# THE NORMALISATION OF TONE

Phil Rose

Department of Linguistics
Australian National University

ABSTRACT - Some considerations in the normalisation of tone are discussed, and some problems in application demonstrated on the fundamental frequency data of 6 speakers of a variety of Wu Chinese.

## INTRODUCTION

The acoustic properties of the radiated speech wave are a unique function of a speaker's vocal tract anatomy, and since speakers' vocal tracts differ, so will their acoustic output - even for auditorily the same sound. The magnitude of between-speaker acoustical variance caused by physiological differences is often enough to swamp the linguistic content of the signal. The perception of this content has therefore to be mediated by a process which separates the Accentual and Linguistic content[1] of the acoustic stimulus from the components determined by the individual speaker's physiology. Normalisation is a mathematical analog of this perceptual process, two main aims of which are firstly to extract and specify the invariant acoustic correlates of the Accentual and Linguistic features within a particular variety, and then to compare varieties with respect to these correlates for typological and universal purposes (Disner 1980:253).

One major physiological source of between-speaker differences in acoustical output is the difference in size (length, mass) of the vocal cords. Such differences result in different preferred, or default values and ranges of the fundamental frequency (F0) of the radiated wave (Nolan 1983:51,59). F0 is the main acoustic correlate of perceived pitch, which functions as a dimension for suprasegmental linguistic systems of intonation, stress, and tone (Lehiste 1970). Thus female speakers, with shorter, less massive cords tend to have higher F0 values than males, and it is possible for a female's phonologically low tone to have a higher F0 than a male's phonologically high tone.

In contrast to the large amount of theoretical and empirical work done on vowel normalisation, very little attention has been paid to the question of normalising the acoustical correlates of suprasegmental categories of tone or intonation. This paper attempts to redress the balance a little, by examining some considerations of tonal normalisation using F0 data from a variety of Chinese.

## DATA

Fig. 1 shows raw F0 shapes of the same six phonemic tones of a variety of Chinese as spoken by six different native speakers under similar circumstances[2]. The $F_0$ shapes are plotted as functions of absolute duration and represent arithmetical mean values of several tokens (the exact number of tokens per tone per speaker is also shown in the figure). The corpus was controlled for the usual intrinsic effects, and consisted of

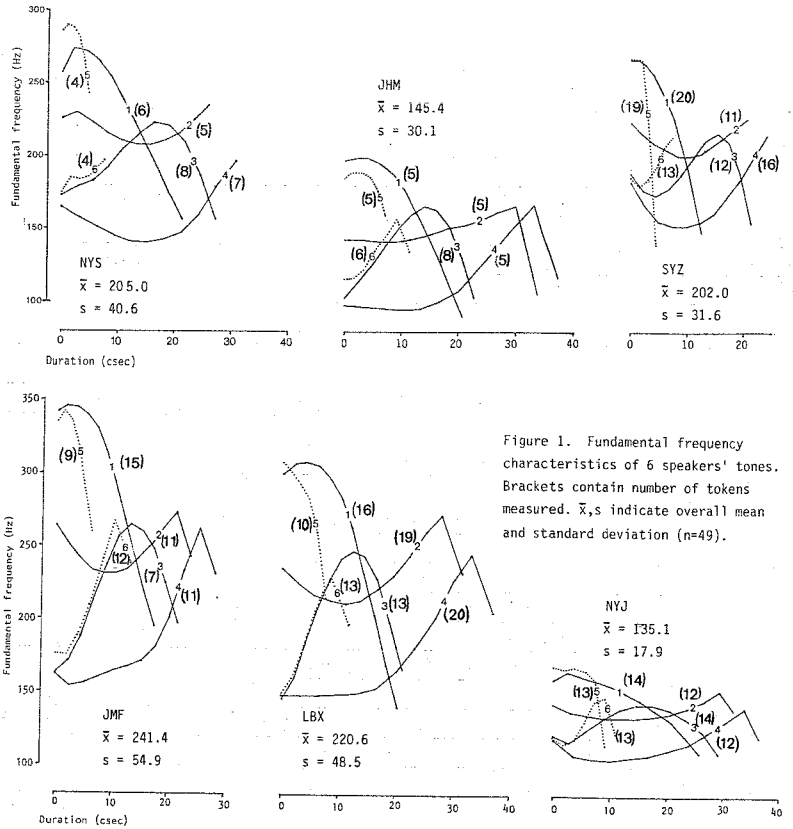


Figure 1. Fundamental frequency characteristics of 6 speakers' tones. Brackets contain number of tokens measured. $\bar{x}$,s indicate overall mean and standard deviation (n=49).

