A Comparison of Tone Normalization Methods for Language Variation Research

Jingwei Zhang
Department of Chinese Language and Literature, University of Macau
Avenida da Universidade, Taipa, Macau, China
jwzhang@umac.mo

Abstract

One methodological issue in tonal acoustic analyses is revisited and resolved in this study. Previous tone normalization methods mainly served for categorizing tones but did not aim to preserve sociolinguistic variation. This study, from the perspective of variationist studies, re-evaluates the effectiveness of sixteen tone normalization methods and finds that the best tone normalization method is a semitone transformation relative to each speaker's average pitch in hertz.

1 Introduction

In comparison with the large amount of work done on vowel normalization (see van der Harst, 2011 p. 90-107 for an overview), tone normalization, or more specifically, the normalization of the fundamental frequency associated with linguistic tone, has received relatively little attention (Rose, 1987). The goal of tone normalization is to eliminate anatomical variation between speakers and allow between-speaker comparison. The anatomical variation resides in the frequency variation mainly caused by different sizes of the vocal folds. Although the variation of vocal fold sizes is the source of the variation of fundamental frequencies, listeners are nevertheless able to neutralize such anatomy-based acoustic differences while retaining phonemic differences (ibid.). Hence, to accurately model the perception of tones by the human auditory system, these anatomy-based differences need to be removed by a proper normalization method. In variationist studies, an effective technique for the normalization of vowels is one that (1) preserves phonemic variation, (2) minimizes anatomical variation, and (3) preserves sociolinguistic variation (Adank, 2003; Adank et al., 2004; van der Harst, 2011). These criteria can be applied to tone normalization as well.

Previous vowel normalization studies suggested that different normalization methods serve different goals (Disner, 1980; Thomas, 2002; Thomas, 2011). Section 2 reviews the aforementioned criteria for evaluating normalization methods and introduces a comparison method used in vowel normalization that has been adapted for tone. In Section 3, 16 normalization methods, ten existing ones and six variants, are briefly described and compared. Section 4 presents the comparison results among 16 methods. Section 5 concludes with a discussion of which normalization method is the best for tonal variationist research.

2 Literature Review

2.1 Previous evaluations of tone normalization methods

Rose (1987) provided the first quantitative comparison of tone normalization methods by evaluating both Z-score and Fraction of Range (FOR) by the method Dispersion Coefficient (DC, equation 1). The DC is the ratio of mean between-speaker variance to overall sample variance and is a measure of the degree to which speakers' values cluster. Rose (1987) found that Z-score is preferable since it has a smaller DC, which indicates a better convergence result.

(1) $$DC = \frac{mean\ between\text{-}spea \quad and\ within\text{-}normalization\text{-}point\ variance}{sample\ variance}$$

32nd Pacific Asia Conference on Language, Information and Computation
Hong Kong, 1-3 December 2018

Zhu (1999) followed Rose's study and further developed four normalization methods by Z-score and FOR: proportion of range (POR), ratio of log semitone distances (LD), logarithmic Z-score (LZ), and logarithmic proportion of range (LPOR). Zhu (1999) compared six methods by a Normalization Index (NI, equation 2) on the data of the Shanghai dialect and suggests the LZ method as the best method. Before describing these six methods, we first review the process of comparison as it is key to the selection of methods.

$$(2)\ NI = \frac{DC\ of\ F0\ values\ before\ normalization}{DC\ of\ normalized\ values}$$

An NI is defined as a ratio between DC of F_0 values before normalization and DC of normalized values (Rose, 1987; Zhu, 1999, p. 49). A large NI indicates that normalized F_0 contours cluster tightly. When tonal studies (e.g. Rose, 1987; Zhu, 1999) prefer normalization methods yielding larger NIs, it suggests that the aim was to categorize tones. High NI priority implies that a good normalization method can eliminate phonetic differences if they are not phonemic contrasts. In sum, the NI functions well for studies drawing up the tone inventory of a given language; the method with the largest NI is preferred.

However, tonal variationist studies not only aim to categorize tones but also aim to record the intermediate status between points on some scale between phonological and phonetic. If a tone is involved in a phonetically gradual change in progress, a proper variationist normalization method would keep the phonetic differences that are eliminated by high-NI normalization methods. In addition to phonemic and anatomical information, the F_0 of a tone carries a great deal of sociolinguistic information, like the speaker's regional background, socioeconomic class, and ethnicity. An inherent flaw of NI is that it is unable to preserve sociolinguistic information while evaluating the success of normalization methods. Consequently, NI is not fit to evaluate and then select the normalization methods for variationist studies.

