CONTINUOUS LINGUISTIC TONETIC REPRESENTATION USING POLYNOMIAL RESIDUALS

Shunichi Ishihara

Japan Centre (Asian Studies) and Phonetics Laboratory (Linguistics, Arts), The Australian National University.

ABSTRACT-Normalised F0 data from the two accent types of Kagoshima Japanese are used to argue for 1) a continuous (polynomial) rather than a discrete-mean representation, and 2) the superiority of parameters derived from polynomial residuals over standard deviation measurements in the modeling of tone. It is claimed that both are preferable for linguistic-phonetics and speech technology.

INTRODUCTION: LINGUISTIC-PHONETIC REPRESENTATION

According to Ladefoged and Maddieson (1986: 1), there are two major aims in linguistic phonetics. These are the identification of the phonetic features which specify the sounds within a given language or variety, and also of those features which underlie sound contrast between varieties. However, it is controversial how these features should be linguistic-phonetically presented. Wide-ranging views were presented from various authors in the theme of "Phonetic representation" in Vol. 18 of Journal of Phonetics (1990). These views include both cognitive and non-cognitive views; one based on production/perception; and a view based on abstract concept (cf. Ladefoged, 1990; Nearey, 1990; Keating, 1990; Rischel, 1990; Pierrehumbert, 1990; etc). Nolan (1982: 1) comments that "at least two kinds of phonetic representation should be kept distinct: one for the descriptive phonologist, and one for modellers of speech production (and, as it turns out, speech perception)". That is, the former kind of (linguistic-) phonetic representation—of which the prime purpose is description—first of all, needs to "contain(s) just enough information to compare the phonetic properties of one language with those of any other languages; and secondly needs to supply clarifications of why synchronic and diachronic sound patterns are as they are" (Nolan, 1982: 3). The latter type of phonetic representation should be beneficial in modelling the human speech mechanism both in production and perception, and further applicable to the field of speech technology. The necessity of different types of phonetic representations depending on what they are used for does not imply their total separation. On both theoretical and practical grounds, both 'linguistic-phonetic interpretation' and 'direct-physical interpretation' of speech are necessary (Nolan, 1998).

With the discussion of "phonetic representation", there has been a strong emphasis on segmental properties of speech. However, suprasegmental or prosodic properties of speech are just as important as segmental properties for the accounts of the human speech faculty. Fundamental frequency—which is one of the main acoustic-phonetic correlates of prosody—is taken to be the primary phonetic dimension expressing tonal contrast. In speech synthesis, it is generally believed that prosody, particularly fundamental frequency, is the key to the naturalness (Carlson and Granström, 1997).

Reflecting the different purposes of speech analysis, different speech encoding schemes have been developed. As with the phonetic representation of intonation or tones, what different schemes seem to have in common is that they seek for the description of the F0 shape by means of stylised approximation ['close-copy stylisation' (Pijper, 1983)]. In the process of stylisation of F0 curve, the F0 curve is represented by a sequence of discrete elements: inflection points (= target points) which are interpolated with a linear or curvilinear function.

On the other hand, in various linguistic-tonetic descriptive works (Rose, 1987, 1993; Phuong, 1981; Zhu, 1999; etc), a discrete-mean representation using mean normalised F0 values and standard deviation values has been used not only to show the contour shape of a certain tone's F0, but also to express the magnitude of expected variation in the population, further aiming at between-language/between variety comparison (Rose, 1993).

In the following sections, I will show a continuous presentation together with some residual related parameters as a possible tool for linguistic-phonetically describing tonal phenomena, and justify its possibility by applying it to actual data obtained from Kagoshima Japanese (KJ).

THE ACCENTUATION OF KAGOSHIMA JAPANESE

KJ exhibits a two way accentual contrast; (L)^{0}HL and (L)^{0}H (cf. Hirayama, 1960). In the former pattern, only the penultimate syllable of a word has a high pitch, and every other syllable has a low pitch. In the latter pattern, only the last syllable of a word has a high pitch, and every other syllable before it has a low pitch. Following Hirayama (1960), (L)^{0}HL type is referred to as Type A and (L)^{0}H type as Type B in this paper.

EXPERIMENT PROCEDURE AND NORMALISATION

Four native speakers of KJ (two females: TY and YN and two males: TT and NK aged between 25 and 35 years of age) participated in this study. One corpus containing 18 noun phrases differing in syllable number with approximately 15 dummy phrases which were scattered at random throughout the corpus was prepared. The informants were asked to read this corpus 10 times in the frame given in Table 1. The recording was conducted in a room in Sydney with very low ambient noise, and the reading material was recorded onto high-quality normal position tapes using professional equipment. The raw material was digitised with Computerised Speech Laboratory (CSL) (sampling rate = 10000 Hz).