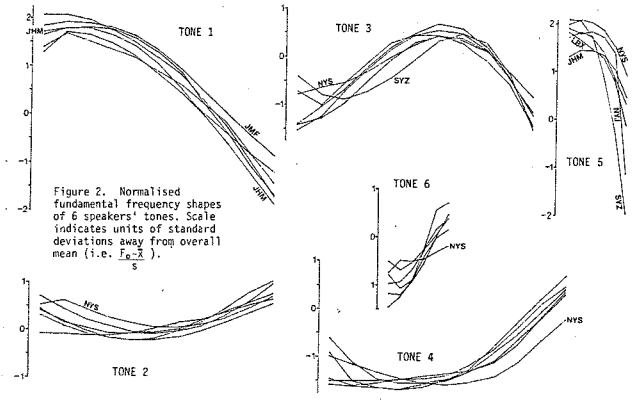


Figure 2. Normalised fundamental frequency shapes of 6 speakers' tones. Scale indicates units of standard deviations away from overall mean (i.e. $\frac{F_0-\bar{x}}{s}$).

CV(ʔ) syllables, where C = voiceless unaspirated obstruent, and V = monophthongal vowel.

Besides pitch, the six phonemic tones are characterised by a variety of co-occurring auditory features including voice quality (i.e. phonation type), voicing onset and offset, length, loudness, vowel quality and manner of syllable-initial consonant. Their pitch characteristics are as follows:
/Tone 1/ - high falling, with short initial level component - [tɑ] : knife
/Tone 2/ - level then rising in mid pitch range - [tɑʔ] : island
/Tone 3/ - convex in low half of pitch range - [d̥ɑ] : to flee
/Tone 4/ - low rising with either level or falling initial component - [d̥ɑʔ] : way
/Tone 5/ - very short high level or high falling (sometimes not possible to say which, possibly because of its extremely short length) - [tɕiʔ] : knot
/Tone 6/ - short low rising - [dz̥ɿʔ] : straight.

The six speakers differ with respect to age, sex, and socio-economic background. LBX is a 62 y.o. businessman; JMF a 30 y.o. male waiter; NYJ a 25 y.o. male student; NYS is NYJ's 30 y.o. student sister; SYZ is a 30 y.o. female labourer. All these informants speak forms of Zhenhai 鎮海 dialect (Zhenhai is a rural county in N.E. Zhejiang Province), and, with the exception of tones 5 and 6, their tones are very similar in pitch: LBX has a falling tone 5, the pitch of SYZ's tone 5 is indeterminate with respect to ± level, and all the others have a level pitch; NYS's tone 6 does not rise as much as the others'. The sixth speaker, JHM, is a 60 y.o. businessman from Cixi 慈溪 county, about 16 miles to the west of Zhenhai. He speaks a variety with slightly different segmental structure, but the same pitch values as the others.

Fig. 1 shows that all 6 speakers share a remarkably similar F0 configuration. This is despite large, statistically significant differences in central tendancy and dispersion, which parameters appear also to be linearly related. (Only the two females do not differ in mean F0; note also that 4 speakers have roughly the same lower limit to their range, and that the highest values are, unexpectedly, not shown by females.) Between-speaker differences in consonantally induced perturbation in the first few centiseconds of the F0 time course can also be seen.

All speakers have three distinct, evenly distributed onset points: tones 3, 4 and 6 have statistically the same low onset; tone 2 onsets in mid range, and tone 5 has a high onset. NYS's tone 1 appears to lie lower, and JHM's higher than their tone 5. For the other speakers, tones 5 and 1 have the same onset. The rapid drop in F0 in the few centiseconds after peak in tones 2, 4, 6 and 5 (which has not been shown for NYS and SYZ's tones 2, 4, and 6) is not audible as a fall in pitch and is presumably one acoustical correlate of the syllable-final [ʔ] which characterises these tones.

