Declination-adjusted Normalisation of Cantonese Citation Tones

Phil Rose

Independent Researcher, Australian National University Emeritus Faculty

https://philjohnrose.net

Abstract

An experiment is described to see if the normalisation of citation tone fundamental frequency (F0) can be improved by taking into account its occasion-specific decay. An optimum baseline z-score normalization of the F0 of ten Cantonese speakers' unstopped citation tones without such adjustment gave about a twenty-fold reduction in the between-speaker tonal variance (normalisation index = 21.3). It is shown that adjusting for F0 decay by simply using the linear slope of the phonologically level tones can improve on this baseline normalization a little, by up to about 11% (NI = 23.6). The result is also a representation closer to the tonal pitch.

Index Terms: normalisation, tonal F0, tonal pitch, F0 declination, Cantonese

1. Introduction

1.1. Declination

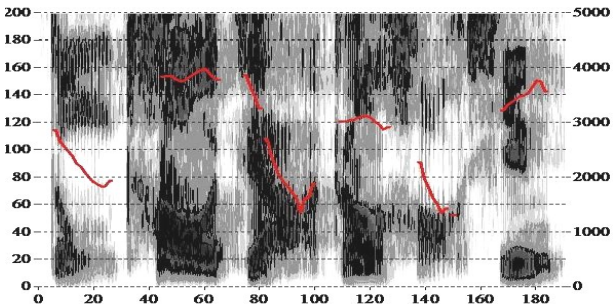

Figure 1: *Alternating sequence of L and H tones in a six-syllable Mandarin utterance (F0 superimposed on wideband spectrogram). Y axis-left = F0 (Hz),Y axis-right = spectral frequency (Hz), X-axis = duration (csec.)*

Fundamental frequency (F0) tends to gradually drop through an utterance: this is a phonetic commonplace [1 p. 53]. This so-called *declination* can be most easily seen and modeled in tone languages, where the language's phonology makes clear which of the utterance's F0 values are tonologically equivalent [1 pp. 69-71, 2 pp. 296-297]. This is illustrated in figure 1 with a Mandarin utterance 董超假装好心 *Dǒng Chāo jiǎzhuāng hǎoxīn* 'Dong Chao pretended to have a kind heart'. The utterance was taken from a recorded narrative [3 p. 116] which happened to have six syllables carrying a sequence of alternating L and H tones (tones 3 and 1). Figure 1 plots the F0 time-course for the utterance superimposed on a wide-band spectrogram to show the relationship of the F0 to the segmental structure. It can be seen that the falling F0 trajectory of the L tones on the odd syllables (which may be an extrinsic part of the tone's underlying fall-rise target or an intrinsic effect of the voiceless obstruent Onset consonants) gradually decreases through the utterance. The higher F0 of the H tones also decreases over the first two even syllables. The higher value of the F0 on the H tone of the final syllable *xīn* [ɕɪn] probably reflects intrinsic vowel effects, as the utterance was played to a native speaker phonetician who said that it was unmarked and the final syllable not perceptually prominent.

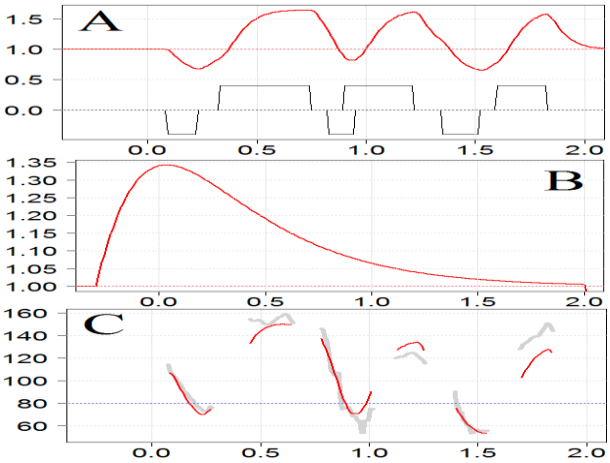

Figure 2: *Declination decomposition with Fujisaki command-response model of F0 in the utterance in figure 1. A = Tonal impulses and response, B = Phrasal Response, C = actual (grey) and estimated (red) F0. X-axis = duration (sec.).*

