Normalization of Zhangzhou Citation Tones

Yishan Huang[1], *Mark Donohue*[1], *Phil Rose*[2], *Paul Sidwell*[1]

[1]ANU Linguistics
[2]ANU Emeritus Faculty

yishan.huang@anu.edu.au; mark.donohue@anu.edu.au; philjohn.rose@gmail.com; paulsidwell@gmail.com

Abstract

The seven citation tones of the Southern Min dialect of Zhangzhou 漳州 are described impressionistically, and a linguistic-tonetic representation of their acoustics derived from the z-score normalization of the tones of 9 male and 12 female speakers. A normalization of the raw mean tonal data is shown to be slightly superior to a $\log_{10}$ transform, delivering about an eight-fold reduction in the between-speaker tonal variance (normalization index = 8.6). The data are then used to do a preliminary *Monte Carlo* investigation on how the normalization index changes with the number of speakers used to normalize.

Index Terms: normalization, tonal F0, tonal duration, Zhangzhou dialect.

1. Introduction: Normalization

Fundamental frequency (F0) as the main acoustic correlate of perceived pitch for linguistic systems of intonation, tone and stress shows high variability attributed to a wide variety of factors, both linguistic and non-linguistic [1]. Linguistically, F0 may be intrinsically perturbed by tautosyllabic segments, but also may be perturbed by other contextual factors, for instance the F0 height of segments, or the tone on adjacent syllables, the presence or absence of stress and dialectal accent [1]. Moreover, the F0 on one tone may alternate with another in a specific morpho-syntactic or phonological environment [2]. The non-linguistic factor for F0 variability is well known to be speaker-specific effects [1]. Anatomical and physiological differences in the individual vocal tract structure may generate dynamic acoustic outputs of difference even for phonologically identical utterances. For instance, female speakers generally have higher F0 values than males from their shorter and less massive vocal cords, thus, it is both theoretically and empirically possible for a female's phonological low tone to have a higher F0 value than a male's phonological high tone [3, pp. 38-41].

Given the acoustic variability resulting from differences of individual physiology, it becomes necessary to perform an effective reduction in the between-speaker variance prior to identification of different linguistic categories in signals. The mathematical analog of this process is just what the theory of normalization concentrates on. The aim of normalization is to abstract the variable Individual content away from invariable Linguistic and Accentual content in speech signals, and thereby derive a quantified representation of the variety in question [1].

The main aim of this paper is to present multi-speaker tonal data on the citation tones of the Chinese dialect of Zhangzhou, which has hitherto been impressionistically described a lot, but has arguably received inadequate attention acoustically.

How many speakers are needed for such a quantified investigation of a particular variety of speech? Numbers have been suggested e.g. [4], but not justified. Therefore a subsidiary aim of the paper will be to use the Zhangzhou data to do a preliminary investigation of how the efficiency of the normalization varies with the number of speakers used.

2. Zhangzhou Tones

2.1. Zhangzhou

Zhangzhou is a prefecture-level city situated in mainland China's southern Fujian province with approximately 4.8 million inhabitants [5]. The variety that native Zhangzhou people speak belongs to the Southern Min dialects of the Sinitic language branch of the Sino-Tibetan language family, which are primarily spoken in Southern Fujian and Taiwan.

2.2. Phonetic descriptions of Zhangzhou tones

There have been quite a few previous descriptions of Zhangzhou citation tones [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. All except two – [6, 10] – agree in describing seven different citation tones. Table 1 lists the descriptions under their Middle Chinese (MC) tonal categories (Ia, IIIb etc. - *a* and *b* stand for historical Yin and Yang registers, respectively). Tonal pitch values are given in the Chao five-point "tone letters" [16]. It can be seen that Zhangzhou does not distinguish separate reflexes of MC IIa and IIb tones (that is the reason only a *II* category is shown), but it otherwise preserves the other six MC categories.

