Speech Communication 6 (1987) 343–351
North-Holland

CONSIDERATIONS IN THE NORMALISATION OF THE FUNDAMENTAL FREQUENCY OF LINGUISTIC TONE

Phil ROSE

Department of Linguistics, The Faculties, Australian National University, Canberra, ACT, Australia.

Abstract. Some considerations in the normalisation of tone are discussed, and their application demonstrated on the fundamental frequency data of seven speakers of a variety of Wu Chinese. It is argued that, although a considerable reduction in between-speaker variance can be achieved by either a *Z*-Score or Fraction of Range normalisation, the former strategy is preferable because of its slightly superior numerical performance, and relative lack of methodological problems. The derivation of a possible Linguistic-Phonetic representation for comparison with other varieties is also illustrated.

Zusammenfassung. Es werden einige Betrachtungen zur Tonnormalisierung diskutiert und ihre Anwendungen auf die Grundtondaten von 7 Sprechern des Wu-Chinesischen demonstriert. Die Ergebnisse lassen den Schluss zu, dass, obwohl eine beträchtliche Reduzierung der Inter-Sprechervarianz durch entweder den *Z*-Wert oder durch eine "Fraction of Range"-Normalisierung erreicht werden kann, der *Z*-Wert zu bevorzugen ist, weil diese Methode eine etwas höhere numerische Leistung erbringt und, relativ gesehen, keine methodologischen Probleme aufwirft. Darüber hinaus wird die Ableitung einer möglichen linguistisch-phonetischen Repräsentation zu Vergleichszwecken mit anderen Varianten des Chinesischen illustriert.

Résumé. Nous discutons de la normalisation du ton et de son application à des données sur la fréquence fondamentale de sept locuteurs parlant une variété du chinois Wu. Une réduction importante de la variabilité inter-locuteur peut être réalisée, soit par une normalisation "*Z*-score", soit par une normalisation par "fraction of range". Nous préférons la première approche à cause de ses performances légèrement supérieures et à cause de l'absence de problèmes méthodologiques majeurs. Nous dérivons une représentation phonétique en vue d'une comparaison avec d'autres variétés du chinois.

Keywords. Chinese, tone, normalisation, fundamental frequency, between-speaker differences.

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 phonetically the same sound. The magnitude of between-speaker (B-S) 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 analogue of this perceptual process, which aims to extract and specify the invariant acoustic correlates of the Accentual and Linguistic features of a particular variety, and then to compare varieties for typological and universal purposes (Disner, 1980, p. 253).

One major physiological source of B-S differences in acoustical output is the difference in size (i.e., length and mass) of the vocal cords. Such differences result in different preferred, or default values and ranges of the fundamental frequency (F_0) of the radiated wave (Nolan, 1983, pp. 51, 59). F_0 is the basic acoustic correlate of perceived pitch, which functions as a dimension for supra-segmental linguistic systems of intonation, stress, and tone (Lehiste, 1970). Thus male speakers,

[1] For a discussion of the types of information present in the speech wave—Accentual vs. Linguistic vs. Personal—see Ladefoged (1967, p. 104). Nolan (1983, pp. 68, 69) criticises the Accentual/Personal distinction.

who tend to have longer, more massive cords than females, tend to have correspondingly lower F_0 values, and it is possible for a female's phonologically low tone to have a higher F_0 than a male's phonologically high tone.

In contrast to the large amount of theoretical and empirical work done on vowel normalisation, the normalisation of the acoustical correlates of intonation has received very little attention, and the normalisation of tone even less. This is a pity, because without normalisation it is not possible to take the first step in defining the phonetic and phonological features of a particular variety, namely, separating those features of the acoustic signal which are characteristic of the individual from those which characterise the particular variety. Moreover, normalisation offers the only way, other than (the unlikely) recourse to bilingual speakers, of testing transcriptionally based hypotheses on the nature of linguistic-phonetic tonal variation. Do, for example, the high level tones of Mandarin (transcribed as [55]) and Suzhou ([44]) really differ in relative height (Yuan et al., 1980, p. X), and is there continuous or discrete cross-linguistic variation in the set of phonetic tones?

The aim of this paper is to look briefly at some important considerations in the normalisation of tone, and demonstrate their application to F_0 data from a variety of Chinese in order to obtain a partial specification of the acoustic correlates of the tones of that variety. To give an idea of the nature of the problem, the F_0 data will be presented first, together with a description of the pitch of the tones.