2.2 Using discriminant analysis to evaluate tone normalization methods

Recent studies of vowel normalization (Adank, 2003; Adank et al., 2004; van der Harst, 2011) used linear discriminant analysis (LDA) to compare different vowel normalization methods to select the most successful one for fulfilling three criteria: (1) preserving phonemic variation, (2) minimizing anatomical variation, and (3) preserving sociolinguistic variation. LDA is a commonly used technique for data classification (Balakrishnama and Ganapathiraju, 1998). It maximizes the ratio of between-class variance to the within-class variance in order to guarantee the maximal separability. LDA generates models on the basis of the group variable and the independent continuous predictor(s). It can predict the group membership of each token from the model and then calculate the model's accuracy of prediction for the actual classification.

The methodological differences of the present study from those vowel studies are the predictor(s) of LDA. Vowel studies use values of F_0, F_1, F_2, and F_3 at a specific point during the vowel because these formant values are crucial to the classification of vowel categories and are independent indexes. The main goal of tonal LDA, however, is to predict tonal categories. In order to differentiate tones, information like pitch height, the direction of pitch movement (contour shape), and duration are crucial. One assumption of LDA is that predictors are independent. Pitch height values on the same tonal contour are strongly correlated with each other so they cannot be predictors at the same time. Thus, the information conveyed by the F_0 values of the tone contour (i.e. pitch height, contour shape, and duration) has to be transformed into independent predictors. Previous studies, like Zhu 1999 and Andruski and Costello 2004, have shown that polynomial equations are effective for synthesizing pitch contours. A polynomial equation is an appropriate tool for describing tones quantitatively. The degree of the polynomial equation depends on the complexity of tonal contours.

The current study uses a dataset of Wu dialects to test the different tone normalization methods. The tonal contour in Wu has maximally one turning point for the horn-shaped, peaking or dipping curves (Liu, 2015). The parabola with an equation of $y=a+bx+cx^2$ is a 2nd degree polynomial equation and the appropriate fitting curve for the Wu tones. A-, b-, and c-coefficients obtained from the quadratic equation can be used as continuant independent variables for studying the tonal variation. Each of the quadratic coefficients represents one characteristic of the parabola. They can provide the information of pitch height as well

as the contour shapes, so they can be used as predictors for LDA. Another independent predictor is the tonal duration. In the present study, a stepwise LDA is performed to select the best predictor(s) among a-, b-, and c-coefficients and duration.

Vowel studies have used sex-related variation to represent anatomical variation and regional variation to represent sociolinguistic variation. If a normalization method can successfully remove the anatomical variation, the success rate for that method to predict speaker's sex would be at chance (i.e. 50%). If a normalization method can successfully preserve the sociolinguistic variation, the success rate for that method to predict the speaker's region would be close to 100%. In the present research, we ask if sex-related variation can represent anatomical variation and regional variation can represent sociolinguistic variation. We test these premises in our own dataset before doing LDA.

2.3 Dataset

This study uses a dataset of Wu dialects to evaluate tone normalization methods. The dataset was collected for the purpose of studying tonal variation in Wuxi and Shanghai dialects and includes 120 participants from 6 localities. The 6 localities are Wuxi urban area, Wuxi suburb Huazhuang, Shanghai urban area, and three Shanghai suburbs: Baoshan, Songjiang, and Nanhui. Twenty participants of each locality were further designated by age[1] (old vs. young) and sex (male vs. female). This design allows investigating the possible regional variation, sex variation, and age variation.

Each participant was directed to read monosyllable words, bisyllable words, a passage, and minimal pairs. Speech was recorded with a portable TASCAM DR-100 digital recorder and an AKG C420 headset microphone. The recordings were sampled at 48kHz (24 bits). The dialects spoken in those six localities have a comparable phoneme: the low rising tone. This study uses the data of words originally pronounced in the low rising tone, including their F_0 values and tonal duration, to quantify anatomical variation and sociolinguistic variation. The F_0 values of the tone contour of each word were measured at 11 equidistant points (P_0, P_1… P_{10}), resulting in a set of time-normalized F_0 values. Considering the effect "F_0 perturbations" (Rose, 1993) due to coarticulation, the first 10% of the tone contour was neglected (Rose, 1987; Stanford, 2008), so P_0 is excluded and only the F_0 values of P_1 to P_{10} were used for further analysis. The low rising tones were either transcribed as /13/ or /113/ in the dialectological studies. Because of a lack of obvious regional variation indicated by the transcription, it is not appropriate to use regional variation to index sociolinguistic variation. Another variable is needed. In the present study, the age cohort effect (old vs. young) is a possible. If ongoing changes exist, we would see the old generation typically using the conservative variant while the young generation uses the advanced variant. In this case, age-related variation can index sociolinguistic variation. However, age-variation also contains possible anatomical variation (Chatterjee et al., 2011). The anatomical F_0 changes caused by aging are mainly changes in pitch heights, so the variation of rising contour shapes constrained by age, rather than the variation of pitch height, could be a reliable index of sociolinguistic variation.

