Tonal Aspects of Languages 2016
24-27 May 2016, Buffalo, New York

Production and Perception of a Tonal Neutralization Case in Taiwan Southern Min

Mao-Hsu Chen

University of Pennsylvania, USA
chenmao@sas.upenn.edu

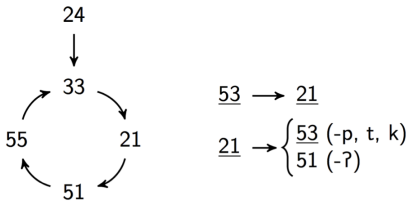
Abstract

This study explored the tonal neutralization of context tones 55 and 24 in Taiwan Southern Min, which are said to be realized both as sandhi tone 33 on surface when occurring in the context positions according to the Tone Circle. A production experiment with a reading task was conducted and the speech samples produced by speakers of two different age groups, old and young generations, were examined. The f0 contours of the two target context tones embedded in the minimal pair of sentences were compared using smoothing spline analysis of variance (SS ANOVA). The result showed an age-based acoustic variation, where the f0 contours of context tones 55 produced by the old speakers were significantly higher in pitch than those of context tones 24 throughout the entire contour, which was absent from the data of the young speakers, where the f0 contours of context tones 55 and 24 were indistinguishable from each other. The result of a preliminary perception experiment with an identification task was in line with the production data.
Index Terms: Taiwan Southern Min, incomplete neutralization, tone sandhi, production, perception

1. Introduction

Southern Min is a variant of the south Min Chinese dialects spoken in southeastern China. The regional dialect spoken in Taiwan called Taiwan Southern Min (TSM), or other aliases including Taiwanese, Xiamen, Amoy, and Hokkien, is noted for its complex tone sandhi (TS) system. Every lexical word in TSM has a base, also called *citation* or *juncture*, tone and a sandhi tone. The tone sandhi referring to the alternations between these two tones depends solely on the position of the word within a prosodic constituent, a tone group (TG). A word is realized with the base tone when occurring in the juncture position. i.e., the right edge of a TG; when occurring elsewhere (context position), it is realized with the sandhi tone. There are seven contrastive tones, including five free tones and two checked tones with CV[p, t, k, ʔ] syllable structure that are usually characterized by shorter duration and glottalized voice quality [1, 2].

The tone sandhi systems of the two group are illustrated in Figures 1(a) and 1(b), the former often referred to as the Tone Circle as suggested by the circular movement. The checked tones are marked with underlines, and the direction of the arrow shows the selection of the surface sandhi form. For instance, the sandhi form of a lexical high level tone 55 is a mid level tone 33, the sandhi form of a lexical mid level tone 33 is a low falling tone 21, and so on. As can be inferred from Figure 1, there exist two possible cases of tonal neutralization in the convoluted tone sandhi system in TSM: 1) both context tones 55 and 24 are realized as sandhi tone 33 on surface when occurring in context positions, and 2) context tone 21 and context checked tone 21 with a glottal stop coda are realized as a high falling sandhi tone 51 in context positions. To set aside the complication of checked tones, this study focuses on the first case.



(a) Free syllables (b) Checked syllables

Figure 1: *The tone sandhi of free syllables (Tone Circle) and that of checked syllables.*

While most of the early neutralization studies have been focused exclusively on the encoding of segmental information such as Port & Crawford's [3] study on the acoustic contrast of German voiced and voiceless stops at syllable-final position, the research interest has extended into autosegmental information, allotones or tone sandhi phenomena, in particular. Peng [4] investigated the Mandarin Third Tone Sandhi in which tone 3 (214)[1] becomes tone 2 (35) when followed by another tone 3. The production experiment found marginally significant effect of tone type (underlying tone 2 vs. the sandhi tone) on the mean f0 values; nevertheless, native speakers in the identification experiment failed to perceive the small pitch difference found in the production experiment, suggesting that the sandhi tone is not completely neutralized with the underlying tone 2 acoustically but is perceptually neutralized to some degree such that it is indistinguishable from tone 2.

