NORMALISATION OF TONAL $F_{\emptyset}$ FROM LONG TERM $F_{\emptyset}$ DISTRIBUTIONS

Phil Rose
Department of Linguistics (Arts), Australian National University

ABSTRACT An attempt is described to ascertain whether the $F_{\emptyset}$ of 7 speakers' tones can be normalised using parameters from their long term $F_{\emptyset}$ distribution. It is shown that normalisation using long term mean and standard deviation is not as effective in reducing the between-speaker variance as with parameters derived from the tones themselves. However, the approach is still successful enough to be worth pursuing, and some suggestions for improvement are indicated.

INTRODUCTION

In a previous paper (Rose 1987), I examined some problems in the normalisation of tonal $F_{\emptyset}$, using data from 7 speakers of a language with 6 tones (Fig.1). It was shown how the between-speaker (B-S) differences in $F_{\emptyset}$ could be reduced by up to a factor of 13 using normalisation parameters (NPs) derived from the isolation tone values. However, several considerations indicated the desirability of using NPs derived independently of the tones. Among these considerations is the fact that normalisation strategies using NPs derived from the tones themselves are inherently circular: one has to have an idea beforehand of what data points to derive the NPs from, because inclusion of non-comparable data points can significantly bias the NPs. Yet which points are in fact comparable between speakers only emerges after a successful normalisation. For example, it is clear from a comparison of ZSC's tones 1 and 3 with the others' (Fig. 1) that her Z-score NPs of mean and standard deviation will be biased by the much lower $F_{\emptyset}$ offset in these tones. In order to derive unbiased NPs from ZSC's isolation tones (and thus achieve the large reduction in B-S variance of 13) it was in fact first necessary to artificially truncate her tones 1 and 3 by the substantial amount of 6-10 csec. (Rose 1987:349).

A more important reason why NPs should not be derived from the tones themselves lies in the potential use of normalised values to facilitate objective and quantified comparison between varieties in order to determine the nature of Linguistic Phonetic variation in acoustical tonal parameters. One dimension in which varieties can differ is the $F_{\emptyset}$ range of their tones (Rose 1985; Phuong 1981). Therefore using $F_{\emptyset}$ range derived from the tones themselves as a NP (either as such or in terms of standard deviation) will automatically obscure or obliterate these differences.

One possible source of independently derivable NPs is a speaker's long term $F_{\emptyset}$ distribution ($LTF_{\emptyset}D$). Jassem (1975) for example has already demonstrated, although not quantified, the normalisation of B-S differences in intonational $F_{\emptyset}$ in Polish using NPs of mean and standard deviation derived from $LTF_{\emptyset}Ds$. The aim of this paper is to investigate the possibility of extracting parameters for normalisation of tonal $F_{\emptyset}$ from $LTF_{\emptyset}Ds$. In particular, it is of interest to see whether the LT data provide more suitable values for the NPs of ZSC's tones than her isolation tone data (i.e. values that will show her to have a lower normalised offset in tones 1 and 3, and an earlier rise in tone 4, but otherwise the same contours as the other speakers (1).

METHOD

It was convenient to use the same 7 speakers (4 male, 3 female) as in the previous investigation (Rose 1987), since the mean values for the detailed $F_{\emptyset}$ time course of their isolation tones were already available, and it was known how well they could be normalised using NPs from the tones themselves. In addition, they represent a good trial for normalisation, since, as can be seen from Fig. 1, they show relatively large B-S differences in $F_{\emptyset}$.