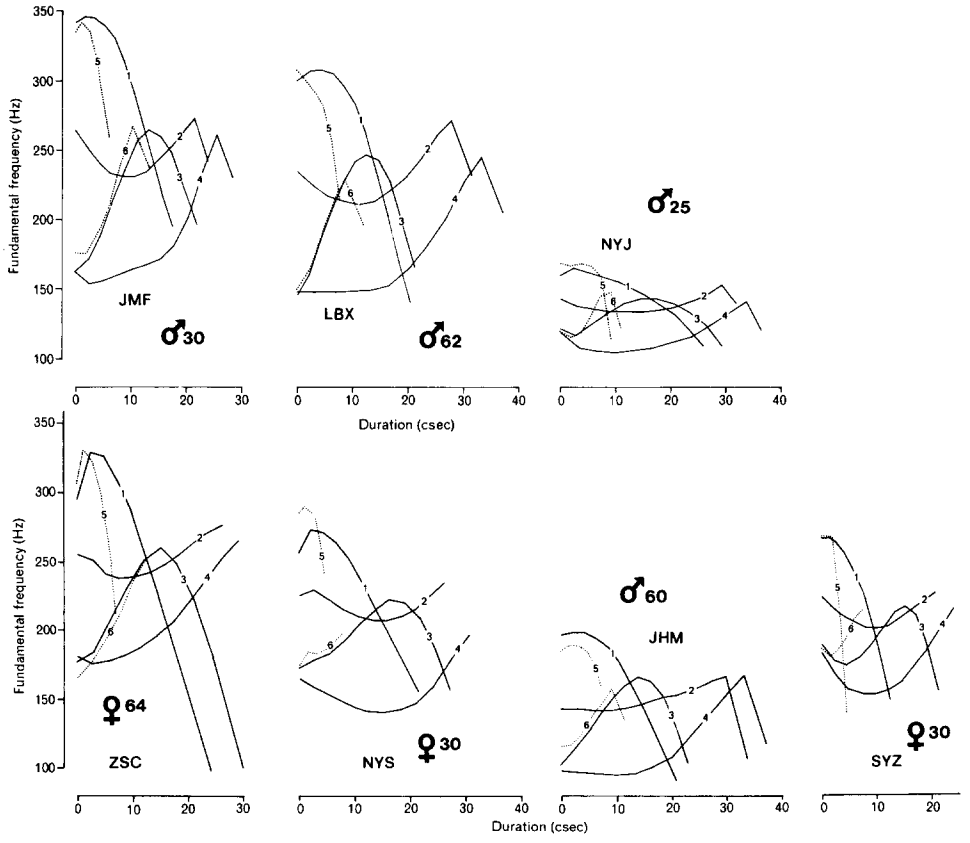

Fig. 1. Fundamental frequency characteristics of seven speakers' tones.

Data

Figure 1 shows raw F_0 shapes of the same 6 phonemic tones of a variety of Chinese[2] as spoken under similar circumstances by 4 male and 3 female native speakers of differing ages. The F_0 values represent arithmetic means of several tokens, and were sampled at percentage points of the duration of the voiced part of the syllable—a mean rate of about 40 samples/sec. (see Table 1). This sampling rate is required to resolve any B-S differences in F_0 contour which would not otherwise emerge (cf. Earle (1975) or Dreher and Lee (1966), who sampled only at onset, offset and mid/inflection point). Except for one speaker (JHM), the corpus was controlled for the well known intrinsic effects of segmental structure on F_0 (Lehiste, 1970, p. 68ff.). The effect of vowel height on F_0 was controlled within a given speaker, and the perturbatory effect of the syllable-initial consonant was controlled both within and between speakers by analysing only syllables with initial voiceless unaspirated obstruent, the vast majority of which were stops.

The tones have the following pitch characteristics:

/T1/ high falling, with initial level component: [tɕi] "to fill";
/T2/ level then rising in mid pitch range: [tɕiʔ] "chicken";
/T3/ convex in low half of pitch range: [d̥ʑi̤] "to ride";
/T4/ low rising into mid pitch range: [d̥ʑ i̤ʔ] "he, she, it";
/T5/ very high level or high falling: [tɕiʔ] "foot";
/T6/ short low rising into mid pitch range: [d̥ʑɛ̆ʔ] "straight".

There are no systematic B-S differences in the pitch of T1, T2, and T3, but some obtain in the pitch of T4, T5, and T6. LBX's T5 has a falling pitch; the pitch contour of SYZ's T5 is ambiguous: it is possible to hear most tokens as both level or falling; the other speakers have level T5. NYS's T4 and T6 do not rise as much as the others'. With the exception of ZSC, all speakers have either a level or falling initial component to their T4; most of ZSC's T4 tokens lack this initial component and simply rise.

Besides pitch, the 6 phonemic tones are characterised by a variety of co-occurring auditory features including voice quality (i.e., phonation type), voicing onset and offset, length, vowel quality, and manner of syllable-initial obstruent.