When treating sex-related variation as an anatomical variation, it is imperative that there be only anatomy-based variation between males and females in the testing data but no significant sociolinguistic variation. Like the age-related variation, the anatomically sex-related variation on F_0 is also the variation of pitch height and not associated with contour shape. Only the variation of rising contour shapes constrained by sex could be a reliable index of anatomical variation.

Therefore, a series of mixed models were conducted on the slope of low rising tones. The independent variables were age (young vs. old) and sex (male vs. female), with speaker (20 levels for each region) and word (33 levels in Wuxi urban and Huazhuang, 52 levels in Shanghai urban, 36 levels in Songjiang, Baoshan, and Nanhui) as random effects. In the rising tones, the slope is an indicator of pitch shape. Results of mixed models show that in Wuxi urban, Shanghai Songjiang, and Shanghai Nanhui there is no significant effect of sex on the slope variation, either as a main effect or in the interaction. This means that these three regions have

[1] Old speakers refer to speakers born before 1950, aged 60 and above at the time of data collection in 2010, while young speakers refer to speakers born between 1987 and 1992, aged between 18 and 23 in 2010.

purely anatomical variation. There is a significant effect of age on the slope variation: Wuxi urban (t=5.41, p=0.031), Huazhuang (t=2.42, p=0.028), and Shanghai Songjiang (t=3.62, p=0.002). In those three regions, sociolinguistic variation can be represented by age-related variation. Only dialects in Wuxi urban and Shanghai Songjiang show variation of tonal shapes between the old and young generations with no variation between males and females. The data of those two localities was chosen to test all of the tone normalization methods for the present study.

3 Description of tone normalization methods

Zhu (1999, p. 45-48) systematically described and compared six tone normalizations. To facilitate understanding, the equations of Zhu's six methods are cited with a brief description and numbered as Methods 1, 2, 3, 4, 6, and 7 in this study.

In the following equations (3) to (11), raw values of F_0 in hertz are represented by x_i, where i can take the value 1 to 10 for measuring points P_1 to P_{10}. $F_0^{Method\ X}$ of each equation stands for normalized value via "Method X"; using the superscript identifies the name of the normalization method.

Method 1: Z-score

Z-score values, $F_o^{Z\text{-}score}$, express "an observed F_0 value as a multiple of a measure of dispersion away from a mean F_0 value" (Rose, 1987, p. 347). Z-score is calculated as follows:

$$F_0^{Z-sco} = \frac{x_i - m}{s} \quad (3)$$

In equation (3), m is the mean value of x_i and s is the standard deviation, both calculated per speaker.

Method 2: Fraction of range (FOR)

F_0^{FOR} expresses "an observed F_0 value as a fraction of the difference between two range-defining F_0 values" (Rose, 1987, p. 347). The normalized F_0 value is calculated by equation (4).

$$F_0^{FOR} = \frac{x_i - x_L}{x_H - x_L} \quad (4)$$

In the Shanghai dialect, Zhu (1999, p. 49) defined x_H as the highest F_0 value of the high level tone /55/ and x_L as the lowest F_0 value of the low rising tone /13/ for each speaker, usually the speaker's highest and lowest F_0 values respectively (Zhu, 1999, p. 49).

Method 3: Proportion of range (POR)

F_0^{POR} also expresses F_0 value as a proportion of a range expressed by the mean (m) and standard deviation (s). The equation is:

$$F_0^{POR} = \frac{x_i - (m - cs)}{(m + cs) - (m - cs)} \quad (5)$$

In equation (5), c is a consonant. Zhu (1999) used c=1 and c=2 in calculating POR. As with Z-score, m is the mean value of x_i and both m and s are calculated per speaker.