In choosing material for the LT analysis, the most important criterion was taken to be maximum comparability between isolation and LT data. The LT data were therefore extracted from running speech which had been recorded in the same session as the isolation tones, and , like the isolation tones, had been read out in an unemotional manner from a prepared text. (One exception to this was JHM, for whom I did not have any suitable running speech data recorded in the same session as his

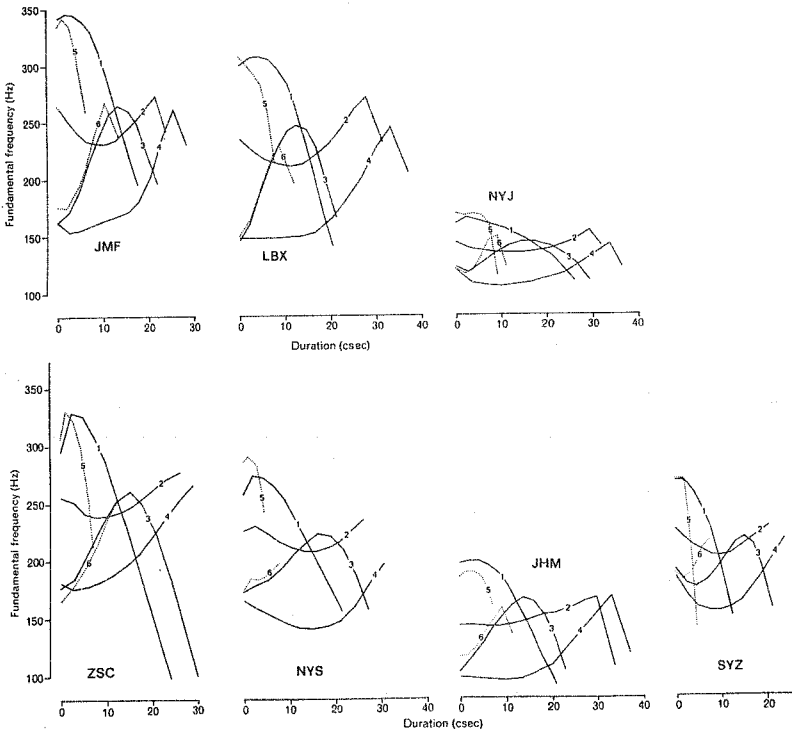

Figure 1. Fundamental frequency characteristics of the isolation tones of seven speakers of Zhenhai/Cixi dialect

isolation tones.) Apart from ZSC and SYZ, it was not possible to use the same text for all speakers, since the different speakers' isolation tones had been recorded over a period of 14 years, during which time the elicitation format had naturally changed. However, it was assumed that the nature of the texts was sufficiently homogeneous with respect to possible confounding variables to ensure B-S comparability. All texts were examples of unemotional narrative consisting of fairly short, grammatically well-formed utterances with unmarked intonation. There was thus a general absence of false starts and repair; of parenthetical intonation; and of contrastive and emphatic stress.

With the exception of SYZ, all texts were read at a tempo which sounded unhurried and natural. SYZ's tempo sounded rushed: her relatively faster tempo is shown by a comparison with ZSC, who took about 7 sec. longer to read the same text.

It was not clear how long the speech sample for the calculation of representative LT $F_{\emptyset}$ parameters should be. Nolan (1983:123) notes "a convergence of opinion that within-speaker variation between speech samples reduces with increasing sample length up to around one minute, and thereafter rather little". However, the proportion of voiced and voiceless segments in a short text differs considerably between languages - Catford (1977:107) cites a proportion of 78% for voiced segments in French, compared to 41% in Cantonese. Therefore an appropriate length of speech might depend on the language under investigation. Jassem (1975:525) considered 60 sec. of Polish sufficient to furnish adequate NPs. This is equivalent to about 33 sec. of voiced speech, given the proportion of 55% which he found for Polish (Jassem et al. 1973:210). In view of this, I decided simply to use all the running speech available. For 5 out of the 7 speakers, this supplied at least 33 sec. of actual voiced speech. For JHM, I only had a relatively short text with 24 sec. of voiced speech; the shortness of SYZ's text - 23 sec. of voiced speech - was due to her abnormally fast tempo (2).