It is obvious from Fig. 1 that, despite the expected B-S differences in central tendency and dispersion, which are quite large, all seven speakers share a remarkably similar F_0 configuration. The fairly strong positive correlation observable between central tendency and dispersion has also been noted for other tone languages: Vietnamese (Earle, 1975, p. 107) and Mandarin (Chen, 1974, p. 168). Contrary to expectation, the highest values are not necessarily shown by females. The rapid F_0 drop in the few centiseconds after peak in T2, T4, T5 and T6 (which does occur but has not been shown in T2, T4 and T6 of NYS, SYZ and ZSC) is not audible as a fall in pitch and is one acoustical correlate of the syllable-final glottal stop which characterises these tones. (In T2, T4 and T6, the duration of the drop is excluded from the sampling base, which therefore extends from phonation onset to F_0 peak.)

Probably the most conspicuous B-S difference in F_0 concerns the offset values of T1 and T3, and in particular their relationship to the other tones. ZSC especially shows exceptionally low offset values, which also correlate with exceptionally long duration values.[3] The speakers also differ in the way their syllable-initial obstruents and syllable-final glottal stop perturb the F_0 over the first and last few centiseconds of its time course.

[2] All the informants except JHM speak varieties of Zhenhai dialect (Zhenhai is a rural county in N.E. Zhejiang province) JHM is from Cixi country, 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. Cixi and Zhenhai belong to the major Wu dialect group of Chinese.

[3] Differences in offset values between ZSC and the rest reflect differences in manner of phonation offset in T1 and T3. For all speakers except ZSC, the point of F_0 offset was marked by a clear discontinuity in increase of period of the pulse train. For SYZ, the phonation then became creaky; for the rest except ZSC it ceased almost immediately. For ZSC however there was no such discontinuity, and her F_0 period continued to increase gradually to a much lower offset value.

Table 1
Means and standard deviations ($\bar{x}$,s) for fundamental frequency and duration of seven speakers' tones. n indicates number of tokens per sample; percent values indicate percentage point of duration at which F_0 was sampled

	Fundamental frequency											Duration	n
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%		
Tone 1													
NYJ	160,10	165,6	162,7	159,7	156,8	151,7	146,7	139,7	132,7	121,8	110,9	25.7,2.8	14
JMF	342,11	346,12	345,14	340,18	331,22	315,23	293,23	268,21	243,21	219,14	195,12	17.5,1.8	15
LBX	300,21	307,19	308,21	306,23	297,21	283,21	263,19	232,18	203,17	169,17	140,11	20.4,2.0	16
SYZ	267,19	267,19	264,17	257,17	249,17	239,16	228,16	211,16	194,16	172,16	149,16	12.4,1.9	20
ZSC	295,39	328,33	326,33	308,31	288,30	258,28	227,27	194,22	161,17	129,14	98,14	24.1,3.5	25
NYS	257,18	274,11	272,9	265,11	254,13	239,15	223,16	206,14	189,11	173,8	157,6	21.3,1.5	6
JHM	197,17	199,12	199,11	195,8	188,5	177,6	161,8	145,11	128,10	109,9	90,6	20.8,2.6	5
Tone 2													
NYJ	143,9	138,6	136,6	134,6	134,6	134,6	135,6	137,7	141,8	146,8	153,9	29.3,2.8	12
JMF	264,13	252,14	241,14	233,14	231,16	231,17	234,18	242,19	251,20	262,22	273,24	21.4,1.7	11
LBX	235,23	225,21	217,19	213,18	211,18	213,19	221,22	230,24	244,24	260,24	272,23	27.8,2.9	19
SYZ	224,12	216,13	209,12	205,11	201,11	201,11	203,11	209,13	215,14	221,16	227,17	20.4,2.9	11
ZSC	255,26	251,28	240,28	237,28	238,29	241,30	247,32	255,34	264,35	271,35	276,34	26.2,5.7	19
NYS	226,29	230,26	223,24	216,25	211,20	208,20	208,20	211,20	216,22	225,22	235,20	26.0,2.8	5
JHM	143,10	143,8	142,8	142,8	144,10	147,9	151,9	153,10	157,12	162,13	166,11	29.8,4.2	5
Tone 3													
NYJ	122,8	117,7	125,8	133,7	140,6	143,6	143,6	140,6	134,6	125,8	110,9	29.6,3.6	14
JMF	163,12	171,7	188,8	214,11	237,17	257,16	265,16	260,15	248,16	225,14	197,10	21.7,1.8	7
LBX	146,13	160,15	185,13	208,14	228,16	242,18	247,21	244,20	228,20	199,19	166,12	21.2,2.6	13
SYZ	189,7	177,10	174,10	179,12	188,14	200,17	212,15	217,14	211,12	191,9	155,15	21.2,2.8	12
ZSC	177,14	184,14	209,18	231,21	250,23	260,25	249,25	223,27	187,25	143,21	100,20	30.0,2.7	21
NYS	173,9	179,14	183,16	193,17	205,17	215,15	223,14	221,15	210,16	187,17	157,17	27.2,3.3	8
JHM	103,9	114,11	125,12	139,15	151,16	160,17	166,20	163,18	152,17	132,16	103,8	22.9,1.4	8
Tone 4													
NYJ	119,6	108,6	106,7	105,7	107,8	108,8	112,8	116,8	124,6	132,6	141,8	33.8,3.9	12
JMF	162,18	154,15	156,14	160,14	164,16	167,14	171,12	181,13	201,14	233,18	262,24	25.5,3.8	11
LBX	148,13	148,10	148,9	149,11	150,12	153,13	165,17	181,19	202,24	227,29	245,29	33.3,4.2	20
SYZ	183,14	167,12	156,8	153,6	153,5	156,6	162,7	174,6	186,6	201,9	216,10	24.0,3.0	16
ZSC	181,29	175,25	177,23	182,24	188,24	197,25	206,27	221,30	237,33	252,35	265,37	29.0,3.4	28
NYS	165,17	158,10	152,8	146,10	142,11	141,15	143,18	148,18	160,20	179,22	197,18	30.3,2.1	7
JHM	98,5	97,4	96,5	95,4	96,5	101,5	108,10	123,16	138,16	153,18	167,17	33.1,2.6	5
Tone 5													
NYJ	169,12		167,10		168,11		166,9		160,8		115,14	9.1,2.4	13
JMF	335,17		342,16		336,9		319,16		289,18		260,15	6.1,1.2	9
LBX	308,11		301,9		293,9		283,11		260,14		215,14	7.6,1.1	10
SYZ	268,16		267,16		257,17		233,16		195,20		140,17	4.5,0.9	19
ZSC	306,42		330,43		321,36		298,31		262,30		212,31	6.9,2.4	18
NYS	286,19		290,16		288,12		281,9		267,11		243,19	4.5,0.6	4
JHM	185,7		189,11		189,10		186,12		177,7		160,14	7.0,1.2	5
Tone 6													
NYJ	121,9		116,7		119,8		131,10		145,9		148,7	9.2,1.4	13
JMF	176,10		175,11		190,8		211,11		241,14		267,20	10.3,1.4	12
LBX	150,11		160,14		177,13		200,12		219,15		228,14	8.7,1.1	13
SYZ	186,10		180,5		185,6		194,7		207,9		214,10	7.3,0.9	13
ZSC	166,31		177,26		193,21		211,21		236,22		253,27	12.5,2.0	11
NYS	175,10		185,23		184,27		186,28		192,27		198,24	7.3,1.8	4
JHM	116,5		117,8		125,9		137,9		147,8		157,8	9.1,1.0	6