Tsay et al. [6] examined four pairs of base and sandhi tones of the same tonal value (e.g. the underlying 21 and the sandhi variant derived from base 33) and showed complete neutralization in f0 and a phrase-final lengthening effect in base tone. They also looked at the two sandhi 33 tones derived from base 24 and base 55 using 3 minimal pairs of sentences but reported insignificant difference in f0. They therefore concluded that TS in TSM confirms the categoricality hypothesis. Myers and Tsay [7] further employed four minimal pairs of sentences along with two discourse contexts (a listener being present or absent) to see whether neutralization occurred between the two sandhi 33 tones in two conditions: a) across-positionally (juncture vs. context), in which the target words occupied different positions in a TG, one in the juncture position (realized with its

[1]The numerical values in parentheses represent pitch height on a five-point scale introduced by Chao [5], where 5 indicates the highest pitch and 1 the lowest.

http://dx.doi.org/10.21437/TAL.2016-18

base tone 33) while the other in the context position (realized as the sandhi tone 33 derived from base tone 55), and b) within-position (context vs. context), where the pairs of sentences had the same TG formation while the target words had different underlying tones (55 or 24) but the same surface tones 33. No significant effect of discourse context on duration and f0 was found, and the difference in overall f0 between context 55 and context 24 was not only insignificant but also extremely small, a mere 2.3 Hz. The duration with context 24 within context position is significantly longer, 8 ms on average, than context 55.

As only speakers of single age group participated in previous studies on the tonal neutralization in TSM, speakers of wider age range, young and old age groups, were recruited for this study. Meanwhile, while at most four minimal pairs were used for the production task in the aforementioned studies, more minimal pairs were incldued as stimuli in the current study. A replicate production experiment and a preliminary perception experiment were designed to examine whether the two mid level sandhi tones (33), one derived from the high level tone (55) and the other from the low rising tone (24) are completely neutralized in the same context position in production and whether native speakers can tell apart these two sandhi variants perceptually.

2. Production experiment

2.1. Subjects

13 native speakers of TSM, 6 males and 7 females, were recruited partially at the University of Pennsylvania and partially in Taiwan for this study, since it was hard to find old native speakers in the US. They were divided into two age groups, young and old. Seven speakers, 4 males and 3 females, were in the young age group (age ranging from 25 to 35 years old with an average of 28.43) and six speakers, 2 males and 4 females, were in the old age group (ranging from 52 to 66 years old with an average of 58.83). All the old speakers and two of the young speakers were recorded in Taiwan, while the others were recorded at the Phonetics Lab at the University of Pennsylvania. Those sampled from Pennsylvania were Taiwanese graduate students studying in Philadelphia who had stayed in the US for no more than two years. All speakers were able to speak and read Mandarin.

2.2. Instruments

Recordings were made either with Audacity on computers with Mac or Windows OS along with the Shure WH30 condenser headset microphone, or with a recording application using 44.1 kHz/16-bit sampling rate on a smartphone with iOS or Android OS using the built-in microphone.

2.3. Materials

There were 10 target monosyllablic minimal pairs embedded in stimuli sentences with a length of four to ten syllables with an average of 6.23 syllables. All the target words were at sentence-medial position. An example is shown in (1) and (2). The two sentences have the same sandhi domain structure and they only differ in the underlying tones of the target monosyllabic words, which are marked in bold. One has base tone 55 while the other has 24 as its base tone, but they are realized the same as 33 on the surface, i. e., the sandhi variant. A randomized reading list including 300 sentences, derived from (10 minimal pairs $\times$ 2 words + 10 fillers) with 10 repetitions), was generated for each subject.

(1) beh kio i **poe24** kau tang-si
want ask him accompany until when
'How long do you still want him to accompany you?'