Author	Year	Tone1(Ia)	Tone2(Ib)	Tone3(II)	Tone4 (IIIa)	Tone5(IIIb)	Tone6(IVa)	Tone7(IVb)
Dong	1959	24	212	53	32	33	32	13
Lin	1992	44	13	53	21	22	32	12
Ma	1994	44	12	53	21	22	32	121
FCCEC	1998	44	13	53	21	22	32	121
ZCCEC	1999	44	13	53	21	22	32	121
Gao	1999	45	23	53	21	33	21	121
Zhou	2006	44	13	53	21	22	32	121
Chen	2007	44	13	53	21	22	32	121
Yang	2008	44	13	53	21	22	32	121
Guo	2014	44	13	53	21	22	31	121

Table 1. *Previous descriptions of Zhangzhou citation tones according to Middle Chinese tonal category (in brackets).*

As is usual with impressionistic descriptions there are areas of both agreement and disagreement. Tone 3 for example is uniformly described with a high falling [53] pitch. Disagreements on the pitch values of other tones include, for instance: tone 1 is represented as a low rise [24], a high level [44] or a high rise [45]. Tone 2 is either a low dipping [212] or a low rise [13]. Tone 7 is transcribed as a low rise [13] or a low convex [121] by most scholars. It is not clear what this sort of variation is due to. It could possibly be ascribed to sub-dialectal variation and/or between-speaker or between-transcriber differences.

The previous descriptions are all impressionistic, but in the most recent decade, two scholars [5, 15] investigated the acoustic properties of Zhangzhou citation tones in terms of F0.

However, there are some inadequate and problematic aspects with regard to their research designs and analyses. For instance, [5] compared data from one male and one female directly in terms of their raw F0 values, ignoring the speaker-dependent effects on the F0 realizations and the importance of normalization for linguistic quantitative studies of speech signals. Normalised acoustic data from four speakers were given in [15] using the *T* algorithm, but the study only addressed sonorant-ending tones while neglecting the tones ending in obstruents, and the variation in tonal durations.

In 2015, as part of her Ph.D, the first author, who is a native speaker of Zhangzhou, collected extensive data in the field from 21 Zhangzhou speakers, including, of course, citation tones. It was clear that there were seven citation tones, but the auditory characteristics of most turned out to be different and more complicated from those available in the literature. In addition to pitch differences, the seven tones are characterized by a variety of co-occurring auditory features including length, vowel quality, voice quality, loudness and manner of articulation of syllable-initial consonants. For this paper, however, we will concentrate on their pitch, which it will now be helpful to generalise as follows:

- /**mid rising**/: rising pitch from the speaker's middle range to high, rather than a high level pitch or a high rising contour as previously described, e.g. /kɔ/ "mushroom 菇", /si/ "poetry 诗", /tɐŋ/ "east 东", /tsɛ̃/ "to contend 争", /tsʰjɐ/ "vehicle 车", /swɛ̃/ "mountain 山", /tĩ/ "sweet 甜".
- /**low level**/: level in the speaker's lower third pitch range with long duration, rather than a low rising contour as described before, e.g. /kɔ/ "paster 糊", /si/ "time 时", /tɐŋ/ "copper 铜", /pɛ̃/ "flat 平", /dɐm/ "male 男", /ɣu/ "cow 牛", /tʰɐw/ "head 头", /tsʰɐ/ "wood 柴".
- /**high falling**/: pitch falling from high in the speaker's range to low, with a short initial level component. This is similar to previous descriptions but with a lower offset, e.g. /kɔ/ "drum 鼓", /si/ "to die 死", /tɐŋ/ "to wait 等", /ɓɛ/ "horse 马", /tsjɐw/ "bird 鸟", /tsʰjõ/ "to rob 抢", /hɐj/ "sea 海", /tsu/ "host 主".
- /**mid falling**/: falling from the middle third of the speaker's pitch range to low, rather than the low falling contour of previous descriptions, e.g. /kɔ/ "to look after 顾", /si/ "four 四", /tɐŋ/ "frozen 冻", /kʰo/ "course 课", /kʰwɛ̃/ "to watch 看", /hi/ "drama 戏", /kʰɛ/ "guest 客".
- /**mid level**/: long and level in the middle third of the speaker's pitch, rather than at a low pitch range, e.g. /hɔ/ "rain 雨", /si/ "yes 是", /tɐŋ/ "heavy 重", /tjan/ "electricity 电", /pɛ̃/ "illness 病", /zi/ "character 字".
- /**short stopped mid fall**/: mid falling pitch as in the mid falling tone, but with salient short duration, similar to previous descriptions; high vowels are diphthongized, and become creaky to most speakers, e.g. /kɔk/ "country 国", /sit/ "colour 色", /kut/ "bone 骨", /kip/ "urgent 急", /hwɐt/ "law 法", /ik/ "one 一", /tsʰit/ "seven 七".
- /**stopped low level**/: similar pitch to the low level tone, but with a slight final fall due to the depressing effect by creaky phonation. Some rhymes lose their obstruent coda and become open with modal phonation. High vowels are diphthongized. This differs from the low convex contour as described previously, e.g. /tɔk/ "poison 毒", /sit/ "cooked 熟", /dɐk/ "six 六", /tit/ "straight 直", /tsɐp/ "ten 十", /zit/ "sun 日", /ɓɔk/ "wood 木".