Considerations in tone normalisation

Vowel normalisation aims for a maximum reduction in B-S variance without sacrificing the desideratum of making perceptual sense. The notion of perceptual sense, which thus serves to evaluate the numerical strategy, can be understood in two ways: normalised values should correctly reflect the transcriber's auditory impression (Disner, 1980, p. 256), and normalisation should ideally model the actual process of the listener's perceptual normalisation. Use of perceptually relevant transforms of the acoustic data such as the mel scale and effective F_2 (Fant, 1973, p. 46ff.) or sones/bark (Bladon and Lindblom, 1981) help meet these requirements. Productionally relevant transforms have also been proposed (Fant, 1973, p. 84).

In the normalisation of tone, strategies for achieving reduction in B-S variance and for quantifying this reduction are relatively straightforward, but the perceptual criterion is difficult to apply. Although F_0 is clearly the basic term in the function relating acoustics to linguistic pitch, the pitch of speech is nevertheless mediated by the interaction of F_0 with the other major acoustic parameters of radiated amplitude, duration, spectrum, and correlates of segmental structure (Rose, in press).

Examples of this complex relationship between F_0 and pitch occur in the tonal data. On the one hand there are clear correlations between pitch and F_0 (in the B-S differences in the pitch of T4 for instance); on the other there are clear (usually within-speaker) discrepancies. The pitch of T5, for example, is higher than the onset pitch of T1, although they both have statistically the same initial F_0 values, and the initial level pitch component in T1 may be due more to a prominent amplitude shoulder than any F_0 feature (Rose, 1982, p. 158).

In order to ensure that tonal normalisations made perceptual sense, it would be necessary to integrate the other perceptually relevant acoustic parameters. This is not yet possible, given the state of our knowledge of the relationship between acoustics and linguistic pitch, and the fact that all acoustic parameters other than F_0 are usually neglected in tonal studies. Moreover, the reliability and consistency of linguists' pitch transcriptions have still to be assessed. The efficacy of a tone normalisation strategy cannot therefore be evaluated primarily in perceptual terms, although a possible heuristic, given that F_0 is the basic determinant of pitch, would be to prefer those normalisation strategies that reflect B-S pitch differences as differences in normalised F_0.