Phrase	Adjective	+	Noun	Accent type	Meaning
AA	nagaka	+	kamikiribasami	Type A + Type A	long paper scissors
AB	nagaka	+	niwatorigoya	Type A + Type B	long hen house
BA	yoka	+	kamikiribasami	Type B + Type A	good paper scissors
BB	yoka	+	niwatorigoya	Type B + Type B	good hen house
Frame	Naomi-wa	_____	to itta	gayo.	
	LL H L		L HL	H L	
	Name-TOP	_____	QTN. say-past	SFP.	QTN = Quotation
	Naomi said	_____.			SFP = Sentence final particle

Table 1: Four target phrases and carrier sentence.

Only four noun phrases which are listed in Table 1 are relevant to this paper. All phrases have the same LHLLLLL(L)H(L) pitch configuration with different accentual types.

F0 was extracted from the HLLLLL(L) sequences—which can be represented by the interpolation of two target tones in Autosegmental-Metrical model (cf. Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988; etc)—using the CSL's Automatic Pitch Extraction. F0 samples were taken at the onset, 50% and the offset of each syllable nucleus except for the initial high pitched syllable from which only the maximum F0 value was sampled.

A logarithmic z-score normalisation—which Zhu (1999) reports the superiority of in F0 normalisation—was used in this study in order to exclude between-speaker differences and specify the invariant features (Rose, 1987). The log z-score normalisation procedure is: $F0_{norm} = (F0_i - x) / SD$, where $F0_i$ is a sampling point, x is the average F0 from all sampling points, and SD is the standard deviation around the mean of those points, all of which are logarithmic terms. Table 2 below contains the normalisation parameters of the four informants.

Speaker	TY	YN	TT	NK
x	2.133	2.013	2.284	2.262
SD	0.053	0.043	0.073	0.043

Table 2: Normalisation parameters in log F0

CONTINUOUS REPRESENTATION OF AA, AB, BA AND BB PHRASES

The F0 contour shapes appearing in the HLLLLL(L) sequences of AA, AB, BA and BB were continuously described using a two-degree polynomial line. A two-degree polynomial line was calculated from normalised F0 values plotted against absolute time using the 'least-square method'. In Figure 1 below, the obtained two-degree polynomial curve for AA is presented together with the mean normalised F0 curve and SDs (discrete-mean) which are plotted against the mean absolute duration.

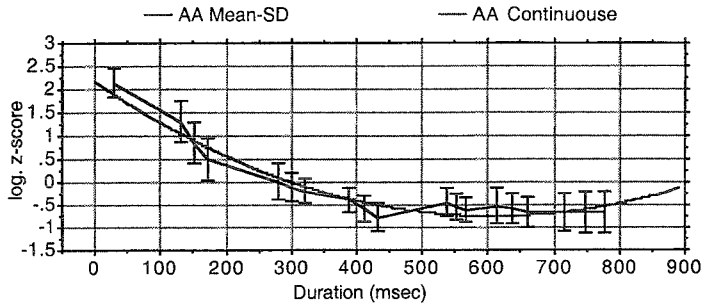


Figure 1: Mean normalised F0 values for AA, compared with two-degree polynomial fit. AA: $y = 2.165 - 0.010x + 7.837E - 6x^2$. $R^2 = 0.770$. X-axis is duration and Y-axis is normalised value.

As can be seen from Figure 1, these two descriptions show very similar contour shapes in that the two-degree polynomial line nicely remains within the SD corridor of the mean normalised F0 curve. Although this sort of mathematical and continuous representation is easily computable and applicable in speech technology, it is not yet adequate for linguistic-phonetics. The presentation using mean normalised F0 values and SDs is still superior in that it is expressive in terms of the magnitude of variations, however it suffers from a shortcoming in that the presentation allows any sort of unusual F0 movement as far as they are within the SD band (i.e. zigzag). Relying on a continuous description using a mathematical formula, there is no way to show the range of variations. Residual values and some parameters derived from them could be used not only to indicate the magnitude of variation appearing in a given linguistic feature but also to overcome the above mentioned shortcoming the discrete-mean representation has.

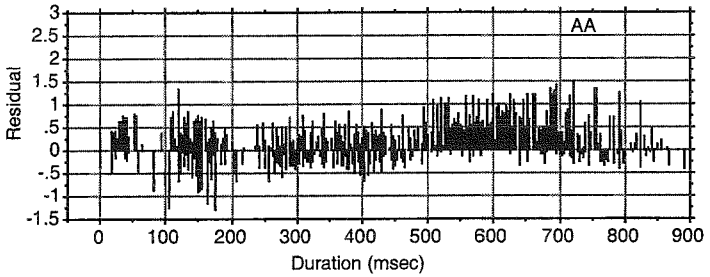


Figure 2: Residual values plotted against time (msec).

In Figure 2, all residual values calculated from all observed normalised F0 values (AA) are presented against time. Figure 2 shows that the residual values are not always distributed evenly around the zero level. That is, the observed normalised F0 values are not evenly scattered around the obtained two-degree polynomial fit. These statistics are useful, particularly at the micro level, not only to show the degree of variation appearing in the same linguistic feature but also to predict the magnitude of

variations if the same experiment is conducted under the same environment. The same residual values are pooled at each measuring point (the onset, 50% and the offset of each low pitched syllable nucleus, but only the point where maximum value was sampled for the initial high pitched syllable) and the mean residual values and SDs were calculated. These mean residual values and one SD above and below them are superimposed on the polynomial line in Figure 3 so that the graph is more visually expressive in terms of magnitude of variations.

