A Comparison of Normalisation Strategies for Citation Tone F0 in Four Chinese Dialects

Phil Rose

Australian National University Emeritus Faculty

philjohn.rose@gmail.com

Abstract

Seven common normalization strategies are compared for unstopped citation tone F0 in the Chinese dialects of Shanghai, Cantonese, Fúzhōu and Zhāngzhōu. A z-score normalization is shown to give clearly superior clustering as quantified by normalization index, but no indication of superiority for a prior log transform of F0 is found.

Index Terms: normalization, tonal F0, tonal duration, Cantonese, Shanghai, Fuzhou, Zhangzhou.

1. Introduction

Quantification, Lord Kelvin famously said, is the first step to science. Speech acoustics, although readily quantifiable, inevitably bear the imprint of the individual vocal tract that produced them, as well, of course, as the various parts of the brain driving that vocal tract. If we are focusing on the speech of the individual, as for example in forensic voice comparison, then this is indeed desirable. But if our focus is Language, then it is often necessary to remove as much speaker-dependent acoustic material as possible so as to arrive at a quantified parametric representation of the variety under question. This is one purpose of normalization: to extract and quantify the Linguistic and Accentual content in the signal by abstracting away from its Individual content. The result should be a quantified representation of the properties of the variety. This is illustrated in figure 1, using data from the 8 male and 8 female Shanghai speakers' unstopped citation tonal F0 in [1].

The left panel of figure 1 plots the 16 speakers' raw mean tonal F0 trajectories as a function of raw mean duration. Apart from the unsurprising fact that the females' tonal F0 generally lies, with a small overlap, higher than the males', the result is rather a mess: it is difficult to see from this figure how many tones there are and what their F0 trajectories are like. The normalisation in the right panel, however, where normalised tonal F0 is plotted against equalised duration, resolves the raw tonal F0 nicely into three groups corresponding to the three unstopped Shanghai tones (often described as high falling, high rising and low rising). Note that the between-speaker differences are reduced, but not eliminated. In particular it appears that there is a sex-related difference in the trajectory of the high-falling tone, which has a much steeper fall in males than females (this may relate to sex-differences in Onset obstruent production and therefore is not necessarily a tonal feature).

A configuration of mean normalised F0 trajectories similar to those in the right panel of figure 1 (shown with thicker black lines) could of course have been obtained without normalization, by simply taking the mean of the raw F0 values in the left panel. However, that would not have allowed an estimate of the variance around the mean normalised curves, which is necessary for many important dialectological, socio-phonetic, typological and even historical purposes, such as quantifying the tones of a variety [1], comparing varieties with respect to their tones [2, 3] or reconstructing tonal acoustics [4]. Neither would it allow for a means of evaluating the efficacy of the normalization, in which variance plays the crucial role.

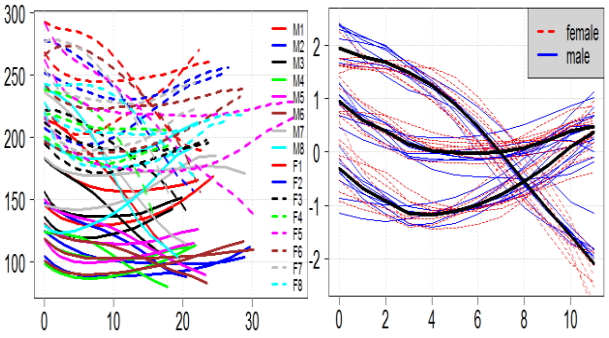

Figure 1: *Normalisation of 16 Shanghai speakers' unstopped tonal F0. Dashed lines = females. Left = mean raw data, right = normalised data. Thicker black lines = mean normalised F0. Axes: left = mean raw duration (csec.) & F0 (Hz.), right = equalised duration (%) & z-score normalised F0 (sds).*

The normalization strategy used in figure 1 is only one of several proposed, but there have not been many attempts to evaluate their performance. The performance of two approaches – z-score and fraction of range – was tested in [5] on Vietnamese, and in [6] on a Yǒngjiāng variety of Wu Chinese. More extensive testing was carried out on Shanghai in [7]. All three studies demonstrated the superiority of a z-score normalization. Even more extensive testing was carried out in [8, p.88ff] on the slope of contour tones from two more Wu dialects, Wúxī and Sōngjiāng. The study, however, was not concerned with overall minimization of between-speaker differences but in determining which normalization minimized sex differences in F0 while preserving age differences *qua* sociolinguistic information. Linear discriminant analysis showed this was best achieved by a simple semitone transform of the tonal F0.