Declination has been widely studied from several aspects, including to what extent it is under a speaker's control and is used to convey linguistic and paralinguistic information; and how best to model it. A useful overview is in [1]. The well-known Fujisaki *command-response* model, e.g. [4 5] provides an insightful way of separating out the declination from the tonal components. This model factors the time-varying F0 into two types of component, both modeled as impulses of given amplitude and duration. A *tonal* component represents the response of the speech production mechanism to impulse commands for implementing tone. Such impulses are shown on the bottom line of panel A of figure 2. They consist of a string of equal amplitude impulses of alternating polarity. These model the low and high tones respectively [6] and are said to relate physiologically to the *pars recta* activity of the crico-thyroid [7, p.4]. The impulse response to these commands of differing polarity is shown in the top line of panel A in red. The second, or *phrasal*, component represents a much slower time-varying response and accounts for the gradual change in F0 – the declination – throughout an utterance or international phrase. Its response is shown in panel B. This is said to correspond physiologically to the *pars obliqua* of the crico-thyroid, but might also represent gradually decaying sub-glottal pressure. Panel C shows the combined tonal and phrasal responses in red and the actual F0

in grey. It can be seen there is a fairly good fit. A better fit could of course have been obtained if the amplitudes of the individual tone impulses had been individually manipulated, but that would have obscured the point of factoring the F0 into a fixed tonal and declinational component (the duration of the tonal components, however, still had to be independently specified to capture the metrical structure of the utterance).

1.2. Citation tone declination

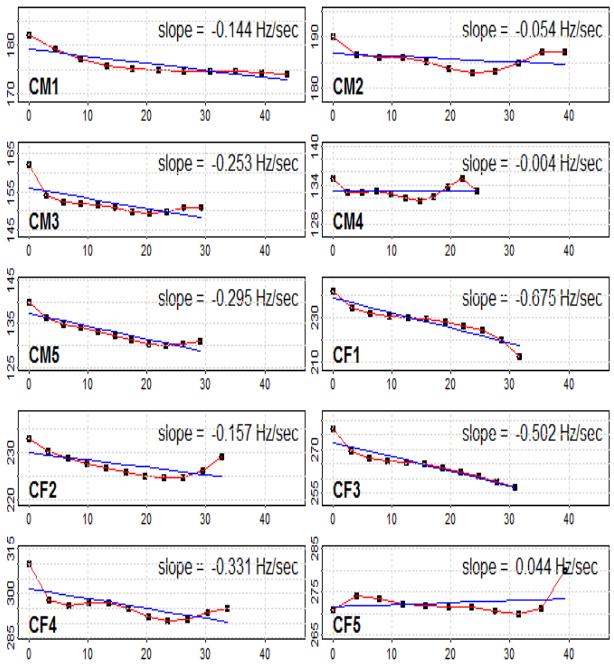

Figure 3: *Mean F0 as function of mean duration for Cantonese high level tone in five male (CM 1 – 5) and five female (CF 1-5) speakers. Blue line = least squares regression. X-axes: mean raw duration (csec.), y-axis = mean F0 (Hz.).*

A gradual decrease in F0 can also be sometimes observed on tones in monosyllabic utterances, as for example when citation tones are elicited. Figure 3 shows F0 plotted as a function of duration for five female and five male Cantonese speakers' high level citation tone. A least-squares regression line has been fitted, and the value of its slope shown. (The data are from a multispeaker acoustic description of conservative Cantonese citation tones [8], where the tones were read out in random order from Chinese characters on prompt cards. Each speaker's F0 shape is the mean of 16 tokens balanced for intrinsic vowel F0 and sampled at 10% points of duration. The mean F0 and duration data may be downloaded from [9].