The complex relationship between F_0 and linguistic pitch would also vitiate perceptually motivated transforms of the F_0 alone (such as the semitone scale advocated by Chiang (1967, p. 108)), even if it were clear which transform best approximated linguistic pitch.[4] It seems, in sum, that a set of normalised F_0 shapes is still best considered as an acoustic representation, and the evaluation of an F_0 normalisation should be based on a principled choice between numerical strategies.

The two most common strategies in the normalisation and scaling of F_0 are of the general linear form:

$$F_{0_{norm}} = (F_{0_i} - F_{0_{ref}})/F_{0_{range}}.$$

Fraction of range (FOR) transforms (eg. Earle (1975), Takefuta (1975), Rose (1982), Ladd et al. (1985)) express an observed F_0 value as a fraction of the difference between two range-defining F_0 values, eg.

$$F_{0_{norm}} = (F_{0_i} - F_{0_{min}})/(F_{0_{max}} - F_{0_{min}}).$$

Z-score transforms (eg. Jassem (1971), Menn and Boyce (1982)), express an observed F_0 value as a multiple of a measure of dispersion away from a mean F_0 value, eg.

$$F_{0_{norm}} = (F_{0_i} - \bar{F}_0)/s,$$

where s is one standard deviation about the mean F_0 ($\bar{F}_0$).

[4] As Menn and Boyce (1982, p. 379) point out, neither musical scale (log to base 2) nor mel scale (effectively linear below 1 kHz), are neccessarily appropriate for the F_0 of the spoken voice. More appropriate reasons for log transforms are statistical (Menn and Boyce, 1982, p. 379) and productional (Fujisaki, 1983, p. 50ff.). Note that although the tonal data appear to display the assumed logarithmicality of the F_0 scale (Ladd et al., 1985, p. 443) a prior log transform will be ineffective because linear normalisation strategies transform log or linear data equally well.

Leather (1983) has demonstrated that an individual speaker's inferred F_0 range plays an important part in the perception of tone, and since both approaches incorporate this notion ($F_{0_{max}} - F_{0_{min}}$,$s$), there is little to choose between them from the point of view of perceptual reality,[5] especially since listeners may differ in the way they compute a speaker's F_0 range (Leather, 1983, p. 379). Also, if the normalisation parameters (NPs) of standard deviation, mean F_0, and ($F_{0_{max}} - F_{0_{min}}$) are closely correlated—as for example in North Vietnamese (Earle, 1975, pp. 104–109)—both strategies might be expected to reduce B-S variance by approximately the same degree, and neither will be numerically superior.

The FOR strategy, although conceptually and computationally simple, is open to several methodological objections. Range-defining (R-D) points are values that are assumed to be equivalent between speakers, but unless they can be identified by external criteria, the normalisation becomes circular. (An FOR normalisation requires R-D points, but which points are equivalent between speakers only becomes clear after normalisation.) One possible external criterion is the property of within-speaker invariance: Ladd et al. (1975, p. 443) propose the F_0 value at the end of an utterance-final fall as a perceptually real R-D point because it is clearly a speaker-constant in several languages. However, the tonal data show that speaker-constant F_0 values (and even values that are otherwise clearly equivalent between speakers, like the low F_0 onset locus) are not necessarily appropriate R-D points. For example, except for NYJ, the offset value of T1 (the end of an utterance-final fall) qualifies as in other languages as a speaker-constant by virtue of its low within-speaker variance (see Table 1). But taking T1 offset as the lower R-D point will normalise ZSC's tones about 25% higher relative to the other speakers—a clearly undesirable result. Note also that potential R-D F_0 maxima and minima often occur on parts of an F_0 contour most likely to reflect individual idiosyncrasies in consonantal perturbatory effect, or phonation offset, and are therefore inappropriate points upon which to base a tonal normalisation. Finally, the common occurrence of R-D values at the same B-S sampling point fixes the normalised B-S variance at that point at zero, thus complicating statistical evaluation.

The Z-score transform avoids the circularity of forcing congruence, since it is the distribution of many F_0 values, not just two, which determines the NPs. Also, the root-mean-square basis of the function will ensure a globally distributed reduction in B-S variance (compared to the local fluctuations in B-S variance that arise artefactually in an FOR normalisation if R-D values occur at the same B-S sampling point). However, as Disner (1980, p. 257) has pointed out for normalisations of different vowel systems, Z-score NPs should only be calculated from samples that are comparable (i.e., transcriptionally equivalent) to avoid biasing. The present data suggest that this might be a difficult requirement to satisfy for tones, even for normalisation within the same variety. Examination of even a few speakers has shown auditory and acoustic differences between some speakers *for most* (5 out of 6) *tones*, thus preventing a simple exclusion of all non-comparable tones from the calculation of the NPs. More importantly, although the Z-score transform is sufficiently robust to neutralise the effect of small B-S differences, some acoustic differences are clearly of sufficent magnitude to significantly bias NPs (ZSC's T1 and T3, for example, will adversely lower her overall mean). If B-S differences are considered too big, an FOR normalisation, which requires only two comparable points, can always be used. Nevertheless, it does seems preferable, both for Z-score and FOR transforms, to be able to use NPs derived from a set of observations other than those that are to be normalised, eg. the speaker's long-term F_0 distribution (Jassem, 1975). This as yet apparently little tried approach looks particularly worthwhile in the case of tone, because if might reveal solutions to the vexed question of how isolation tone values relate to those of normal speech. It also seems indispensible for comparing tones across varieties with different systems, especially since varieties can differ significantly in range (Phuong, 1981).