This paper aims to contribute to the evaluation of normalization procedures by seeing which, if any, performs the best on the citation tonal F0 of three more dialects from the major groups of Sinitic: Cantonese, Fuzhou and Zhangzhou, as well as revisiting Shanghai. These data are described in the following section. Section 3 summarizes the main approaches to tonal F0 normalization and section 4 explains the methods for numerically evaluating their performance. Section 5 has the results.

2. Tonal Data

The tonal F0 data used in this paper were taken from previous studies on four varieties from three of the major so-called dialect groups of Chinese: Shanghai (from Wú 吳), Cantonese (Yuè 粵), Fuzhou and Zhangzhou (from Mǐn 閩). Only tones on sonorant-final syllables (舒聲) were used. All the data were controlled to a large extent for intrinsic vowel effects on F0, and are well-balanced for sex.

On syllables ending in a sonorant, conservative Hong Kong **Cantonese** contrasts six tones: three with level pitch, two with rising pitch, and one with falling pitch. The three level-pitched tones are located at the top, in the middle and just below the middle of the speaker's pitch range. Both rising tones start low in the pitch range, with one rising to high and one to mid. The falling tone starts low and falls still lower, such that its phonation type usually becomes non modal as it falls below the speaker's normal pitch range. The data were taken from a linguistic-tonetic description [9] of five female and five male young students recorded in the late nineties. Each tone had 24 tokens balanced for vowel height.

On syllables ending in a sonorant, **Shanghai** contrasts just three tones: one with pitch falling from high in the speaker's range to low; one with dipping pitch in the speaker's mid-pitch range; and one with pitch rising from low in the speaker's pitch range to mid, with or without an initial delay. The data were taken from a linguistic-tonetic description [1] from eight female and eight male students recorded in the late nineties. Each tone had 16 tokens balanced for vowel height.

On syllables ending in a sonorant, conservative **Fuzhou** 福州 is usually described as contrasting five tones [10, pp. 8-9]. There is consensus on the pitch of three: one with pitch falling from high in the speakers' range to low; one with convex pitch in the lower half of the speaker's range and one with level pitch high in the speaker's range. The two remaining tones are in the lower half of the pitch and are variously described as level, falling or rising. In the variety used here, both tones have slightly falling pitch, one in the mid and one in the low range. The Fuzhou data are from five males and five females. Although taken from two studies separated by about 20 years, ([10,11]), the speakers are of comparable age, having been born in the early sixties. The first study had 18 tokens per tone divided equally between [i~ei], [u] and [a] vowels. The second had 3 tokens per tone, all on [a] vowels.

The **Zhangzhou** 漳州 data are taken from a recent study [12] using 12 females and 9 males which shows a five-way tonal contrast on sonorant-final syllables: high and mid falling, mid and low level, and mid rising. Each tone had ca. 20 tokens reasonably well balanced between high, mid and low vocalic nuclei.

Table 1. *Normalisation strategies tested*

strategy	scale	
z-score	linear	$\log_{10}$
FoR_T *max-min*	linear	$\log_{10}$
$FoR_{T\,PT}$	linear	$\log_{10}$
ST_{meanF0}	semitone	

2.1. Preprocessing

The various sources of the data had used different strategies to sample tonal F0, so all F0 data were pre-processed by first removing any obvious offset perturbations, and then modeling their trajectories with 5th order polynomials. The resulting smoothed trajectories were then resampled at 10% points of duration, as well as at the 5% point, and normalised with different strategies using R code written for the purpose.