It can be seen firstly that speakers differ in their decay, but that, overall, males have less F0 decay than females. Since the F0 of males is usually lower than females, this suggests that the rate of decay may be related to overall F0. But there are clearly within-speaker differences that do not relate to overall F0. CM4 and CM5, for example, differ the most in F0 decay but have very similar F0 for their high level tone; and CF1 and CF3 have similar decay but rather different F0. Importantly, there is also no obvious correlation of F0 decay with duration (one might expect that the longer the tone were sustained the further its F0 would drop). This suggests that the F0 decay may be occasion-dependent, and that adjusting for this may help to improve tonal F0 normalization. It is the purpose of this paper to investigate this hypothesis using citation tonal data from Cantonese.

2. Normalisation

An even more basic phonetic commonplace than declination is that speech acoustics inevitably bear the imprint of the individual vocal tract that produced them, as well, of course, as the brain that drove that vocal tract. If we are focusing on the speech of the individual, as for example in forensic voice comparison [10], 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 in turn is necessary for many important dialectological, socio-phonetic, typological and even historical applications, such as quantifying the tones of a variety [11], comparing varieties with respect to their tones [12] or even reconstructing tonal acoustics [13].

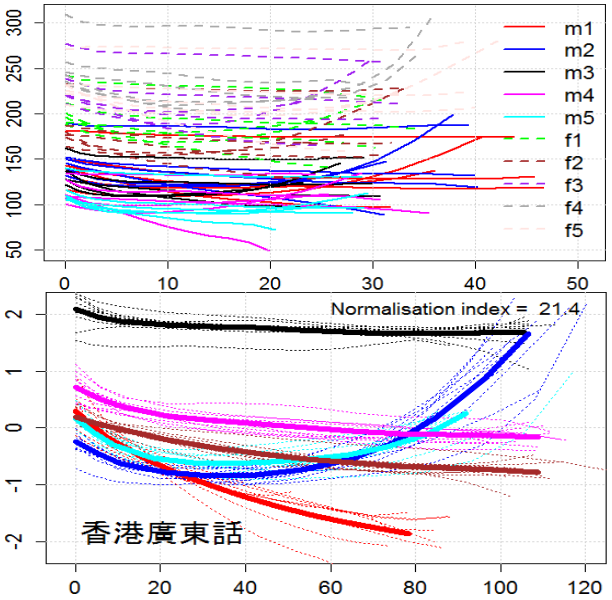

Figure 4: *Normalisation of Cantonese unstopped citation tonal F0. Top = raw mean tonal F0 for 10 speakers' tones. Dashed lines = females. X-axis = mean duration (csec.), y-axis = mean F0 (Hz). Bottom panel = z-score normalised F0. Thicker lines = mean normalised F0. X-axis = normalised duration (%), y-axis = normalised F0 (sds around mean).*

Figure 3 shows the normalisation of the tones of the same 10 Cantonese speakers shown in figure 3. 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 term *pitch* is used here in its proper perceptual sense, not as a synonym of F0.) The three level-pitched tones are located at the top, in the middle and just below the middle of the speaker's pitch range. Examples are [si: high level pitch] *poem* 詩; [si: mid level pitch] *try* 試, [si: lower-mid level pitch] *event* 事. Both rising tones start low in the pitch range, with one rising to high and one to mid: [si; low pitch rising to high] *shit* 屎, [si: low pitch rising to mid]

市 *market*. The falling tone starts low and falls still lower, such that its phonation type usually becomes non modal (breathy or creaky) as it falls below the speaker's normal pitch range: [si: low falling pitch] *time* 時.

The top panel of figure 4 plots the 10 Cantonese speakers' six tones' raw mean F0 trajectories as a function of raw mean duration. Apart from the unsurprising fact that the females' tonal F0 generally lies, with an 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 bottom panel of figure 4 shows a z-score normalisation of the data [14] which resolves the raw tonal F0 nicely into six groups. It was shown in [15] that this type of normalisation easily outperforms other types of normalisation proposed. The performance if the normalisation in figure 4 was therefore used as a baseline.