It will be noted that the two stopped tones are explicitly treated as separate tonemes, as is usually done in Chinese tonology. The results of the normalization will present acoustic evidence to suggest that they might also be considered allotonically related to two unstopped tones.

3. Procedure

3.1. Speakers and elicitation

The speech data used in this study was collected from a linguistic field trip by the first author in Zhangzhou urban areas of Longwen and Xiangcheng during April 15th to May 9th in 2015. Data was collected from 21 native speakers: 9 males and 12 females. Their ages ranged from 38 to 65, with an average of 54 for males and 53 for females at the time of recording. None had physical difficulties in producing or perceiving speech sounds, and also no difficulties recognizing the words to be elicited.

The recording session was conducted in an acoustically absorbent room of the Zhangzhou Hotel with very little background noise and echo. The words to be read out were shown in simplified Chinese characters by means of *PowerPoint* with one slide per word. One major advantage of using this method is to make sure speakers produce the words in a clear and not-exaggerated voice with balanced and well-controlled intensity and speech rate.

All recordings were digitized at a sampling frequency of 44100 Hz in *Praat* [17] using a professional recorder (*Bluebaby* brand) kindly provided by Huaqiao University. This to a large extent ensured high quality recordings for further linguistic-phonetic processing and analysis.

3.2. Acoustic measurements

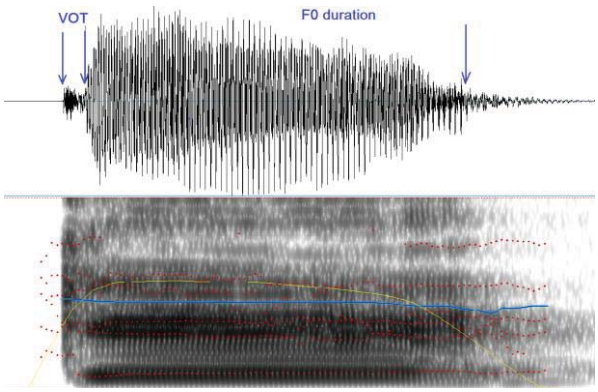

Figure 1: *A Praat labelled example* [kɐ̯44] "low" *by male Zhangzhou speaker WYF.*

Acoustic measurements were made with *Praat*. The rhyme portion of each monosyllabic token was considered as the tonally relevant F0 duration in this study as shown in Figure 1. The rhyme onset was set at the second strong glottal pulse where waveform amplitude begins to increase and formant patterns for vocalic sounds appear to be stable. The offset was judged to occur at the point where periodicity ceases in the waveform and periodically-excited formant patterns also cease to be visible in the spectrogram.

A *Praat* script was further run to automatically extract tonal duration values and F0 values at 10 equidistant sampling points along the labelled F0 duration. Manual corrections and

measurements were made where necessary in the *Praat* pitch window. The extracted raw F0 and duration values were then processed and plotted in *R* for further linguistic-phonetic analyses.