3. Typology of normalization strategies

Three basic normalisation strategies can be distinguished for tonal F0. As pointed out in [6], two involve ranges, being of the general form:

$$F0'_i = (F0_i - F0_{ref})/F0_{range} \quad (1)$$

where $F0_i$ is the value to be normalised, $F0_{ref}$ is a shifting factor and $F0_{range}$ is a scaling factor. Of these, a z-score normalization, as its statistical name implies, involves subtracting the value to be normalised $F0_i$ from a sample mean $F0_{mean}$, and dividing by a sample standard deviation $F0_{sd}$, so that the raw F0 values are transformed to multiples of so many standard deviations around a mean of zero:

$$F0.z_i = (F0_i - F0_{mean})/F0_{sd}. \quad (2)$$

A *z-score* normalisation of tone was first demonstrated on Vietnamese several decades ago [5] and has been used in many studies since. In the second, so-called *Fraction of Range* (FoR) approach, $F0_i$ is expressed as a fraction of the difference between two range-defining points $F0_{upper}$, $F0_{lower}$:

$$F0.For_i = (F0_i - F0_{lower})/(F0_{upper} - F0_{lower}) \quad (3)$$

A version of FoR called *T* [13] is very commonly used in the normalization of Chinese dialect tones, where $F0_{upper}$ and $F0_{lower}$ are a speaker's maximum and minimum F0 values, and the resulting fraction of range value is multiplied by 5 to help mapping onto the well-known Chao five point scale.

The main advantage of z-score normalization is that it ensures a global reduction of between-speaker variation. *FoR*, on the other hand, will force congruence at the range-defining points and thus compromise evaluation of effectiveness in terms of variance reduction [6]. In addition there is the problem of selecting the reference points to use in *FoR*, since it is not known *apriori* which points on a tonal F0 trajectory are comparable between speakers.

The third class of normalization strategy involves a single reference point, usually for semitone transforms. As explained in [8, pp.108-109], reference values can be both fixed, e.g. 100 Hz, or relative to the speaker. Examples of the latter include a speaker's F0 floor, or F0 ceiling, or mid-way between these two; or their overall mean value.

The choice of scale, of course, is logically independent of the normalization approach and can be considered as an additional typological option. A common approach, for example, is to z-score normalise log-transformed F0 values [7]; and the FoR_T normalization also employs a log scale. A final option concerns the source of the normalization parameters: *intrinsic* normalization uses parameters from the data to be normalised; *extrinsic* gets them from elsewhere, for example long-term data [14, 15].

In this paper the three basic approaches were tested: z-score, FoR, and single semitone reference, the first two both with and without prior $\log_{10}$ transforms. Two versions of FoR were tested: a standard FoR_T version, with a speaker's maximum and minimum F0 as range-defining points (FoR_T *max-min)*, and an additional version (FoR_{PT}) with putative comparable pitch target values as more suitable range-defining points. The upper range-defining point was the mean of the high level tone (Cantonese), and the peak of the high falling tone (Shanghai, Fuzhou Zhangzhou); the lower range-defining

point was the lowest point of the low to high/ low rising tone (Cantonese, Shanghai), and the mean of the minima in falling tones (Fuzhou, Zhangzhou). The semitone reference method used a speaker's mean tonal F0 as the reference ST_{meanF0} as this was found to be the most suitable in [8]. Table 1 summarises the seven different normalisations tested.

4. Numerical evaluation of normalization

The effectiveness of the normalisation is assessed by the method used in the first tonal normalisation study, on Vietnamese [5, p.133ff.]. Before normalisation the between-speaker variance in raw tonal F0 values will tend to be large because of between-speaker differences in tonal F0 caused by between-speaker differences in mass and length of the vocal folds. A female's high tone may have twice the F0 of a male, for example. After normalisation it is hoped that the between-speaker differences in tonal values will be minimised. Consequently, evaluation of the normalisation strategy involves quantifying how much the normalisation reduces the between-speaker tonal variance in the unnormalised data, a quantity called the *normalisation index* (NI). The idea is to estimate, for both raw and normalised data, the proportion of the overall variance in the data that is due to the between-speaker variance within tones. This is called the *dispersion coefficient*. Since the point of normalisation is to minimise between-speaker differences in tones, the proportion of the overall variance that is due to between-speaker tonal differences is expected to be smaller after normalisation, and so the ratio of the dispersion coefficients for the raw and normalised data – the *normalisation index* – quantifies by how much the between-speaker tonal variance has been reduced and between-speaker differences in tonal F0 have been minimised.