Method 4: Ratio of log semitone distances (LD) & Method 5: T value

F_0^{LD} is the logarithmic version of Method 2: FOR. LD is calculated with equation (6). The x_H and x_L of LD are the same as in FOR.

$$F_0^{LD} = \frac{log_{10}^{x_i} - log_{10}^{x_L}}{log_{10}^{x_H} - log_{10}^{x_L}} \quad (6)$$

Shi and Wang (2006) developed an adapted version of LD: T value (F_0^T). The T value has been used in the Chinese literature on tone more than any other normalization method, as T values range from 1 to 5 and match Chao's 5-point scale (Chao 1968). The T value is calculated as:

$$F_0^T = 5 \times \frac{\log_{10}^{x_i} - \log_{10}^{x_{min}}}{\log_{10}^{x_{max}} - \log_{10}^{x_{min}}} \quad (7)$$

In equation (7), x_{max} is the highest F_0 value of one speaker rather than the constant P_{max}; x_{min} is the lowest F_0 value and also different from the constant P_{min}. T value uses each speaker's two extreme F_0 values to make sure the result ranges from 0 to 5. T values between 0 and 0.99 are converted to Chao's tone letter 1, values in the range of 1 to 1.99 match Chao's tone letter 2, and so on.

Method 6: Logarithmic Z-score (LZ)

F_0^{LZ} transform is the logarithmic version of the Z-score. The equation for calculating F_0^{LZ} is:

$$F_0^{LZ} = \frac{y_i - m_y}{s_y} = \frac{\log_{10}^{x_i} - \frac{1}{n} \times \sum_{i=1}^{n} \log_{10}^{x_i}}{\sqrt{\frac{1}{n-1} \times \sum_{i=1}^{n} \left(\log_{10}^{x_i} - \frac{1}{n} \times \sum_{i=1}^{n} \log_{10}^{x_i} \right)^2}} \quad (8)$$

As shown in equation (8), $y_i=log_{10}{}^{xi}$, m_y and s_y are the mean and standard deviation of y_i (i= 1, 2, 3 … 10) respectively.

Method 7: Logarithmic proportion of range (LPOR)

$F_0{}^{LPOR}$ is the logarithmic version of Method 3: POR. It is calculated via equation (9).

$$(9) \qquad F_0^{LPOR} = \frac{y_i-(m_y-cs_y)}{(m_y+cs_y)-(m_y-cs_y)}$$

y_i, m_y, and s_y in equation (9) are the same as in equation (8), $y_i=log_{10}{}^{xi}$, m_y and s_y are the mean and standard deviation of y_i. Like in POR, c is a consonant in LPOR.

$F_0{}^{FOR}$, $F_0{}^{LD}$, and $F_0{}^{T}$ values transform the original F_0 to a value relative to the range between x_H and x_L or between x_{max} and x_{min}, two range-defining (R-D) points. R-D points should avoid tonal variation, otherwise the range could vary due to the tonal variation in addition to the anatomical differences. Rose (1987) pointed out that using this method should avoid the circularity of forcing congruence as R-D points are assumed to be equivalent among the speakers. However, their equivalence will only be clear after normalization if no external criterion is applied for evaluation beforehand. One possible external criterion is that R-D points are speaker-constants by their low within-speaker variance. In sum, the definition of R-D points is crucial. Their selection should meet the following requirements: (1) they must have consistent linguistic meanings across speakers to avoid between-speaker variations and (2) they must be speaker-constant values to avoid within-speaker variations.

Apart from the six methods reviewed by Zhu (1999), some tonal studies have used semitone scales to match human pitch perception (see Stanford 2016 for a review). The basic interval for pitch perception is the octave because the human auditory system perceives tones in a logarithmic way rather than a linear way. Equation (10) is used to transform hertz into semitone.

$$(10) \qquad F_0^{\text{ST-ref}} = \frac{12}{\log_{10} 2} \times \log_{10} \frac{x_i}{\text{ref}}$$

Method 8: ST-100	ref=100 Hz
Method 9: ST-AvgF₀	ref=$AvgF_0$ (i.e. each speaker's average pitch)
Method 10: ST-x_L	ref=x_L (i.e. the mean of speaker-constant P_{min})
Method 11: ST-x_H	ref=x_H (i.e. the mean of speaker-constant P_{max})
Method 12: ST-$\frac{x_H+x_L}{2}$	ref=$\frac{x_H+x_L}{2}$
Method 13: ST-x_{min}	ref=x_{min} (i.e. the P_{min} of each speaker's data)
Method 14: ST-x_{max}	ref=x_{max} (i.e. the P_{max} of each speaker's data)
Method 15: ST-$\frac{x_{max}+x_{min}}{2}$	ref=$\frac{x_{max}+x_{min}}{2}$

Many studies use the reference value of 100 Hz, thus getting $F_0{}^{ST\text{-}100}$=0 at 100 Hz, $F_0{}^{ST\text{-}100}$=12 at 200 Hz, and $F_0{}^{ST\text{-}100}$=−12 at 50 Hz. In this case, $F_0{}^{ST\text{-}100}$ is ***Method 8***. The change of reference value will not change the pitch shape. However, the semitone scale, centered to a fixed value, cannot reduce any between-speaker variance. Some studies (e.g. Chen, 2008; Howard, 1997) used split references for males and females to reduce some sex-based differences in physiology, like 50 Hz for male and 100 Hz for female. These choices of split references require further justification.