In this study, there were usually 20 tokens for each citation tone. Some tones for some speakers had less than 20 tokens as a consequence of speech errors or unreliably extracted F0 trajectory due to non-modal phonation. This resulted in a total of roughly 29,400 F0 measurements (= 20 examples × 7 tones × 21 speakers ×10 sampling points). Some of the tokens were exemplified in section 2.2 for reference. The tokens were chosen to include as many (sub-)minimal pairs as the phonotactics allowed, and also to maximally reflect phonetic realisations of linguistic tonal categories while balancing the intrinsic perturbation effects on F0 from tautosyllabic segments. Intrinsic vowel F0 was controlled by having comparable numbers of high mid and low vowels.

4. Normalization

Several normalization strategies have been proposed and compared for achieving an effective reduction in the between-speaker variation in tone. For instance, [18] proposed both a z-score normalization on the tones of Standard Vietnamese and a way of evaluating its performance with the normalization index. [19] proposed a T-value transform approach for single speakers of Chinese languages and [20] proposed a revised T-value normalization for a large corpus with multiple speakers. [1] used data from Wu Chinese to argue that z-score normalizations are preferable both on the balance of theoretical considerations and on numerical performance in comparison with fraction of range (FOR) transform strategy. [21] reported the superiority of logarithmic z-score normalization of Shanghai unstopped tones by comparing six different approaches which included z-score transforms, fraction of range (FOR) transforms, proportion of range (POR) transforms, ratio of Logarithmic semitone distances (LD) transforms and logarithmic proportion of range (LPOR) transforms.

As previous studies [1, 22, 25] have demonstrated the superiority of the z-score normalization, this was used on the 21 speakers' raw mean Zhangzhou data, both with and without prior transformation of raw mean F0 values to $\log_{10}$F0. The z-scored normalized F0 values z_i were calculated using the formula:

$$z_i = (x_i - m)/s \quad (1)$$

where x_i is an observed F0 value at one sampling point, *m* the mean F0 value and *s* is the standard deviation estimated from all the sampled F0 values of a given speaker's tones.

Normalization parameters of mean and standard deviation were estimated from all sampling points of all tones of all speakers. As proposed in [18], the efficiency of a normalization was quantified with the normalization index (NI), a measure of how well tonal values cluster after normalization. The NI reflects how much the normalization reduces the proportion of variance in a sample due to between-speaker differences in tonal values. The higher the NI value, the greater degree of reduction in between-speaker differences and the clearer the linguistic-phonetic content of the signal.

It was found that the normalization with prior log transform performed slightly worse (NI = 7.7) than the normalization with raw F0 (NI = 8.6). The results for the latter are therefore shown in figure 2, which also plots the normalized F0 values as a function of normalized duration to preserve the relationship between the tonal trajectories [22]. Duration was normalised with reference to the mean duration of all tones [23].

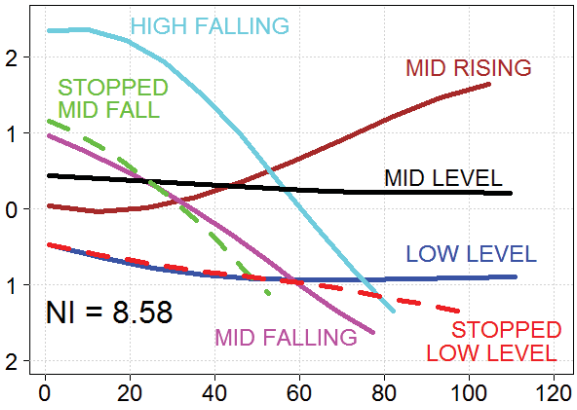

Figure 2: *Intrinsic Z-score normalized F0 for 21 Zhangzhou speakers' citation tones. Y-axis = normalized F0, x-axis = normalized duration (%).*