Using the Shanghai data in figure 1 as an example, the calculation of the NI can be formulated thus. Let $F0_{ijk}$ be the F0 value for the i^{th} speaker's j^{th} tone at the k^{th} sampling point. In the Shanghai data for example, i = 1… 16 speakers; j = 1… 3 tones; and k = 1… 12 sampling points (0%, 5%, 10% 20% … 100%). Then the mean F0 value over all speakers at a given sampling point in a given tone $\overline{F0}_{.jk}$ is:

$$\overline{F0}_{.jk} = \frac{1}{16}\sum_{i=1}^{16} F0_{ijk} \qquad (4)$$

with the variance around the mean F0 value over all speakers at a given sampling point in a given tone $S^2_{\overline{F0}_{.jk}}$ being:

$$S^2_{\overline{F0}_{.jk}} = \frac{1}{16}\sum_{i=1}^{16}\left(F0_{ijk} - \overline{F0}_{.jk}\right)^2 \qquad (5)$$

The mean of the variances $S^2_{\overline{F0}_{.jk}}$ at all 12 sampling points of all tones, called *between-speaker tonal variance* $\overline{S}^2_{\overline{F0}_{.jk}}$ is taken as an estimate of the variance representing between-speaker differences in tonal values:

$$\overline{S}^2_{\overline{F0}_{.jk}} = \frac{1}{36}\sum_{j=1}^{3}\sum_{k=1}^{12} S^2_{\overline{F0}_{.jk}} \qquad (6)$$

For the raw Shanghai data this was ca. 2479. In order to quantify the proportion of the overall variance taken up by variance associated with between-speaker differences in tone, the between-speaker tonal variance is then normalised with respect to the overall variance of the data. This is the mean of the between-speaker variances at each sampling point, i.e. ignoring the tonal differences. For the raw Shanghai data, this was ca. 2804. The ratio of the between-speaker tonal variance to the overall variance is called the *dispersion coefficient* (DC). In this case its value of ca. 2479/2804 = 88% indicates that there is almost as much variation *between* the Shanghai speakers' raw tonal values as in the data overall, and that they effectively do not cluster.

Since normalisation is intended to reduce the between-speaker differences in tonal F0, one expects the DC for the normalised data to be substantially smaller than the DC for the raw data. It is calculated, *mutatis mutandis*, in the same way as the raw DC, namely as the ratio of *between-speaker normalised tonal variance* to *overall normalised variance*. The DC for the normalised Shanghai data was ca. 9%, indicating that only a small amount of the overall variance was taken up by between-speaker differences in tone. The normalisation index (NI) is then defined as the ratio of normalised DC to raw DC. For this normalisation, the NI was 88%/9% = 9, meaning that normalisation has resulted in about a nine-fold reduction in the proportion of variance in the raw data due to between-speaker differences in tone.

5. Results & Discussion

Results are shown in table 2. They agree with previous studies in showing a clear superiority for z-score normalization. Unlike previous results, however, a prior log transform for the z-score is not always preferable: there is nothing to choose between log and linear NIs for Shanghai and Fuzhou; and linear NI is much better than log in Cantonese and a little better in Zhangzhou. This may be because a log transform is over-sensitive to the density of tonal trajectory shapes in the lower pitch range. A FoR_T normalization clearly performs badly if the range is automatically set between a speaker's maximum and minimum F0 values, but can achieve between ca. 50% - 80% of the effectiveness of a z-score with judiciously chosen range-defining pitch targets. The mean semitone transform also performs badly, presumably reflecting large between-speaker differences in their tonal semitone ranges.