So the question this paper asks is: can an improvement on this baseline normalization be made by incorporating the type of F0 decay evident in figure 3? Although the z score normalisation is well-known, its evaluation may not be. The following section thus addresses the numerical evaluation necessary to determine an improvement in normalisation.

2.1. Numerical evaluation of normalisation

The effectiveness of normalisation is currently estimated by the method used in the first tonal normalisation study of some decades ago, on Vietnamese tones [16, p.133ff.]. Variance plays a crucial role. 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: this effect of secondary sexual bimorphism can be seen in figure 4. 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 how much between-speaker differences in tonal F0 have been minimised.

Using the 10 Cantonese speakers' data in the top panel of figure 4 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 Cantonese data for example, i = 1… 10 speakers; j = 1… 6 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:

$$\overline{F0}_{.jk} = \frac{1}{10}\sum_{i=1}^{10} F0_{ijk} \qquad (1)$$

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

$$S^2_{\overline{F0}.jk} = \frac{1}{10}\sum_{i=1}^{10}\left(F0_{ijk} - \overline{F0}_{.jk}\right)^2 \qquad (2)$$

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:

$$\overline{S}^2_{\overline{F0}.jk} = \frac{1}{72}\sum_{j=1}^{6}\sum_{k=1}^{12} S^2_{\overline{F0}.jk} \qquad (3)$$

For the raw Cantonese data in the top panel of figure 4 this was 2580.0. 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 Cantonese data, this was ca. 2887.9. The ratio of the between-speaker tonal variance to the overall variance is called the *dispersion coefficient* (DC). In this case its value of (2580.0 / 2887.9 =) ca. 89.3% indicates that there is almost as much variation *between* the Cantonese 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 Cantonese data was (0.04 / 0.94 =) ca. 4.2%, 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 (89.3% / 4.2%) = ca. 21.3, meaning that normalisation has resulted in about a twenty-fold reduction in the proportion of variance in the raw data due to between-speaker differences in tone.

3. Declination-Adjusted Normalisation

3.1. Procedure

For non-rising tones, negative and positive F0 offset perturbations can obviously effect estimation of the slope. In order to control for this, the F0 at the last two sampling points in these tones was removed. The remaining raw F0 trajectory was then modeled with a 4th degree polynominal to permit resampling of F0 at 0%, 5%, 10%, 20%, …, 100% of duration. A least-squares regression line was then fitted to the F0 trajectory, and the raw F0 adjusted by its slope and intercept according to (4)

$$F0_t' = F0_t - (mt + b) \qquad (4)$$

where $F0_t'$ = declination-adjusted F0 at time t, $F0_t$ = mean F0 at time t, m, b = slope, intercept of least squares regression line, t = time t. This process is illustrated in the left panel of figure 5, which shows how the tonally relevant F0 for CF1's high level tone (solid grey line) is adjusted by the slope of its least squares regression line (black dotted line) to its declination-adjusted value (solid red line).

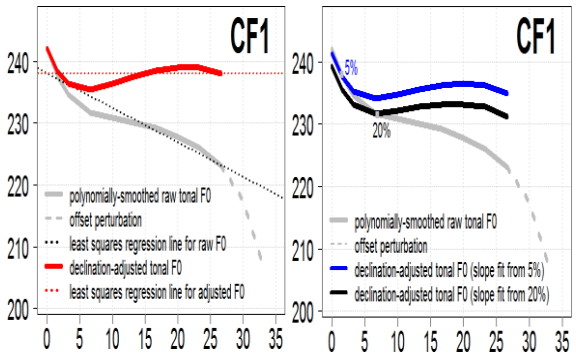

Figure 5: *Illustration of F0 declination adjustment for Cantonese female 1's high level tone. = X-axis = duration (csec.) Y axis = F0 (Hz).*

