Taiwan Mandarin. Language Change and Contact NWAV XVI, Austin: University of Texas.

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

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. Interspeech. DOI: https://doi.org/10.21437/Interspeech.2017-1386

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. DOI: https://doi.org/10.1017/S0954394513000161

Nycz, J. (2013). New contrast acquisition: Methodological issues and theoretical implications. *English Language and Linguistics*, *17*(2), 325–357. DOI: https://doi.org/10.1017/S1360674313000051

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. J. Hopper (Eds.), *Frequency Effects and Emergent Grammar* (pp. 137–158). John Benjamins. DOI: https://doi.org/10.1075/tsl.45.08pie

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95. DOI: https://doi.org/10.1016/j.specom.2004.09.001

R Development Core Team. (2020). *R: A language and environment for statistical computing*. In (Version 4.0.3) R Foundation for Statistical Computing. http://www.r-project.org/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. DOI: https://doi.org/10.3758/PBR.16.2.225

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. DOI: https://doi.org/10.1017/S0047404503324030

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics*, *41*(3), 249–263. DOI: https://doi.org/10.1016/j.wocn.2013.03.007

Seyfarth, S., Buz, E., & Jaeger, T. F. (2016). Dynamic hyperarticulation of coda voicing contrasts. *The Journal of the Acoustical Society of America*, *139*(2), EL31–EL37. DOI: https://doi.org/10.1121/1.4942544

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

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. DOI: https://doi.org/10.1111/j.1467-9841.2008.00370.x

Tse, J. K.-p. (1992). Production and perception of syllable final [n] and [ŋ] Mandarin Chinese: An experimental study. *Studies in English Literature and Linguistics*, *18*, 143–156.

Tse, J. K. P. (1998). Taiwan diqu nianqingren ㄓㄔㄕ yu ㄗㄘㄙ zhende bu fen ma? [台灣地區年輕人ㄓㄔㄕ與ㄗㄘㄙ真的不分嗎?] [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*, 151–164. DOI: 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. DOI: https://doi.org/10.5334/labphon.54

Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, *100*, 61–88. DOI: https://doi.org/10.1016/j.jml.2018.01.001

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

Yu, A. C. L. (2007). Understanding near mergers: the case of morphological tone in Cantonese. *Phonology*, *24*, 187–214. DOI: https://doi.org/10.1017/S0952675707001157

Żygis, M., Pape, D., & Jesus., L. M. T. (2012). (Non-)retroflex Slavic affricates and their motivation: Evidence from Czech and Polish. *Journal of the International Phonetic Association*, *42*, 281–329. DOI: https://doi.org/10.1017/S0025100312000205