Given the large NI difference between Cantonese, with six tones and Shanghai, with three, it is natural to speculate on relationships between NI and number of citation tones or nature of their contrasts. This would be conjecture because NI is a function also of the number of speakers involved and because these are results from single trials: different results would occur purely by chance from further trials with different speakers [12].

Table 2. *Normalisation Indices for strategies tested*

Strategy	Sh	Ca	Fu	Zh
z-score raw	9.8	**21.4**	6.7	**13.0**
z-score $\log_{10}$	**9.9**	18.1	**6.9**	12.0
FoR$_T$ *max-min*	4.2	5.9	4.8	7.3
FoR$_T$ log_{10}*max-min*	4.2	4.6	4.6	6.0
FoR$_T$ *PT*	7.8	15.9	4.0	7.2
FoR$_T$ log_{10}*PT*	7.9	11.9	4.4	6.1
ST*meanF0*	5.8	10.4	3.5	8.0

Figure 2 plots the linear z-score normalised F0 shapes for the four varieties. In anticipation of the paper's final point they are plotted as a function of normalised, not equalised duration (raw duration was transformed to a percentage of a reference duration from the mean duration of all tones [9, 16]). The plots have also been colour-coded to show the different reflexes of Middle Chinese tones in the four dialects.

It looks from figure 2 that, with the possible exception of the high falling tone in Shanghai, Fuzhou and Zhangzhou, and the mid level tone in Cantonese and Zhangzhou, there are no other shared tones between the three varieties. For example, the high level tone appears higher in a speaker's range in Cantonese than in Fuzhou, and the low falling tone in Fuzhou appears to fall at a slower rate than in Cantonese. This may very well be the case, but before one can reliably use such representations to infer linguistic-tonetic (non-) equivalence across dialects one must be sure the normalisation parameters are comparable. Given the different number of tones in each variety and the different way they are distributed, this is unlikely. More research is needed into how normalised representations may be mapped onto each other.

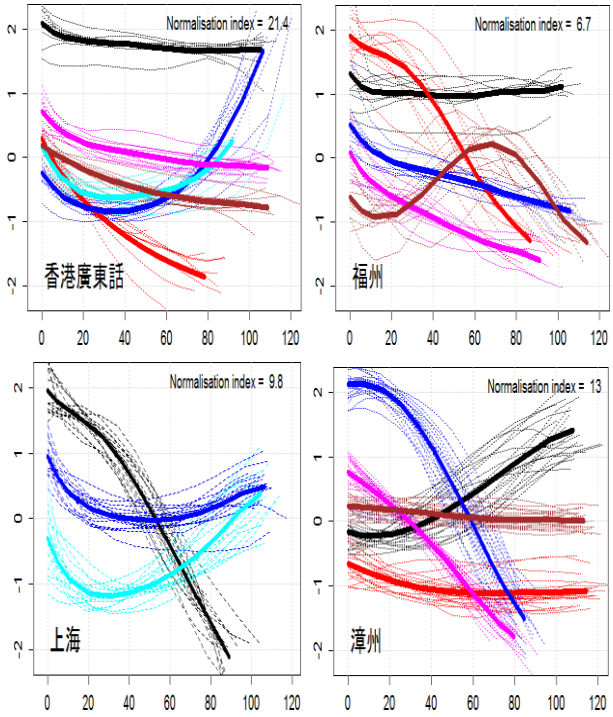

Figure 2: *Linear z-score normalisation of unstopped tones in (clockwise from top left) Cantonese, Fuzhou, Zhangzhou, Shanghai. Middle Chinese tonal reflexes are colour-coded:* **Ia Ib IIa IIb IIIa IIIb**. *Thicker lines = mean normalised F0. X-axes: = normalised duration (%), y-axes = z-score normalised F0 (sds).*

6. Summary & Way Forward

The results of the paper increase the strength of evidence in favour of the superiority of z-score normalization of tonal F0.

The normalisation strategies evaluated in this paper only apply to F0 as a function of *equalised* duration. It may be the case that between-speaker variance in F0 is reduced even more if the normalised F0 is considered a function of *normalised* duration.