(2) beh kio i **poe55** kau tang-si
want ask him fly until when
'How long do you still want him to travel by flight?'

2.4. Procedure

During the recording, one experimenter was either on-site or remotely monitoring the experiment to provide detailed instructions and ensure that the participants understood and performed the task as expected. The randomized stimuli were presented one by one with Chinese characters on a computer screen to elicit subjects production. Subjects were asked to first read the number of the order for each token followed by the stimuli sentence as naturally as possible, and they were allowed to rest at any time. The recording lasted for about 45 minutes with 300 tokens recorded from each speaker, 200 of which were the intended target used in further analyses.

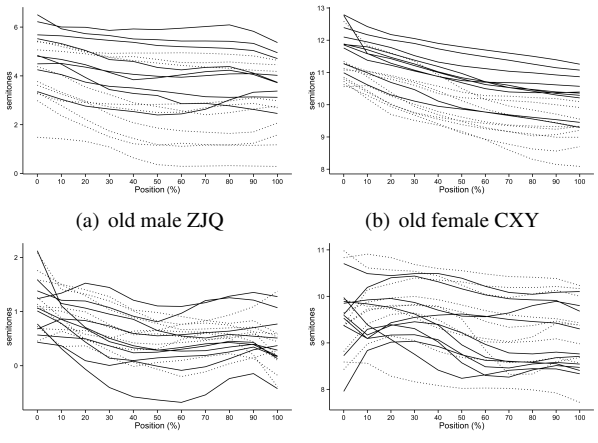
2.5. Data analysis

The nucleus of each target syllable was hand-labeled in Praat. The f0 values at every tenth in time of the syllable nucleus were extracted with VoiceSauce [8] such that the f0 contours were time normalized. The values were further converted into semitones using 100 Hz as the reference. The data of the two age groups were analyzed separately in order to see the generational difference. As there is no exact correspondence between words in TSM and Chinese characters, for some tokens other words were produced instead of the target words during the production task. This kind of mispronounced tokens were excluded from the statistical analysis. In total, 2,389 f0 contours, 1,258 tokens from the younger group and 1,131 tokens from the older group, were analyzed.

Each f0 contour was represented by eleven values for building the smoothing spline analysis of variance (SS ANOVA) models implemented in R [9] using the *gss* package [10]. The SS ANOVA has been applied to determine similarities and differences of multiple curve shapes, such as circadian rhythms [11] and tongue shapes from ultrasound imaging [12]. The advantage of this is that SS ANOVA models take entire curve shape into consideration and that they report statistical significance on the interaction term even when the difference lies only in a small portion of the curves. Bayesian confidence intervals are used to determine at which point in the comparison the curves are statistically different.

2.6. Results

Figure 2 shows the f0 contours of the token "poe" produced by one male and one female speakers in each age group. The solid lines indicate the f0 contours for tokens with base 55 whereas the dotted lines refer to the f0 contours of tokens with base 24. A difference in the realizations of the sandhi tone 33 between different age groups can be observed here: in Figures 2(a) and 2(b), most of the solid lines are above the dotted lines, suggesting that the old speakers pronounced the target words of base tone 55 with a higher pitch than those of base tone 24. However, in the lower two graphs, the solid and the dotted lines overlap with each other to a great degree, indicating that there is no clear

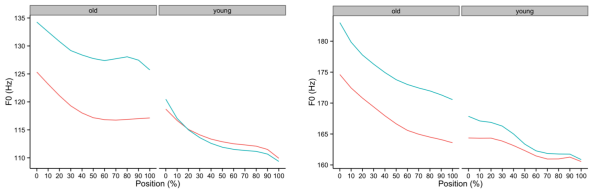


(a) old male ZJQ (b) old female CXY

(c) young male TSY (d) young female CSH

Figure 2: *The f0 contours in semitones of the token "poe" produced by one male and one female speakers in each age group. The solid lines indicate the f0 contours for tokens with tone 55 as the base tone whereas the dotted lines refer to the f0 contours of tokens with base 24 tone.*

division between the two sandhi tones produced by the young speakers.