It was decided to measure $F_{\emptyset}$ by hand from expanded narrow-band spectrograms rather than use an automatic digital $F_{\emptyset}$ extraction method. This was for three reasons, apart from the wish to avoid the occasional error in $F_{\emptyset}$ estimation which inevitably accompanies automatic extraction. This dialect has an unusual 'epiglottal' phonation type in tones 3 4 and 6, which is reflected in polychrotic time-domain waveforms (Rose in press: fn. 2). With automatic extraction, there is the danger that $F_{\emptyset}$ would be estimated from the associated subharmonics, thus yielding values between one-half and one-quarter of the effective $F_{\emptyset}$. The second reason was to maximize comparability with the isolation values, which had also been measured from expanded narrow-band spectrograms. Perhaps the most important reason for direct measurement is that in so doing one is able to observe potentially important relationships between $F_{\emptyset}$ values and linguistic units. (For example, B-S differences in prolongation of voicing during the hold phase of phonologically voiceless obstruents, or whether a speaker associates particular $F_{\emptyset}$ values with a specific tone, or even a specific word.)

$F_{\emptyset}$ was measured using a digitising pad in conjunction with the "pitch" algorithm developed by the MacQuarie University SHLRC centre, modified to allow up to 100 $F_{\emptyset}$ measurements per spectrogram. The estimated accuracy is +/- 5 Hz at the 90% confidence level. $F_{\emptyset}$ was sampled at 40 Hz (the same rate as the mean sampling rate of the isolation tones) using an overlay calibrated in 25 msec. strips. The sampled $F_{\emptyset}$ values were stored on disc and processed by the ILS HIS command to provide histograms and associated statistical data.

RESULTS

All speakers have positively skewed $F_{\emptyset}$ distributions. Although the 3rd and 4th moments were not calculated, it is possible to assume that deviations from normality are primarily in degree of skewness, because χ^2 values show on average a 50% better fit (range:33%-67%) to a log-normal curve than to a linear. Speakers appear to fall into 3 groups according to degree of assumed skewness: LBX, JMF, NYJ, with log-normal χ^2 values of 133, 122, and 119; ZSC, JHM, NYS with values of 65, 60, and 54; and SYZ, with 35.

Fig. 2 shows that a fairly clear relationship exists between the standard deviations and means of the LT data and those of the speakers' isolation forms. For 5 of the 7 speakers, the LT standard deviation does not differ significantly from that of the isolation tones. SYZ's 6.1 Hz significantly smaller standard

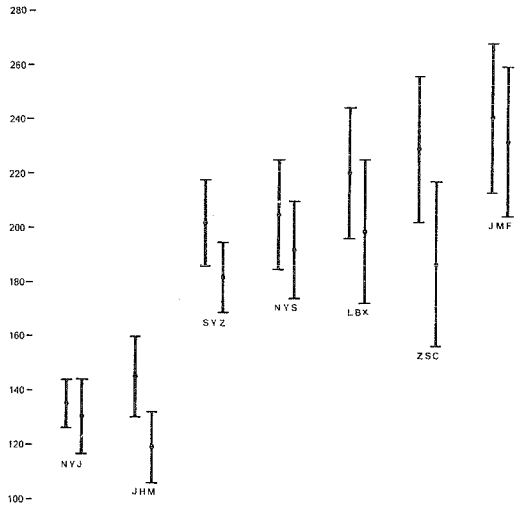

Figure 2. Long term fundamental frequency means and standard deviations of the seven speakers in Figure 1 (right) compared with the means and standard deviations of their isolation tones (left).

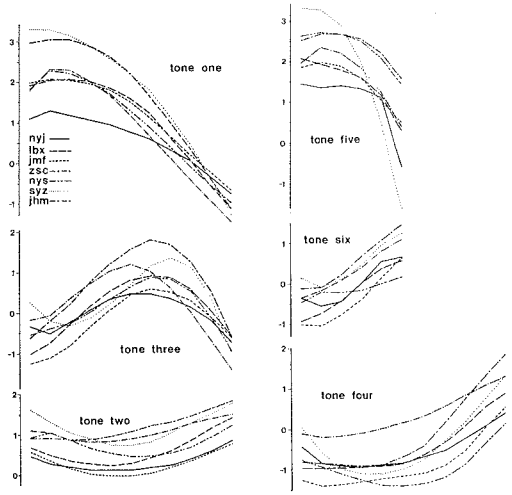

Figure 3. Z-score normalised fundamental frequency shapes for the seven speakers' tones plotted against equalised duration. Scale indicates units of long term standard deviation away from long term mean ($F_{\phi LTi} - \overline{F}_{\phi LT} / s_{LT}$)

deviation is perhaps related to her faster tempo; NYJ's 8.8 Hz significantly larger standard deviation may be related to a higher proportion of questions in his text.