In addition to the fixed references, Andruski and Costello (2004) converted hertz to semitones relative to each speaker's average pitch, that is $F_0{}^{ST\text{-}AvgF0}$ (***Method 9***). Each speaker has their reference based on their production data. This method makes each speaker's data comparable by not only the slope but also the pitch height. This method needs adjusting for the variationist study as well. If the reference is individual average pitch, the average pitch itself is inevitably affected by the tonal variation. A reference is needed that (1) is relatively independent of the phonological variation and (2) reflects anatomical differences and other style differences (relaxed/stressed, excited/calm).

Different unfixed references form several different methods. ***Method 10*** is semitone transformation centered to x_L, the mean of the lowest F_0 of the low tone in the dialects examined. ***Method 11*** uses the speaker-constant P_{max}, the mean of highest F_0 of the high-level tone. ***Method 12*** uses the mean of x_H and x_L. ***Method 13*** is semitone transformation centered to x_{min} of each speaker; ***Method 14*** centered to x_{max} while ***Method 15*** centered to the mean of x_{min} and x_{max}.

Besides a series of semitone transformations, intonation research also uses mel, bark, and ERB-rate as psycho-acoustic scales. For frequencies below 500 Hz (i.e. the main region for F_0 of the speech signal), mel and bark are nearly linear transformations of F_0 in hertz (Nolan, 2003), so they will not be discussed in this study. The ERB scale (Equivalent Rectangular Bandwidth) was proposed by Moore and Glasberg (1983). The equation used in the present study was proposed and shown in Glasberg and Moore 1990. Its transformation is logarithmic at higher frequencies but between linear and logarithmic below 500 Hz (ibid.). This method is indexed as ***Method 16*** in the present study and calculated by equation (11).

(11) $F_0^{ERB} = 21.4 \times log_{10}(0.00437 \times F_0 + 1)$

4 Results

Following Adank and colleagues (2004) and Van der Harst (2011), the sixteen normalizations are first investigated to determine the extent to which they preserve phonemic variation (Section 4.1). Second, the normalizations are tested for their ability to minimize anatomical variation (Section 4.2). Finally, it is examined to what extent the normalizations preserve sociolinguistic variation (Section 4.3). All three criteria are evaluated by Linear Discriminant Analyses (LDAs). In these analyses, a certain category (i.e. toneme, age or sex) is predicted using a-, b-, and c-coefficients from the quadratic equation and tonal duration as predictors. The quadratic equations are calculated with the normalized or raw F_0 values. Section 4.4 discusses which method is optimal for the variationist study.

4.1 Preserving phonemic variation

LDA 1 was conducted in Wuxi and Songjiang separately to investigate the extent to which the normalizations preserve phonemic variation. Besides low rising tone /13/ (Wuxi 845 tokens, Songjiang 655 tokens), a few variants are also found in the impressionist transcription: peaking tone /131/ (Wuxi 60 tokens, Songjiang 266 tokens) and falling tone (Wuxi 27 tokens, Songjiang 34 tokens). In both regions, three tonal categories are predicted by LDA 1. A stepwise LDA was conducted to choose among four potential predictors: a-, b-, and c-coefficients and tonal duration. Table 3 presents the accuracy (abbreviated as ACCY in the table) and rank for Wuxi LDA 1 and Songjiang LDA 1.

Table 3 Results for Wuxi and Songjiang LDA 1 (stepwise): percent correctly classified tonal shapes based on quadratic coefficients predictors: a, b, and c and tonal duration

	method	Wuxi		Songjiang	
		ACCY	rank	ACCY	rank
0	Hz	95.4	16	86.4	17
1	Z-score	95.5	13	93.5	2
2	FOR	95.4	16	93.0	5
3	PORc1	95.5	13	93.5	2
	PORc2	95.5	13	93.5	2
4	LD	95.4	16	94.1	1
5	T	95.9	9	92.8	6
6	LZ	95.7	12	89.4	15
7	LPORc1	95.8	10	90.3	14
	LPORc2	95.8	10	90.3	14
8	ST-100	96.2	1	91.3	11
9	ST-AvgF_0	96.2	1	91.7	9
10	ST-x_L	96.0	7	91.2	12
11	ST-x_H	96.2	1	92.3	7
12	ST-$\frac{x_H+x_L}{2}$	96.2	1	91.4	10
13	ST-x_{min}	96.0	7	91.2	12
14	ST-x_{max}	96.2	1	93.2	4
15	ST-$\frac{x_{max}+x_{min}}{2}$	96.2	1	91.8	8
16	ERB	95.5	13	89.2	16

All of the normalization methods have a high overall accuracy of prediction which is higher than that of the raw F_0 in hertz. The difference between the highest and lowest accuracy is less than 5% (Songjiang: 94.1%-89.2%=4.9%).