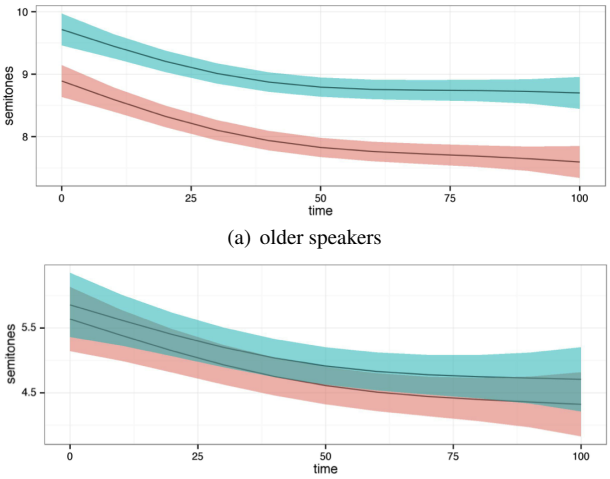


(a) male speakers (b) female speakers

Figure 3: *The average f0 contours of the two sandhi tones poe33 produced by all male and female speakers in each age group. The green lines are for tokens with base 55 and the red lines for tokens with base 24.*

Figure 3 shows the average f0 contours of the two sandhi tones of the token "poe" produced by male and female speakers in each age group. The green lines are for tokens with base 55 and the red lines for tokens with base 24. The graphs on the left indicate that the sandhi tones 33 derived from base tones 55 and 24 produced by the old male speakers are more distinct from each other than those produced by the young male speakers, where the green and the red lines are very close to each other. Similar result is observed for the average f0 contours produced by the female speakers, as shown on the right.

The result of SS ANOVA is shown in Figure 4. The solid lines are the smoothing splines for the main effects (here the two types of 33 tones) curves, representing the estimates of the population mean f0 contours, while the green and the red shaded areas around the smoothing splines are the corresponding 95% Bayesian confidence intervals. An overlap between the confidence intervals suggests insignificant differences between the f0 contours of the two target sandhi tones, as exemplified by the data of the young speakers in Figure 4(b). Figure 4(a), on the other hand, reveals a significant difference between the two types of f0 contours produced by the old speakers.



(a) older speakers

(b) younger speakers

Figure 4: *Smoothing spline estimates and 95% Bayesian confidence intervals for comparison of the f0 contours of the two target sandhi tones, where the one deriving from base 24 is represented by red color, and the one from base 55 by green.*

3. Perception experiment

3.1. Subjects

All speakers of the old age group, 2 male and 4 female, were recruited for the preliminary perception experiment, as their production data showed significant f0 differences between the sandhi tones 33 derived from 55 and those derived from 24.

3.2. Instruments and stimuli

The identification test was generated with a PHP script for each subject such that the participants could access it online. The script was run on a browser on a computer with all the audio stimuli embedded. The stimuli used in this task were sentences with the target tokens produced by the subjects themselves without any fillers; thus, there were 200 sentence tokens for each listener (10 minimal pairs $\times$ 2 words $\times$ 10 repetitions). Similarly, the mispronounced sentences were excluded, and the number of the stimuli ranged from 140 to 200 sentences, with an average of about 180 stimuli per subject.

3.3. Procedure

Listeners were run individually in a quite room, with the stimuli played to them through headphones. In each trial, subjects would first hear a self-produced audio stimulus and then be asked to choose between the corresponding minimal pair of sentences. They were able to replay the audio stimuli as many times as they wanted. They were also asked to specify their confidence level when making the forced-choice on a 5-point Likert scale ranging from "not confident at all" to "completely confident".