3.2. Parameters and results

Two parameters were manipulated in the declination-adjustment. The initial few centiseconds of all speakers' F0 show consonantally-induced onset perturbations, and, as these will affect the tone's F0 slope, normalisation was tested with increasing portions of initial F0 trajectory (0%, 5%, 10%, 20%) removed to estimate the slope. Allowing the onset to vary from 0% to 20% of duration tests whether a better normalization is achieved if the onset perturbation is not included in the F0 slope. This is illustrated in the right panel of figure 5, which shows the adjusted F0 for CF1's high level tone when the slope is based on the raw F0 with the first 5% removed (blue line), and with the first 20% removed (black line). Because of the shape of the initial part of the raw tonal trajectory, removing these portions gradually decreases its slope. As can be seen in figure 5, the effect of this is to pivot the declination-adjusted F0 around the percent sampling point onset. This parameter is therefore called *pivot*, and slopes were generated using four pivot values at 0% 5% 10% and 20% of the raw tonal duration.

Table 1: *Normalisation indices from different normalisation trials.*

pivot	0%	5%	10%	20%
option 1	slope from high-level tone			
	23.0	**23.1**	23.0	22.4
option 2	slope from mid-level tone			
	23.1	22.8	22.4	22.1
option 3	slope from lower-mid-level tone			
	21.6	21.5	21.2	20.8
hybrid combinations				
option 4	23.5	**23.6**	23.3	22.9

It would, with six different tones, be possible to adjust each tone by a different slope. Apart from the overfit that this is likely to cause (if you torture the data long enough, it will confess!), it is also counterintuitive to imagine six different declination factors at play; and in addition a declination slope can only reasonably be estimated from a citation tone which can be assumed to have a level pitch target (of which there are three in Cantonese). Therefore declination-adjusted normalisation was first tested with three options, where all tones were adjusted by a single slope estimated from each of the three level tones (high-level, mid-level, lower-mid-level). Results are shown in table 1, where it can be seen, firstly, that the pivot does have a slight effect. Generally the performance decreases with increasing amounts removed from the initial part of the tone trajectory: apparently it is not a good idea to remove onset perturbations. As far as between-tone differences are concerned, the slope from the lower-mid-level tone (option 3) performs the worst, and only marginally better than the baseline value of 21.3. The high-level tone slope (option 1) outperforms the mid-level tone slope (option 2) for three out of four pivots, although both achieve the same maximum normalisation index of 23.1. This is a small improvement on the baseline value of 21.3, indicating that adjusting for declination can marginally improve clustering.

It was also tested whether a better performance can be achieved with a hybrid approach using both slopes from the high- and mid-level tones to separately adjust different sets of tones (option 4). It was found that when all tones except the mid-level tone were adjusted by the slope of the high-level tone, and the mid-level tone was adjusted by its own slope, a slightly better NI of 23.6 is obtained – an improvement over the baseline of about 11%. This suggests that F0 decay and F0 height may be related in a more complex way such that slope may increase with distance from a central F0 value. This is an obvious thing to test.

Figure 6 shows the resulting normalisation with option 4. It can be seen that the level tones now have correspondingly level normalised F0 over much of their duration.

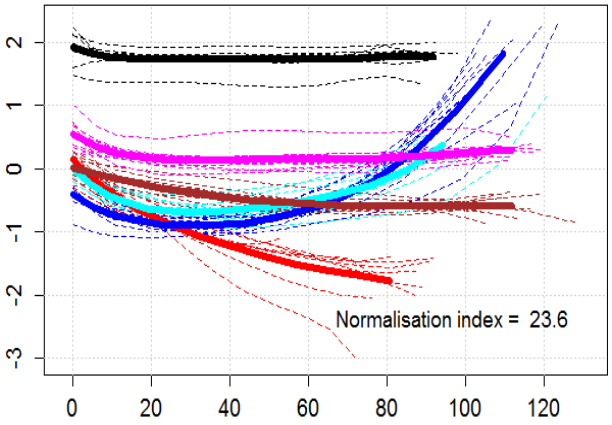

Figure 6: *Declination adjusted z-score normalisation for Cantonese unstopped tones = X-axis = normalised duration (csec.) Y axis = normalised F0 (sds).*