[5] According to this criterion, normalisations not making use of a range (eg. Phuong, 1981; Dreher and Lee, 1966) are not as good.

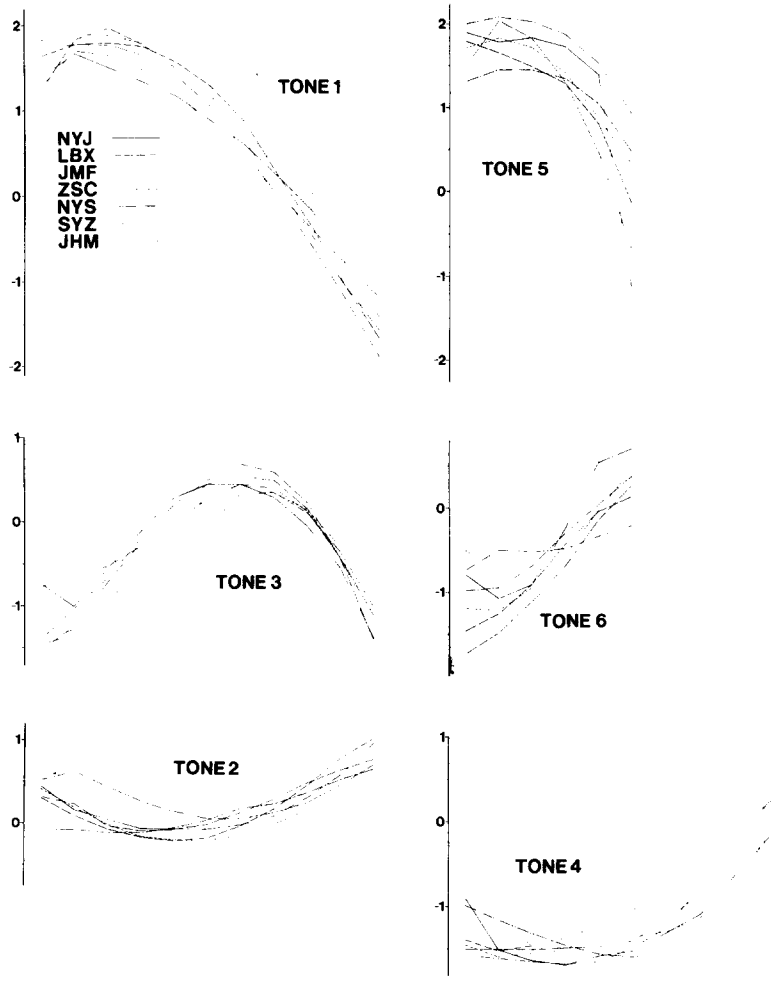

Fig. 2. Z-score normalised fundamental frequency shapes for seven speakers' tones plotted against equalised duration. Scale indicates units of standard deviation away from overall mean ($F_{0_i} - \bar{F}_0 / s$).

Application

Several different normalisations were performed on the data to ascertain which most effectively reduced the B-S variance. Figure 2 shows the results of the most successful: a Z-score normalisation using NPs of arithmetic mean and one standard deviation. The NPs were calculated for each speaker from the set of his sampled mean F_0 values in Table 1, excluding values at onset (0%), and offset (100%) in T5 (i.e., 49 data points for each speaker).[6] Exclusion of these values is theoretically advisable, because they are the ones most likely to reflect B-S differences associated with the effect of syllable-initial and -final consonants, rather than tones. The problem posed by ZSC's T1 and T3 was solved by ignoring the last part of the F_0 time course in these tones, resampling F_0 as a function of the new, shorter, duration bases, and including the resampled F_0 values in the calculating of her NPs.[7] This implies, plausi-

[6] Thus the normalised value at 0% in T1 for NYJ ($\bar{F}_0 = 135.1$, $s = 17.9$) is $(160 - 135.1)/17.9 = 1.39$.

[7] The new duration base for the resampling was calculated by reducing the duration of her T1 and T3 to 64% and 79% respectively of the duration of her T4. These figures represent the mean ratios for the other 6 speakers, and yield new duration values for her T1 and T3 of 18.6 and 20.3 csec.