Figure 2 shows a fairly simple configuration of mean normalized tone trajectories corresponding closely to their impressionistic description. Of interest is the relationship between the stopped and unstopped tonal F0: it can be seen that the F0 of the two stopped tones – stopped mid fall and stopped low level (dashed lines) - is very similar to the F0 of the unstopped mid falling and low level tones respectively. The stopped tones have a slightly shorter duration and a falling offset perturbation, both of which are presumably related to their following voiceless unaspirated coda obstruents or extrinsic laryngealisation. These stopped and unstopped pairs could easily be said to be acoustic allotones of the same tonemes and the configuration could thus be said to comprise two falling, two level and one rising toneme. It is of further interest to note that the pairs of tones with the same contour are not maximally separated. Thus the level tones are not high and low but mid and low; and the falling tones do not fall from high and low but from high and upper-mid. A not too procrustean representation of these shapes in Chao's five point tone letters might be: high falling [51], mid falling and stopped mid falling [41], mid level [33], low level and stopped low level [22], mid rising [35], although this would need to be checked by an appropriate semitone transform, since Chao's tone letters are pitch descriptors and figure 2 is acoustic [24].

5. Optimum speaker number

How many speakers are necessary for achieving an effective estimation of the between-speaker variation and an accurate representation of a particular variety of speech? This is a question that does not seem to have received much attention. For example, in his book *Phonetic Data Analysis*, Ladefoged [4, p. 14] wrote:

> "Ideally you want about half a dozen speakers of each sex. … If you can eventually find 12 or even twenty members of each sex, so much the better."

Ladefoged's figures (24 – 40) were repeated in the *Handbook of Descriptive Linguistic Fieldwork* [p.254]. An ideal database of 30 speakers was recommended in [21] from a statistical perspective, possibly because 30 is about the number where a *t*

distribution becomes normal.

In order to investigate how the normalization index changes as a function of the number of speakers, a quasi-*Monte Carlo* approach was used. The F0 trajectories of the tones of the 12 female and 9 male Zhangzhou speakers were first modeled separately with cubic polynomials. The multivariate random command in *R* was then used to generate synthetic sets of 5 unstopped tonal F0 trajectories for 400 male and 400 female speakers from the normal distributions of each polynomial coefficient (multivariate random generation was necessary to take into account any correlation between the coefficients of the tonal trajectories). Random samples of size increasing from 1 male-female pair to 40 were then taken from the 800 speaker data, normalised, and their NI calculated. This was repeated for 30 trials. Figure 3 plots the means and flat-prior credible intervals for the NIs over the thirty trials.

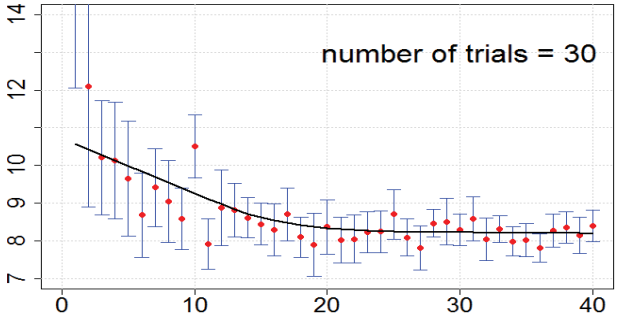

Figure 3: *NI values as function of number of random male-female speaker pairs normalised in 30 trials. Red = means, black = lowess smoothed means, blue = 95% credible intervals. X-axis = number of male-female speaker pairs, y-axis = NI.*

Figure 3 shows that the NI continues to decrease between 1 and ca. 20 speaker pairs (i.e. to 40 speakers), after which the rate of change decelerates. The gradually decreasing credible intervals reflect expected precision increasing with increasing *n*. According to these data, then, the NI is likely to be overestimated in a study with Ladefoged's lower estimate of six male-female pairs, and his upper estimate of 20 pairs would be needed to give a more accurate estimate. The uncertainties with lower speaker numbers look to be of the order of 3 NI, decreasing to about 1^{+} NI by 40 speakers. The effect of normalizing with data unbalanced for sex, as in this paper, remains to be investigated.

6. Summary

This paper has given both impressionistic and acoustic descriptions of the seven Zhangzhou citation tones from 21 speakers. The z-score normalization yielded an eight-fold reduction in the raw between-speaker differences in tonal values in order to extract and specify the Linguistic content of the tonal acoustics. It has also been shown that the two stopped tones have similar enough F0 trajectories to justify an allotonic interpretation of their relationship with the most similar unstopped tones.