4. Summary

This paper has used two phonetic commonplaces – F0 declination and between-speaker differences in acoustic output – to show that normalisation of tonal F0 can be improved a little by taking into account, with very simple modeling, what is probably occasion-specific decay in tonal F0. Whether this is appropriately called declination is moot. The next questions to address are whether further improvement can be achieved by more accurate modeling of the decay – perhaps with Fujisaki parameters – and how to estimate an appropriate slope in tonal systems (like Shanghai) that lack tonologically level tones.

5. Acknowledgements

Very many thanks to my three anonymous referees whose comments let me see where I was not making myself adequately clear. I have tried where possible to incorporate their advice.

6. References

[1] D.R. Ladd, "Declination: a review and some hypotheses," *Phonology Yearbook*, vol. 1, pp. 53–74, 1984.

[2] P. Rose, "Where the science ends and the law begins: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud," *Int'l Journal of Speech Language and the Law*, pp. 277–324, 2013

[3] Editing Team, 水滸傳 *Water Margin*, Mandarin Audio-Visual and Resources Centre of Hong Kong, no date.

[4] H. Fujisaki, "Dynamic Aspects of Voice Fundamental frequency in Speech and Singing," in F. MacNeilage [Ed], *The Production of Speech*, pp. 39–55, Springer, 1983.

[5] H. Fujisaki, S. Ohno, and W.T. Gu, "Physiological and Physical Mechanisms for Fundamental Frequency Control in Some Tone Languages and a Command Response Model for generation of Their F0 Contours," International Symposium on Tonal Aspects of Languages, Beijing, 2004.

[6] H. Fujisaki, K. Hirose, P. Hallé, and H. Lei, "Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese," in *ICSLP 1990 - 1st Int'l Conf. on Spoken Language Processing, Kobe, Proceedings, 1990*, pp. 841–844, 1990.

[7] H. Fujisaki, "In Search of Models in Speech Communication Research," in *INTERSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association, Brisbane, Proceedings*, 2008, pp. 1–10.

[8] P. Rose, "Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study', in M. Barlow [Ed] *SST 2000 – 8th Australian Int'l Speech Science and Technology Conference*, *Canberra, Proceedings*, 2000, pp. 198–203.

[9] https://philjohnrose.net/numerical_data/index.html

[10] P. Rose, "Likelihood ratio-based forensic voice comparison with higher level features: research and reality," in E. Lleida & L. J. Rodriguez-Fuentes [Eds] *Recent Advances in Speaker and Language Recognition and Characterisation*, pp. 476–502, *Computer Speech and Language,* Special Issue, 2017.

[11] P. Rose, "A Linguistic-Phonetic Acoustic Analysis of Shanghai Tones", *Australian Journal of Linguistics*, vol. 13, pp. 185–220, 1993.

[12] W. Steed and P. Rose, "Same tone, different category: linguistic-tonetic variation in the areal tonal acoustics of Chu-qu Wu", in *INTERSPEECH 2009 - 8th Annual Conference of the International Speech Communication Association, Brighton, Proceedings*, 2009, pp. 2295–2298.

[13] P. Rose, "Oujiang Wu tones and Acoustic Reconstruction," in C. Bowern, B. Evans, and L. Miceli [Eds.], *Morphology and Language History,* pp. 235–250, John Benjamins, 2008.

[14] P. Rose, "Some considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech Communication*, vol. 6, no. 4, pp. 343–352, 1987.

[15] P. Rose, "Comparing Normalisation Strategies for Citation Tone F0 in Four Chinese Dialects," in C. Carignan & M. D. Tyler [Eds], *SST 2016 - 16th Australasian Int'l Conf. on Speech Science & Technology*, *Sydney*, 2016, pp. 221–224.

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