As yet, the evaluation of the performance of durationally normalised F0 remains unaddressed. One possible solution is suggested in figure 3, which plots the Shanghai low rising tone F0 as a kernel density distribution. The plot was generated by modelling the distribution of the different speakers' F0 values at every centisecond with a kernel density and then plotting the resulting set of densities as a surface. The tonal density of course increases after normalisation (compare the density axes), and it would be possible to compare the relative amount of increase in durationally normalised and durationally equalised F0.

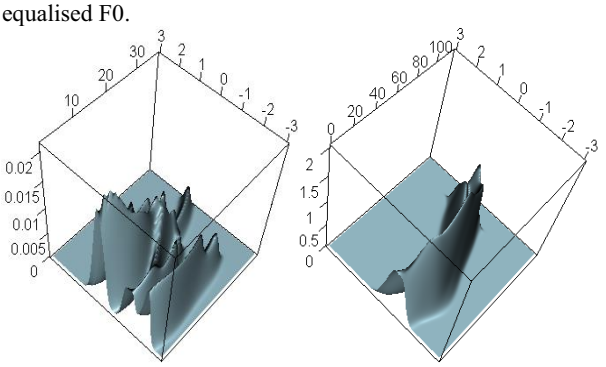

Figure 3: *Density of Shanghai low rising tone F0, left = raw, right = normalised, x-axes = duration, y-axes = z-score normalised F0 (sds), z axes = density.*

7. Acknowledgements

Many thanks to my two anonymous referees for their help!

8. References

[1] Rose, P., "A Linguistic Phonetic Acoustic Analysis of Shanghai Tones", Australian Journal of Linguistics 13:185-219, 1993.

[2] Zhu X., and Rose, P., "Tonal Complexity as a Dialectal Feature: 25 Different Citation Tones from Four Zhejiang Wu Dialects", 5th Int'l Conf. on Spoken Language Processing 3:919-922, 1998.

[3] Steed, W. and Rose, P., "Same tone, different category: linguistic-tonetic variation in the areal tonal acoustics of Chu-qu Wu", Interspeech: 2295-2298, 2009.

[4] Rose, P., "Oujiang Wu tones and Acoustic Reconstruction", in Bowern, Evans, Miceli [Eds.] Morphology and Language History, 235-250, John Benjamins, 2008.

[5] Earle, M.A., An acoustic phonetic study of North Vietnamese tones, Monograph 11, Speech Communication Research Laboratories Inc., Santa Barbara, 1975.

[6] Rose, P., "Considerations in the normalisation of the fundamental frequency of linguistic tone", Speech Communication, 6(4):343-352, 1987.

[7] Zhu X. 朱晓农, "基频归一化 – 如何处理声调的随机差异" F0 normalization – how to deal with Between speaker Tonal Variations, 语言科学 [Linguistic Sciences] 3(2):3-19, 2004.

[8] Zhang J., A Sociophonetic Study on Tonal Variation of the Wúxī and Shànghǎi Dialects, LOT Netherlands Graduate School of Linguistics, 2014.

[9] Rose, P., "Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study", Proc. 8th Australian SST Conf., 198-203, 2000.

[10] Donohue, C., Fuzhou tonal acoustics and tonology, Lincom, 2013.

[11] Peng G., A Phonetic Study of Fuzhou Chinese, Ph.D. thesis, City University of Hong Kong, 2011.

[12] Huang Y., Donohue, M., Rose, P., Sidwell, S. "Normalisation of Zhangzhou tones", Proc. 16th Australasian SST Conf., 2016.

[13] Shi F. 石锋, "天津方言双字祖声调分析" [An analysis of tone in Tianjin disyllables], 语言研究 [Yuyanyanjiu] 1, 1986.

[14] Rose, P., "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?", Speech Communication 10:229-247, 1991.

[15] Rose, P., "Mr. White Goes to Market - Running Speech and Citation Tones in a Southern Thai Dialectal", Proc. 15th Australasian SST Conf., 110-113, 2014.

[16] Zhu X., Shanghai Tonetics, Lincom Studies in Asian Linguistics 32, Lincom, 1999.