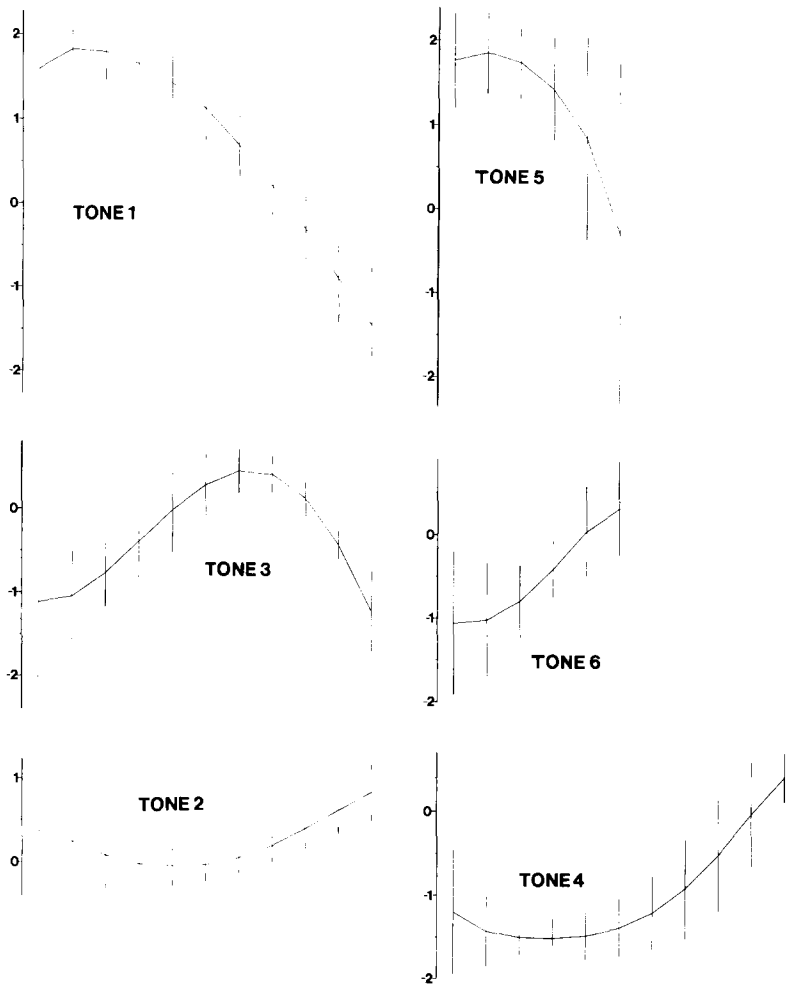

Fig. 3. Linguistic-phonetic representation of the fundamental frequency characteristics of Zhenhai/Cixi dialect isolation tones. Solid line shows mean normalised F_0; vertical lines indicate 4 standard deviations around mean. Scale as for Fig. 2.

bly, that ZSC has the same basic F_0 contours for her T1 and T3 as the other speakers, only differing from them in mode of phonation offset. The apparent success of this approach indicates that the duration of the voiced part of the syllable does not necessarily constitute an appropriate sampling base, since it can yield sampled F_0 values which are not comparable between speakers. Future normalisation strategies must therefore address the problem of finding an appropriate sampling base from raw data.

The effect of the normalisation can best be appreciated by comparing Figs. 1 and 2. Most of the transformed F_0 shapes cluster tightly, and at least some of the tones with different B-S pitches (eg. ZSC's T4 with its immediately rising pitch, and NYS's T4 and T6 with their lower offset) have been kept separate in a satisfying way. Quantitatively, the normalisation has reduced the amount of variance due to B-S differences by a factor of 12.9—from 65.8% in the unnormalised data to 5.1% in the normalised data.[8] The same normalisation on log-transformed F_0 values reduced the

[8] These are the values for the Dispersion Coefficient (DC) of the raw and normalised data. The DC is the ratio of mean B-S variance to overall sample variance, and is a measure of the degree to which speakers' values cluster. An almost identical DC (65%) was found for the unnormalised F_0 of the 6 tones of 11 (4 female and 7 male) N. Vietnamese speakers (Earle, 1975, p. 133).

B-S variance by the slightly smaller amount of 11.4, which indirectly reflects the inherent logarithmicality of the raw data. A version of an FOR transform (on the 49 data points, with R-D values of the speaker's highest mean value and the lowest mean value in T4—the latter chosen because it is also a near speaker-constant) fared a little worse than the best *Z*-score, with a reduction factor of 10.8. A *Z*-score normalisation on Earle's (1975) N.Vietnamese data was also found to perform better than his original FOR results by a small difference of 1.2. These results suggest that *Z*-score normalisations are preferable, both on the balance of theoretical considerations and, marginally, on numerical performance. The particular F_{min} value chosen also seems to be an effective R-D speaker-constant.