## CONSIDERATIONS & PROBLEMS

As with vowels, tonal normalisation should ideally satisfy both numerical and linguistic criteria, with the latter taking precedence: it should achieve a maximal reduction in between-speaker variance without sacrificing the desideratum of making phonetic sense. The degree of reduction can easily be quantified by the ratio of the dispersion coefficients[3] of the normalised and unnormalised data - the Normalisation Index (Rose 1982:145). One prerequisite for the calculation of this statistic, however, is the

availability of a large number of sampled data points. Ideally, the F0 shape of each tone should be specified by at least 5 sampled values, although this can clearly be unrealistic when the duration of the tone is as short as Zhenhai tone 5. (The F0 data in fig. 1 were sampled at 10% points of duration for the 'long' tones 1-4, and at 20% points of duration for the 'short' tones 5 and 6 - a mean sampling rate of about 40 Hz.) A relatively detailed F0 curve has the additional merit of resolving any between-speaker differences in contour which would not emerge at a lower sampling rate, (as for example in Earle (1975) or Dreher & Lee (1966) who sampled only at F0 onset, offset and mid/inflection point).

There are two ways in which normalisation can make phonetic sense. Firstly, there is the requirement that normalised values should correctly reflect the auditory percept (Disner 1980:256). In other words, it shouldn't make what is auditorily different appear the same, and vice versa. In vowel normalisation, this requirement is at least partially guaranteed by the use of perceptually relevant transforms such as the mel scale, effective F2' etc. However, the application to tone is more problematic: tonal normalisations are performed exclusively on F0 data, because F0 is the main acoustic cue to pitch. But the pitch of speech is also mediated by acoustical cues other than F0: amplitude (Rossi 1977); spectral properties (Hombert 1978), and possibly duration (Lehiste 1970) can influence the way F0 is perceived as pitch. In Zhenhai dialect, for example, there is a tonal contrast between falling and rising-falling pitch after a low level tone, which appears to be cued by differential amplitude distribution on essentially the same F0 shape (Rose 1984a), and the amplitude contour of tone 1 which has a fairly prominent shoulder, may be partly responsible for its initial level pitch percept (Rose 1982:158). Finally it can be noted that the pitch of tone 5 is higher than the onset pitch of tone 1, although they both have very similar F0 values for most speakers. (This apparent discrepancy between F0 and pitch could also be the result of some masking effect, however (Hombert 1978)).

In order to ensure, therefore, that tonal normalisations make perceptual sense, it would be necessary to find a way of incorporating amplitude, and possibly also spectral and duration data as well as masking effects. As yet this goal seems distant: amplitude and duration are still generally neglected in acoustic studies of tone, even though they can also provide valuable evidence for productional inferences (Rose 1984b). Moreover, the reliablity of pitch transcriptions has still to be assessed - perhaps along the same lines as the evaluation of vowel quality transcription in Laver (1965). For the present, then, we have to be clear that a set of normalised F0 shapes is still an acoustic representation, and cannot easily be evaluated in auditory terms.

The second way in which a normalisation strategy can make phonetic sense is in the degree to which it models the actual process of perceptual normalisation. The notion of 'range' plays an important part in most tone and intonation normalisations, and there is evidence that listeners' perceptual judgements of tone are in fact made with reference to an individual speaker's F0 range (Leather 1983). (In this sense, then, normalisations not making use of a range (Phuong 1981; Dreher & Lee 1966) are not as good.)

Normalisation range has been defined in different ways: Takefuta (1975) uses a range determined by a speaker's absolute highest and lowest F0 values - a method rejected by Earle (1975) in favour of a range defined by

mean maxima and minima. Both these approaches are criticised by Rose (1982:138,139) because maxima and minima are often on parts of an F0 contour which are most likely to reflect individual idiosyncracies in consonantal perturbatory effects - witness the differences at F0 onset in fig. 1 - and are therefore inappropriate points upon which to base a normalisation. Jassem (1975) defines range as one standard deviation about the arithmetical mean F0 of a speaker measured over some 60 seconds of running speech.

From the point of view of perceptual reality, there is very little to choose between these methods since it may be the case that listeners differ in the way they compute a speaker's range (Leather 1983). It might be possible for example to derive a range either from direct computation of maxima and minima, or from a speaker's mean F0 value (since there is a clear relationship between the two (Earle 1975:107, and above)). Jassem's approach, although computationally more complex, does have the attraction of avoiding the circularity of forcing congruence (i.e. deciding beforehand which two range-defining points are 'the same' between speakers, in order then to assess the degree of sameness between them). With Jassem's model, it is the distribution of all the instantaneous F0 points that determines the value of the normalisation parameters.

However, in cases where there is some indication that a range is best defined by two points - such an indication may be the equidistance of the low, mid and high onset points in the data above - the Jassem approach can in fact introduce undesired artefacts, as is demonstrated below.