For all of the speakers, the LT means are lower than the isolation means (JMF and NYJ are the only speakers whose LT and isolation means do not differ significantly). The relative amount by which the LT means are lower than the isolation means varies between 16% and 80% (mean = 48%) of a speaker's LT standard deviation. It is not clear why JHM and ZSC have relatively large differences (of 80% and 78% respectively). They may be related to the different factors of, in ZSC's case, her lower $F_{\emptyset}$ offset in tones 1 and 3, and, in JHM's case, the fact that his isolation tones were recorded in a different session from his LT material.

The isolation tone NPs of mean and standard deviation in a Z-score normalisation will reduce the B-S variance in the raw data by a factor of 7.5 (from a dispersion coefficient of 64.3% in the raw data to one of 8.6 for the normalised data (3). Because the LT mean and standard deviation bear a fairly constant relationship to the isolation NPs, it is to be expected that they will also constitute effective NPs. Fig. 3 shows the Z-score normalised $F_{\emptyset}$ curves using NPs of LT mean and standard deviation. It can be seen from a comparison with the raw data in Fig. 1 that this transform has in fact caused the raw $F_{\emptyset}$ shapes to cluster to a considerable extent. Quantitatively, this normalisation achieves a reduction in B-S variance of 3.2 (from 64.3% to 20.3%). This is therefore about half as good as the normalisation using NPs derived from the isolation tones themselves. As far as ZSC is concerned, the hopes for a better resolution from her LT parameters are not realised. Although her tones 1 and 3 are resolved with a slightly lower offset, the LT normalisation does not improve on the configuration resulting from a Z-score normalisation using NPs from the non-truncated isolation tones. It can be seen in fact that her LT mean was too low, because 4 of her tones appear to be resolved somewhat higher than average.

CONCLUSION

The results of this study are encouraging, in that they indicate that long term $F_{\emptyset}$ parameters do in fact constitute possible normalisation parameters for tonal $F_{\emptyset}$. Their use can therefore avoid the methodological objection of circularity.

Nevertheless, I think that the degree of reduction in B-S variance achieved by this particular LT normalisation must be considered inadequate on two counts. The normalised dispersion coefficient of 20.3% is still so large that the magnitude of the standard deviation around the mean normalised curves would almost certainly be too great to allow comparison across varieties. Also, although the identity of most of the tones of this variety is ensured by distinctive $F_{\emptyset}$ contours, the amount of scatter is such that there is an (admittedly very small) possibility of confusion between tones 2 and 4, which have similar contours but are separated by relative $F_{\emptyset}$ height. For example, the normalised values for ZSC's tone 4 after 50% of duration fall within one standard deviation below the mean of the normalised values for tone 2. In this connection, future normalisation studies could exploit this evaluation metric by using tone languages with a larger number of potentially confusable $F_{\emptyset}$ contours, for example Cantonese, which has 3 tones with quasi level $F_{\emptyset}$ shapes.