4.2 Minimizing anatomically-based variation

To test the extent to which the normalization methods minimize sex-related anatomically-based variation, LDA 2 was conducted for the rising tones in Wuxi and Songjiang separately. If a normalization method successfully removes the variation, the success rate for that method will be at chance level (i.e. 50%). The results are presented in Table 4.

In Table 4, LDA 2 was not completed for the methods Z-score, PORc1, and PORc2 for Wuxi data; they are marked with "-" in the table. These methods removed almost all of the sex-related anatomically-based variation, making further analyses not computable. For the sake of method

comparison, their predicted accuracy of sex will be calculated as chance level (50%).

Table 4 Results for Wuxi and Songjiang LDA 2 (stepwise): percent correctly classified tonal shapes based on quadratic coefficients predictors: a, b, and c and tonal duration

	method	Wuxi		Songjiang	
		ACCY	rank	ACCY	rank
0	Hz	90.3	15	89.9	17
1	Z-score	-	1	69.4	8
2	FOR	56.2	5	70.9	12
3	PORc1	-	1	69.4	8
	PORc2	-	1	69.4	8
4	LD	57.1	10	70.2	11
5	T	60.3	12	68.6	7
6	LZ	55.1	4	73.9	13
7	LPORc1	54.9	3	73.9	13
	LPORc2	54.9	3	73.9	13
8	ST-100	90.3	15	89.1	15
9	ST-AvgF_0	56.7	6	66.9	1
10	ST-x_L	59.2	11	68.2	6
11	ST-x_H	61.5	13	66.9	1
12	ST-$\frac{x_H+x_L}{2}$	56.7	6	67.8	3
13	ST-x_{min}	64.9	14	67.8	3
14	ST-x_{max}	56.7	6	69.4	8
15	ST-$\frac{x_{max}+x_{min}}{2}$	56.7	6	67.8	3
16	ERB	90.5	17	89.1	16

In the LDA 2 procedure, the raw values carry a great deal of anatomically-based variation, as evidenced by the high success rate for predicting speaker sex (around 90% in both regions). The method showed similar results with ST-100 and ERB regarding the raw values, which carried much anatomically-based variation. Apart from ST-100 and ERB, other normalizations removed the variation successfully. In Wuxi data, Z-score, PORc1, and PORc2 performed best, whereas ST-x_{min} (64.9%) and ST-x_H (61.5%) removed the variation to a lesser extent. In Songjiang, ST-AvgF_0 and ST-x_H predicted speaker sex closest to chance level (both 66.9%), whereas LZ, LPORc1, and LPORc2 removed the variation to a lesser extent (73.9%). In general, different normalizations, except for ST-100 and ERB, do not have large differences in predicting speakers' sex, they all performed well in removing sex-related anatomically-based variation.

4.3 Preserving sociolinguistic variation

To investigate the extent to which normalization methods preserve sociolinguistic variation, that is age-related variation, LDA 3 was conducted to predict whether the rising tones were spoken by the old people or by the young people in Wuxi and Songjiang. If a method shows a success rate significantly above chance level 50%, it shows great ability to preserve sociolinguistic variation. The results are given in Table 5.

Table 5 Results for Wuxi and Songjiang LDA 3 (stepwise): percent correctly classified tonal shapes based on quadratic coefficients predictors: a, b and c and tonal duration

	method	Wuxi		Songjiang	
		ACCY	rank	ACCY	rank
0	Hz	75.7	7	89.9	6
1	Z-score	72.6	14	87.7	11
2	FOR	74.4	10	85.6	15
3	PORc1	72.6	14	87.7	11
	PORc2	72.6	14	87.7	11
4	LD	74.6	8	85.9	14
5	T	71.9	18	87.4	13
6	LZ	73.7	11	85.6	15
7	LPORc1	73.5	12	85.4	17
	LPORc2	73.5	12	85.4	17
8	ST-100	67.3	19	90.4	5
9	ST-AvgF_0	86.5	3	93.6	3
10	ST-x_L	76.0	6	89.8	9
11	ST-x_H	86.4	4	95.7	1
12	ST-$\frac{x_H+x_L}{2}$	84.4	5	94.1	2
13	ST-x_{min}	72.5	17	89.9	6
14	ST-x_{max}	87.3	1	91.2	4
15	ST-$\frac{x_{max}+x_{min}}{2}$	86.6	2	89.6	10
16	ERB	74.6	9	89.9	6