The *Z*-score normalised data can now be used to specify part of a possible linguistic-phonetic acoustical representation of the Zhenhai/Cixi isolation tones. Such a representation is shown in Fig. 3, which plots the mean of the normalised tonal values, together with two standard deviations above and below the mean.[9] Since a range of four standard deviations around the mean will include about 95% of all normally distributed observations, this indicates the magnitude of expected variation in the normalised F_0 values of Zhenhai/Cixi tones. Interestingly, this amount is almost identical to that obtained for Earle's (1975) N.Vietnamese data (for a *Z*-score normalisation based on all 56 data points, the mean $4 \times s$ is 28% of the total range, compared to 27% for the Vietnamese data). This is perhaps suggestive of a limit on the amount of reduction in B-S variance that can be achieved by normalisation.

[9] Note that this is still strictly speaking incomplete as an acoustic representation of the Zhenhai/Cixi tones, since the initial part of the mean F_0 contour still reflects the effect of only one type of syllable-initial consonant. Also, in order to provide a detailed basis of comparison with other varietes, the normalised F_0 parameters should be expressed as functions of normalised duration parameters (rather than equalised duration, as here). Normalised amplitude should also be specified to facilitate inferences on B-S similarities in production.

Summary

This paper has discussed and illustrated aspects of the normalisation of tone, including problems of evaluation, relative merits and implications of two common strategies, and linguistic-phonetic application. It has argued that, at the present state of our knowledge, *Z*-score normalisations are preferable, provided that between-speaker comparability of tones is maximised by ignoring intrinsic consonantal effects and identifiable speaker-dependent features in the calculation of the normalisation parameters. Five directions for future research are indicated: the relationship between acoustics and linguistic pitch (both transcribed and perceived); the normalisation of other tonally relevant acoustic parameters like duration and amplitude; the relationship between isolation tone normalisation parameters and statistical parameters from long-term data; the problem of appropriate sampling bases; and comparison of linguistic-phonetic specifications to reveal the nature of cross-linguistic dimensions of variation in phonetic tone.

References

R.A.W. Bladon and B. Lindblom (1981), "Modelling the judgement of vowel quality differences", *J. Acoust. Soc. Am.*, Vol. 69, pp. 1414–1422.

Gwang-tsai Chen (1974), "The pitch range of English and Chinese speakers", *J. Chinese Linguistics*, Vol. 2 (2), pp. 159–171.

H.T. Chiang (1967), "Amoy-Chinese tones", *Phonetica*, Vol. 17, pp. 100–115.

S. Disner (1980), "Evaluation of vowel normalisation procedures", *J. Acoust. Soc. Am.*, Vol. 67, pp. 253–261.

J.J. Dreher and P.C. Lee (1966), Instrumental investigations of single and paired Mandarin tonemes, Paper 4156, Douglas Advanced Research Laboratory, California.

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

G. Fant (1973), *Speech Sounds and Features*, (MIT Press, Cambridge, MA).

H. Fujisaki (1983), "Dynamic characteristics of voice fundamental frequency in speech and singing", in: *The Production of speech*, ed. by P. MacNeilage (Springer, Berlin) pp. 39–55.

W. Jassem (1975), "Normalisation of F_0 curves", in: *Auditory Analysis and Perception of Speech*, ed. by G. Fant and M. Tatham (Academic Press, London) pp. 523–530.

P. Ladefoged (1967), *Three Areas of Experimental Phonetics* (Oxford University Press, London).

D.R. Ladd, K. Silverman, F. Tolkmitt, G. Bergmann and K. Scherer (1985), "Evidence for the independent function of intonation contour type, voice quality, and F_0 range in signalling speaker affect", *J. Acoust. Soc. Am.* Vol. 78, pp. 435–444.

J. Leather (1983), "Speaker normalisation in perception of lexical tone", *J. Phonetics*, Vol. 11, pp. 373–382.

I. Lehiste (1970), *Suprasegmentals*, (MIT Press, Cambridge, MA).

L. Menn and S. Boyce (1982), "Fundamental frequency and discourse structure", *Language and Speech*, Vol. 25, pp. 341–383.

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

V.T. Phuong (1981), "The acoustic and perceptual nature of tone in Vietnamese", Ph.D. Thesis, Australian National University.

P. Rose (1982), "An acoustically based phonetic description of the syllable in the Zhenhai dialect", Ph.D. Thesis, Cambridge University.

P. Rose (in press), "On the non-equivalence of fundamental frequency and pitch in tonal description" *Pacific Linguistics.*

Y. Takefuta (1975), "Method of acoustic analysis of intonation", in: *Measurement Procedures in Speech Hearing and Language*, ed. by S. Singh (University Park Press, Baltimore) pp. 363–378.

Jia Hua Yuan et al. (1980), *Hanyu Fangyan Gaiyao*, [a synopsis of Chinese dialects] second ed. (Wenzi Gaige Chubanshe, Peking).