Fig. 2 shows the F0 curves of fig. 1 normalised with a variant of the Jassem approach. The normalisation parameters of mean F0 and standard deviation (which are given in fig. 1) are calculated from the sampled F0 values of the tones themselves. A small additional reduction in between-speaker variance has been achieved by excluding all onset F0 values, and offset F0 values in tones 2, 4, 5 and 6 with final [2]. (This is justified on the grounds that these values reflect between-speaker differences associated with syllable-initial and -final consonants rather than tones; they have also been ignored in calculating the Normalisation Index.)

As can be seen, the normalisation is rather effective: quantitatively, it has reduced the amount of variance due to between-speaker differences by a factor of 13.5 - from 68.7% in the unnormalised data to only 5.1% in the normalised data. It also shows a possible correlation between sex and contour in tone 3, and that the low offset to NYS's tone 6 is not anomalous, since her tone 4 has it as well. However, it has obscured the relative position of tone 1 to tone 5 in NYS and JHM: if it is the case that tone 5 does define the range maximum, we should want the normalised 5 tones to cluster more tightly at the expense of tone 1. It is also clear that a normalisation based on such a range would shift NYS's tone 2 down, and JHM's tone 1 up - and this would contribute to an additional drop in between-speaker variance.

As far as the method and evaluation of tonal normalisation is concerned, then, perceptual reality indicates that range-based normalisations are preferable, and their statistical evaluation is easy, given enough data points. However, the difficulty of auditory evaluation often makes the choice of a particular strategy problematic.

NOTES

(1) For a discussion of the types of information present in the speech wave - Accentual vs. Linguistic vs. Personal - see Ladefoged (1967:104). For a criticism of the distinction between Accentual and Personal, see Nolan (1983:68,69).

(2) The data for JMF, LBX, and NYJ are from Rose (1982); other speakers were specially measured for this paper. Thanks to Cathy Wildermuth for measuring NYS, and to Xu Weiyuan for arranging the recording of SYZ.

(3) The dispersion coefficient is the ratio of mean between-speaker variance to overall sample variance, and is a measure of the degree to which speakers' values cluster (Earle 1975).

REFERENCES


DISNER, S. (1980) "Evaluation of Vowel Normalisation Procedures" JASA 67, 1, 253-261.

DREHER, J.J. & LEE. P.C. (1966) "Instrumental Investigations of Single and Paired Mandarin Tonemes", Douglas Advanced Research Laboratory Paper 4156, California.

EARLE, M.A. (1975) "An Acoustic Phonetic Study of Northern Vietnamese Tones", Speech Communication Research Laboratories Inc. Monograph 11, Santa Barbara.

JASSEM, W. (1975) "Normalization of FO Curves" in Fant & Tatham (eds.) "Auditory Analysis and Perception of Speech", Academic Press, 523-530.

HOMBERT, J-M. (1978) "Consonant Types, Vowel Quality and Tone" in Fromkin (ed.) "Tone A Linguistic Survey" Academic Press, 77-111.

LADEFOGED, P. (1967) "Three Areas of Experimental Phonetics" OUP.

LAVER. J.D.M.H. (1965) "Variability in Vowel Perception", Language & Speech 8, 95-121.

LEATHER, J. (1983) "Speaker Normalization in Perception of Lexical Tone" Journal of Phonetics 11, 373-382.

LEHISTE, I. (1970) "Suprasegmentals", MIT Press.

NOLAN, F. (1983) "The Phonetic Bases of Speaker Recognition", CUP.

PHUONG, V.T. (1981) "The Acoustic and Perceptual Nature of Tone in Vietnamese", Ph.D. Thesis, Australian National University.

ROSE, P.J. (1982) "An Acoustically Based Phonetic Description of the Syllable in the Zhenhai Dialect', Ph.D. Thesis, Cambridge University.

--------- (1984a) "A Re-evaluation of the Acoustical Correlates of Pitch in Tonal Contrasts" Australian Linguistic Society Conference Paper.

--------- (1984b) "The Role of Subglottal Pressure and Vocal Cord Tension in the Production of Tones in a Chinese Dialect", in Hong (ed.) "New Papers on Chinese Language Use", Canberra, 133-168.

ROSSI, M. (1977) "Les configurations et l'interaction des Pentes de FO et de I", Proc. 9th Int. Cong. Phonetic Sciences, vol. 1, Copenhagen 246.

TAKEFUTA, Y. (1975) "Method of Acoustic Analysis of Intonation", in Singh (ed.) "Measurement Procedures in Speech Hearing and Language" University Park Press, Baltimore, 363-378.