7. Acknowledgements

The first author would like to express her thanks to Mark and Paul for their untiring supervisory help and encouragement over the last three years; to Siva for his marvelous help with *R*; and to Phil for the preliminary MC analysis and help in other things *R* and tonal. We also thank our two anonymous reviewers for their time and very helpful comments which we have tried to incorporate.

8. References

[1] Rose, P., "Considerations in the normalisation of the fundamental frequency of linguistic tone," Speech Communication, 6:343-352, 1987.
[2] Chen, M. Tone Sandhi, Cambridge University Press, 2000.
[3] Rose, P. Forensic Speaker Identification, Taylor & Francis, 2002.
[4] Ladefoged, P., Phonetic data analysis: an Introduction to fieldwork and Instrumental techniques, Wiley-Blackwell, 2003.
[5] Yang X., Studies of tones and regional cultures of Zhangzhou dialect 漳州方言声调与地域文化研究, Beijing: Zhongguo Shehui Kexue Chubanshe 中国社会科学出版社, 2008.
[6] Dong T., Four Southern Min varieties 四个闽南方言, Taipei: Zhongyang Yanjiuyuan 中央研究院, 1959.
[7] Lin B., "Zhangzhou vocabularies 漳州方言词汇", Fangyan 方言 ,1-3, 1992.
[8] Ma C., Studies of Zhangzhou dialect 漳州方言研究, Hongkong: Zongheng Chubanshe 纵横出版社, 1994.
[9] FCCEC, Fujian chorography-dialect volume 福建省志-方言志, Beijing: Fangzhi Chubanshe 方志出版社, 1998.
[10] Gao R., "Introduction to the sound system of Zhangzhou 漳州方言音系略说", in Minnan dialect-studies of Zhangzhou variety 闽南方言-漳州话研究, Beijing, Zhongguo Wenlian Chubanshe 中国文联出版社,109-116, 1999.
[11] ZCCEC, Zhangzhou chorography-dialect 漳州市志-方言 49, Beijing: Zhongguo Shehui Kexue Chubanshe 中国社会科学出版社, 1999.
[12] Zhou C.,The great Southern Min dictionary 闽南方言大词典, Fuzhou: Fujian Renmin Chubanshe 福建人民出版社, 2006.
[13] Chen Z., Southern Min dictionary of Zhangzhou variety 闽南漳州腔辞典, Beijing: Zhonghua Shuju 中华书局, 2009.
[14] Guo J., Zhangzhou Southern Min 漳州闽南方言, Zhangzhou: Zhangzhou Library 漳州图书馆, 2014.
[15] Yin X., "Acoustic analysis of tonal patterns in Zhangzhou 漳州话声调格局的分析", Journal of Chifeng University 赤峰学院学报, 30 (6):31-33, 2009.
[16] Chao Y., "A system of tone letters", Le Maître Phonétique 45: 24-27, 1930.
[17] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International5:9/10, 341-345, 2001.
[18] Earle, M., An acoustic phonetic study of North Vietnamese tones, Monograph 11, Santa Barbara: Speech Communication Research Laboratories Inc., 1975.
[19] F. Shi, "Tonal studies of disyllaibc words in Tianjin dialect 天津方言双字组声调分析", Studies in languages and linguistics 语言研究, 77-90, 1986.
[20] Shi F., Ran Q., and Wang, P., "On sound pattern 论语音格局" , Nankai Linguistics 南开语言学刊, 1-14, 2010.
[21] Zhu X., "F0 normalization-how to deal with between-speaker tonal variations 基频归一化-如何处理声调的随机差异", Linguistic Sciences 语言科学, 3-19, 2004.
[22] Rose, P., "A linguistic phonetic acoustic analysis of Shanghai tones," Australian Journal of Linguistics, 13:185-219, 1993.
[23] Rose, P., **"**Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study", *8th Australian Int'l. Conf. on Speech Science and Technology,* 198-203, 2000.
[24] Rose, P., "Transcribing Tone – A likelihood-based quantitative evaluation of Chao's tone letters", Interspeech, Singapore: 101-105, 2014.
[25] Rose, P. "A Comparison of Normalisation Strategies for Citation Tone F0 in Four Chinese Dialects", Proc. 16th Australasian SST Conf., 2016.