3.4. Data analysis

The ratio of the correct responses (e. g., the setnece with the target token of underlying tone 55 was opted for when the audio stimuli contained the surface tone 33 derived from the target with underlying tone 55) and the average confidence level were calculated for each token and for each subject. The val-

ues of confidence level were transformed into percentage values with 0% corresponding to "not confident at all" and 100% "completely confident". To control the response bias, the non-parametric analog of d', the measure A' [13], ranging from 0 to 1, was calculated for each listener using the formula shown in 1, where x is the ratio of correct responses and y is the ratio of false alarms. An A' near 1.0 refers to good discriminability, while a value near 0.50 indicates chance performance.

$$A' = 0.5 + [(x - y)(1 + x - y)/4x(1 - y)] \quad (1)$$

3.5. Results

Table 1 lists the result of the perception task. The average correct response rates pattern with the A' indices. Using the chance level (A' = 0.5) as the reference, one could see that three of the subjects, CJS, ZJQ, and WCM, were able to distinguish the two sandhi tones. The female subject YMA performed at chance level, while the other two female subjects, CXY and WMA, mixed the two tones and performed way below the chance level. The confidence levels were intended to see whether native speakers experienced any difficulty in distinguishing between the two tones in certain contexts, but the result shows high values everywhere.

Table 1: *The average ratio of the correct responses (CR), the A' index, and the average confidence levels (CL) for each subject.*

gender	subject (age)	avg. CR	A'	avg. CL
male	CJS (66)	0.8286	0.8967	92.80%
	ZJQ (60)	0.5455	0.5834	80.81%
female	CXY (57)	0.2921	0.2059	87.99%
	YMA (52)	0.50	0.50	84.11%
	WCM (60)	0.60	0.6704	94.60%
	WMA (58)	0.315	0.2221	91.80%

4. Discussion

For the production experiment, distinct acoustic patterns are observed for different age groups in terms of the effect of the base tone on the surface f0 values. While the young speakers did not maintain a pitch distinction between the two sandhi tones 33, one derived from base tone 55 and the other from base tone 24, the old speakers preserved the difference in the underlying forms and consistently produced the sandhi tone 33 with 55 as the base tone with higher pitch than the one derived from base tone 24. This suggests an age-based production difference, where there is incomplete neutralization for the old speakers and complete neutralization for the young speakers. As incomplete neutralization is closely related to the phenomenon of near mergers, the age-based acoustic variation could also signal an ongoing change towards complete merger, which is worthy of a longitudinal study.

One may argue that the production difference could possibly be due to the inherent register contrasts, where tones 55 (*yingping*) and 24 (*yangping*) differ in what is often called 'register'. Historically, words with base 24 have a voiced onset and words with base 55 have a voiceless onset. Thus, it could be that the old speakers merged 'pitch contours' but not 'register' while the young speakers merged both. As it is widely assumed that after Middle Chinese each of the four tones (*ping, shang, qu*, and *ru*, 'even, rising, falling, entering') split into two phonetic registers, *ying* and *yang* 'upper and lower' conditioned by the absence or presence of initial voicing/murmur on the syllables, this register contrast remained a phonetic difference until the initial voicing contrast was lost, which is thought to have begun sometime in the Tang period (618-907 A.D.) in at least some dialect families and have resulted in an eight tone system consisting of paired upper and lower ping, shang, qu, and ru tones [14]. It is questionable whether native speakers of current TSM could have preserve the register contrast that has been lost for at least one thousand years.

An issue followed by the observed production difference from the data of the old speakers related to whether this acoustic difference could be perceived by native speakers, and thus brought forth the preliminary perception experiment. The result shows a general pattern of male subjects outperforming female subjects. Among them, the oldest male subject CJS had the highest A' index value (0.8967). This suggests that despite inter-subject variation, native speakers of the old age group are capable of perceiving the small pitch difference and that they can distinguish the two, which is contrary to what was found with the case of Mandarin Third Tone Sandhi. On account of their highly overlapped f0 contours in the production data, the young speakers were not recruited in the preliminary identification task. Yet they should be included in the future work in order to present a more comprehensive study. Another possible revenue for futher investigation would be to include multispeaker audio stimuli in the perception task for both young and old listeners.