Table 5 presents some interesting results. First, the raw F_0 in hertz preserves most age-related variation, better than the majority of normalization methods (rank 7 in Wuxi and rank 6 in Songjiang). This is reasonable since age-related variation is the steepness variation of rising tone, which is the variation of pitch shape, which should be reflected by the raw F_0. Second, considering the performance of raw F_0 as a baseline, the normalization methods can be split into two groups: the group preserving less sociolinguistic variation than the baseline and the group preserving sociolinguistic variation more than or equal to the baseline. As evidenced by Table 5, most methods of semitone transformation can preserve more sociolinguistic variation than baseline whereas the six methods presented in Zhu

1999, as well as T values, all perform poorly in comparison. It is not surprising that those seven methods (i.e. Z-score, FOR, POR, LD, T, LZ, and LPOR) cause attrition of sociolinguistic variation because their transformations are principally composed of two steps: (1) parallel shift of tones on the coordinate using (logarithmic) mean or (speaker-constant) P_{min} (i.e. x_{min} or x_L) as reference and (2) range compression or expansion based on standard deviation or R-D range[2] (c.f. Zhu, 1999, p. 46[3]). The compression or expansion of the range will consequently reduce or enlarge the within-speaker variance, causing a change of speaker weight in the cross-speaker variance (i.e. the sociolinguistic variation). These methods cannot preserve as much sociolinguistic variation as the baseline.

The group of normalizations preserving sociolinguistic variation more than the baseline are mainly semitone transformations centered to varying references. They create a parallel shift of tones on the coordinate but do not result in range compression or expansion. Comparing the results of Wuxi and Songjiang, only ST-$AvgF_0$, ST-x_H, ST-$\frac{x_H+x_L}{2}$, and ST-x_{max} show better performance than the baseline in both regions. Therefore, it appears that the best method for tone normalization will be one of these.

4.4 Summary of the LDA results: the selection of ST-AvgF0

In the previous sections, LDA results were presented concerning the preservation of phonemic variation and sociolinguistic variation and the reduction of anatomically-based variation. LDA 1 results in Table 3 show that all of the normalization methods have a high overall accuracy of predicting tonal shapes. LDA 2 results in Table 4 show that, apart from ST-100 and ERB, all of the normalization methods performed well in removing sex-related variation. However, LDA 3 results in Table 5 suggest that the series of semitone normalizations perform better than the other normalization methods. Specifically, ST-$AvgF_0$, ST-x_H, ST-$\frac{x_H+x_L}{2}$, and ST-x_{max} show better performances than the baseline in both regions. Since the three aims of the normalization methods are equally important, LDA 3 results help to reduce the scope of comparison into four methods, ST-$AvgF_0$, ST-x_H, ST-$\frac{x_H+x_L}{2}$, and ST-x_{max}. They outperform the baseline hertz in all respects.

The four methods are evaluated for maximum generalizability and stability in converting hertz to semitones. Among the four methods, ST-x_H and ST-$\frac{x_H+x_L}{2}$ use x_H and/or x_L in calculation. x_H and x_L are not simply P_{max} and P_{min} of a given speaker but P_{max} and P_{min} in a particular tone. This means that different languages can have different definitions for x_H and x_L, ultimately complicating cross-linguistic comparison. Comparing the stableness of x_{max} – the highest F_0 value of one speaker – and $AvgF_0$, we find that x_{max} is more easily influenced by the falsetto register, thus being less stable than $AvgF_0$.

For these reasons, ST-$AvgF_0$, or the semitone transformation relative to each speaker's average pitch, is the best normalization method.

5 Conclusion

In order to make accurate statements about between-speaker differences in tone, the fundamental frequency associated with the linguistic tone needs to be normalized. Previous tone normalization methods mainly served to categorize tones, but did not aim to preserve sociolinguistic variation. It is necessary to re-evaluate the effectiveness of tone normalization methods from the perspective of variationist studies. Following the sociophonetic studies of vowel variation (Adank, 2003; Adank et al., 2004; van der Harst, 2011), three criteria were used to evaluate a tone normalization method: (1) preserves phonemic variation, (2) minimizes anatomical variation, and (3) preserves sociolinguistic variation. The current study compared sixteen normalization methods by linear discriminant analysis (LDA).