The magnitude of the dispersion coefficient for the LT data could reflect one or a combination of the following factors. (1) It may reflect true B-S variability in the relationship between LT parameters and isolation tones. In this case, some other source, or additional source of NPs must be sought. (2) It may reflect an inadequacy of the particular normalisation transform. Two other possible improvements spring to mind. Since it is known that the isolation tone NPs effectively reduce the B-S variance better than their LT counterparts, and this study has shown a relationship between LT and isolation values, one could try using LT-based estimates of the isolation NPs (for example LT standard deviation as best estimate of the isolation standard deviation, and LT mean plus 48% of LT standard deviation as best estimate of the isolation mean (48% is the mean amount by which the LT mean is lower than the isolation mean)). Since the LT $F_{\emptyset}$ distribution shows clear logarithmicality, and higher tonal $F_{\emptyset}$ values do not normalise as well as lower (at least in tone 1), the incorporation of a log transform might also contribute to an additional reduction in B-S variance. (3) Finally, perhaps my initial assumption of B-S comparability in LT texts was incorrect, and the relatively high scatter of normalised curves reflects lack of adequate control in selection of LT material. It is worth noting that the two speakers who show fairly extreme normalised values in all tones except 5 are precisely those who differed from the others in not

having the same recording session for LT and isolation tones (JHM), or using a faster tempo (SYZ). So perhaps a more accurate indication of the LT approach might be given by excluding the data for these two speakers. In any case, it would obviously be advisable to use a single text for all speakers, and perhaps try also to control for tempo. An investigation into the normalisation of Shanghai tone $F_{\emptyset}$ along these lines is already in progress.

NOTES

(1) ZSC's tone 4 has an audibly earlier rise in pitch than the others', and so this difference should be preserved in the $F_{\emptyset}$ normalisation. There is of course no optimal normalisation of her tones 1 and 3 without truncating them.

(2) Duration values for the individual speakers were (total duration of utterances (sec.); total duration of voiced speech (sec.); percent of voiced speech): LBX: 88; 68; 77%. JMF: 85; 65; 76%. NYS: 72; 57; 79%. NYJ: 39; 34; 87%. ZSC: 38; 33; 87%. SYZ: 31; 23; 74%. JHM: 30; 24; 80%.

(3) The dispersion coefficient (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. Comparing the DCs for the raw and normalised data provides a measure of the reduction in B-S variance achieved by a normalisation. The different values of 65.8% and 5.1% for the raw and normalised DCs given in Rose (1987:350) reflect a corpus containing ZSC's resampled truncated tones 1 and 3: If ZSC's tones 1 and 3 are not truncated, the efficacy of the Z-score normalisation is considerably reduced from 12.9 to 7.5, which is the appropriate value for comparison here.

REFERENCES

Catford, J.C. (1977) *Fundamental Problems in Phonetics*, (Edinburgh University Press).

Jassem, W. et al. (1973) 'Statistical Characteristics of Short-Term Average $F_{\emptyset}$ Distributions as Personal Voice Features', in W. Jassem (ed.) *Speech Analysis and Synthesis* Vol. 3, (Polish Academy of Science: Warsaw), 209-225.

Jassem, W. (1975) 'Normalisation of $F_{\emptyset}$ curves.' in G. Fant and M. Tatham (eds.) *Auditory Analysis and Perception of Speech* (Academic Press: London), 523-530.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition,* (Cambridge University Press).

Phuong, V.T. (1981) 'The acoustic and perceptual nature of Tone in Vietnamese', Ph.D. Thesis, Australian National University.

Rose, P.J. (1985) 'Comparing the Tones of Central and Southern Thai - Evidence from a Bilingual Speaker'. Paper at the 18th Intl. Conf. on Sino-Tibetan Languages and Linguistics, Bangkok.

Rose, P.J. (1987) 'Considerations in the Normalisation of the Fundamental Frequency of Linguistic Tone', *Speech Communication* 6, 343-352.

Rose, P.J. (in press)..'Tonology through Acoustic Phonetics - An Analysis of Disyllabic Lexical Tone Sandhi in Zhenhai [in Chinese], in Xu Baohua (ed.) *Zhao Yuanren "Xiandai Wuyude Yanjiu" chuban 60 nianji jinian zhuanhao. Wuyu Luncong Dierji.* [Special edition commemorating the 60th anniversary of the publication of Yuen Ren Chao's "Studies in the Modern Wu Dialects". Wu Dialect Papers Vol 2].