The across-the-board high values for the confidence levels selected by the subjects in the perception experiment could be interpreted as either the subject believed that the difference between the two tones did exist but may or may not perceive it, or the subjects generally assigned high values for the confidence levels all the time, which appeared to be the case for the oldest male subject CJS. Despite the small sampling size, subjects performances appear to correlate with their age, suggesting another possible age effect.

5. Conclusions

The production experiment examines the two types of sandhi tones 33, one derived from base 55 and the other from base 24, and shows a gradual merger in f0 contours and that the old speakers preserve the difference in the underlying forms, where 55 is inherently higher than 24 throughout the entire f0 curve while the young speakers do not maintain such acoustic contrast. The result of the perception experiment exhibits a gender-based perceptual variation, where the male subjects performed better than the female subjects at distinguishing between the two sandhi tones, along with a potential age effect, where older subjects in their 60s performed better than those in their 50s. This study suggests an age-based effect on the move from a near-merger towards a complete merger, along with a potential gender effect. If we think of the complete neutralization as the trend, the perception result seems to show that the neutralization interacts with not only age but also gender. A possible implication of this finding is the prevalent observation in sociolinguistics that women are more likely to lead the sound change; more data are certainly needed to draw that conclusion.

6. References

[1] R. L. Cheng, "Tone sandhi in Taiwanese," *Linguistics*, vol. 6, no. 41, pp. 19–42, 1968.

[2] ——, "Some notes on tone sandhi in Taiwanese," *Linguistics*, vol. 11, no. 100, pp. 5–25, 1973.

[3] R. F. Port and P. Crawford, "Incomplete neutralization and pragmatics in german," *Journal of Phonetics*, vol. 17, no. 4, pp. 257–282, 1989.

[4] S.-H. Peng, "Lexical versus phonological representations of Mandarin Sandhi tones," *Papers in laboratory phonology V: Acquisition and the lexicon*, pp. 152–167, 2000.

[5] Y. R. Chao, *Mandarin primer: An intensive course in spoken Chinese*. Harvard University Press, 1948.

[6] J. Tsay, J. Charles-Luce, and Y.-S. Guo, "The syntax-phonology interface in taiwanese: acoustic evidence," in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 3, 1999, pp. 2407–2410.

[7] J. Myers and J. Tsay, "Neutralization in taiwan southern min tone sandhi," *Interfaces in Chinese phonology: Festschrift in honor of Matthew Y. Chen on his 70th birthday*, pp. 47–78, 2008.

[8] P. K. Yen-Liang Shue and C. Vicenik, "Voicesauce: A program for voice analysis," *Journal of the Acoustical Society of America*, vol. 126, p. 2221, 2009. [Online]. Available: http://www.ee.ucla.edu/spapl/voicesauce

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[10] C. Gu, "Smoothing spline anova models: R package gss," *Journal of Statistical Software*, vol. 58, no. 5, pp. 1–25, 2014. [Online]. Available: http://www.jstatsoft.org/v58/i05/

[11] Y. Wang, C. Ke, and M. B. Brown, "Shape-invariant modeling of circadian rhythms with random effects and smoothing spline anova decompositions," *Biometrics*, vol. 59, no. 4, pp. 804–812, 2003.

[12] L. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variancea)," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 407–415, 2006.

[13] J. B. Grier, "Nonparametric indexes for sensitivity and bias: computing formulas." *Psychological bulletin*, vol. 75, no. 6, p. 424, 1971.

[14] W. S. Coblin, *A handbook of Eastern Han sound glosses*. Chinese University Press, Hong Kong, 1983.