The results show that ST-x_H (a semitone transformation relative to the mean of speaker-constant P_{max} in hertz), ST-x_{max} (a semitone transformation relative to the P_{max} in hertz of each speaker's data), and ST-$AvgF_0$ (a semitone

[2] R-D range is the range between two range-defining points (R-D). R-D points can be either x_{max} and x_{min} or x_H and x_L.

[3] Zhu (1999, p. 46) pointed out that Z-score, POR, LD, and LZ are composed of these two steps. According to the equations of FOR, LPOR, and T, these three transformations are also composed of these two steps.

transformation relative to each speaker's average pitch in hertz) are the top three normalization methods. Considering cross-linguistic comparison and the stability of x_H (the mean of speaker constant P_{max} in hertz), x_{max} (the P_{max} value in hertz of each speaker's data), and $AvgF_0$ (each speaker's average pitch in hertz), ST-$AvgF_0$ was found to be the best tone normalization method and hence was used to normalize all the F_0 values in this study.

Acknowledgments

The author would like to thank Utrecht Institute of Linguistics of Utrecht University, Chinese Scholarship Council and the University of Macau (Start-up Research Grant SRG2018-00131-FAH) for supporting this study. Thanks also go to Prof. René Kager and Dr. Hans van de Velde for their very helpful comments and suggestions.

References

Adank, Patricia Martine. 2003. Vowel Normalization: A Perceptual-Acoustic Study of Dutch Vowels. Ph.D. dissertation of Radboud University.

Adank, Patricia Martine, Smits, Roel & Van Hout, Roeland. 2004. A Comparison of Vowel Normalization Procedures for Language Variation Research. The Journal of the Acoustical Society of America, 116(5): 1729–1738.

Andruski, Jean E, and Costello, James. 2004. Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong. Journal of the International Phonetic Association, 34(2): 125-140.

Balakrishnama, S. and Ganapathiraju, A. 1998. Linear discriminant analysis - A brief tutorial. [Online]. Available: https://www.isip.piconepress.com/publications/reports/1998/isip/lda/lda_theory.pdf. Accessed on 30 September 2018.

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese.* Berkeley: University of California Press.

Chatterjee, Indranil, Halder, Hindol, Bari, Sayani, Kumar, Suman and Roychoushury, Amitabha. 2011. An Analytical Study of Age and Gender Effects on Voice Range Profile in Bengali Adult Speakers using Phonetogram. International Journal of Phonosurgery and Laryngology, 2(1): 65-70.

Chen, Yiya. 2008. Revisiting the phonetics and phonology of Shanghai Tone Sandhi. Proceedings of Speech Prosody 2008. Campinas, Brazil.

Disner, Sandra Ferrari. 1980. Evaluation of Vowel Normalization Procedures. The Journal of the Acoustical Society of America, 67(1): 253-261.

Glasberg, Brian R. and Moore, Brian CJ. 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Research, 47(1): 103-138.

Howard, David M. 1997. Practical Voice Measurement. In Harris T., Harris S., Rubin J.S. and Howard D.M. (Eds.), The Voice Clinic Handbook: 323-382. London: Whurr Publishing Company.

Moore, Brian CJ. and Glasberg, Brian R. 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The Journal of the Acoustical Society of America, 74: 750.

Rose, Phil. 1987. Considerations in the Normalisation of the Fundamental Frequency of Linguistic Tone. Speech Communication, 6(4): 343-352.

Rose, Phil. 1993. A linguistic-phonetic acoustic analysis of Shanghai tones. Australian Journal of Linguistics 13: 185-220.

Stanford, James. 2008. A sociotonetic analysis of Sui dialect contact. Language Variation and Change, 20(3): 48-81.

Stanford, James. 2016 Sociotonetics using connected speech. Asia-Pacific Language Variation, 2(1): 409-450.

Thomas, Erik R. 2002. Instrumental Phonetics. In Chambers, J., Trudgill, P. & Schilling-Estes, N. (Eds.), The Handbook of Language Variation and Change: 168-200. Oxford, UK/Malden, MA: Blackwell.

Thomas, Erik R. 2011. Sociophonetics: An Introduction. Basingstoke, UK: Palgrave Macmillan.

van der Harst, Sander. 2011. The Vowel Space Paradox: A Sociophonetic Study on Dutch. LOT.

Zhu, Xiaonong 1999. Shanghai Tonetics, Lincom Europa.