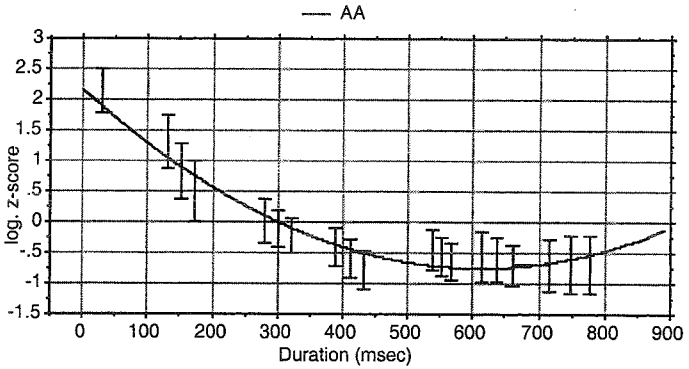


Figure 3: F0 curve and residual distribution around it.

Assuming normally distributed normalised F0 values, one SD above and below the mean will include approximately 68% of all observations. As seen from Figure 3, however, observed normalised F0 values are *not* evenly distributed around the polynomial fit. The group of F0 values sampled at the first measuring point, for example, shows a negative skew (= -0.445). In addition to Figure 3 which indicates the magnitude of variation at the micro level, it would be more appropriate if we could access the information concerning the magnitude of variation at the macro level. Three parameters were calculated from residual values. These are the sum of negative residual values (NegRSum), that of positive residual values (PosRSum) and the sum of Absolute residual values (AbsRSum) appearing in one token. These parameters indicate the degree of dispersion which is accepted in one token.

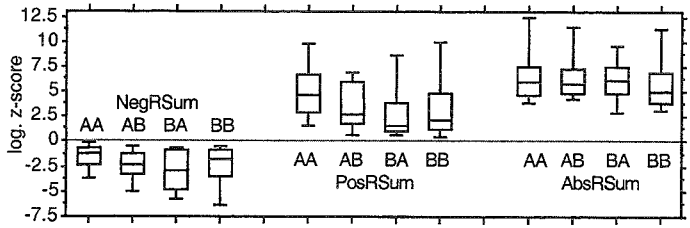


Figure 4: Box plots showing the percentiles of NegRSum, PosRSum and AbsRSum for each phrase type.

Figure 4 contains box plots showing the percentiles of NegRSum, PosRSum and AbsRSum for each phrase type. The numerical information of Figure 4 is presented in Table 3. According to Table 3, for example, -2.481 SDs appearing in the cross section of Column 75% and Row NegRSum AA means that 75% of all AA tokens have smaller dispersion value than -2.481 SDs for NegRSum. What one might notice from Figure 4 is that different phrase types exhibit rather different NegRSum and PosRSum values at each percentage point, whereas all phrase types share very similar AbsRSum values at each percentage point (particular similar at 25, 50 and 75% point). This result indicates that

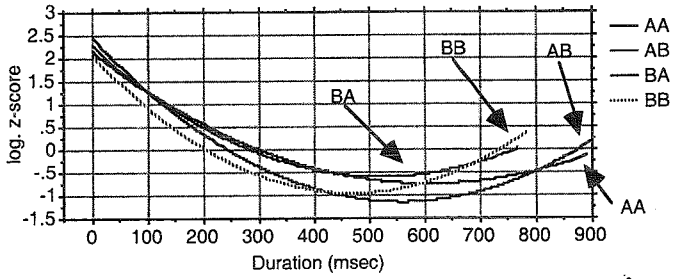
the sum of dispersions which are legitimately found in one token is more or less the same regardless of phrase types. These three parameters calculated from residual values have a macro aspect because they provide the information concerning the overall degree of dispersion appearing in one token. Therefore, these parameters can be used to specify the range in which all appearing variations can be perceived as utterances sharing the same linguistic information. The range specified by these three parameters is further useful for cross-language or variety comparisons.

Residual	Phrase	10 %	25 %	50 %	75 %	90 %
NegRSum	AA	-0.162	-0.601	-1.265	-2.481	-3.783
	AB	-0.512	-1.242	-2.419	-3.515	-5.047
	BA	-0.596	-0.936	-3.003	-5.018	-5.880
	BB	-0.450	-0.915	-1.819	-3.640	-6.378
PosRSum	AA	1.534	2.745	4.621	6.789	9.832
	AB	0.676	1.611	2.838	5.954	6.987
	BA	0.725	0.891	1.668	3.807	8.692
	BB	0.461	0.950	2.114	4.895	9.950
AbsRSum	AA	3.978	4.485	5.907	7.547	12.418
	AB	4.184	4.735	5.731	7.381	11.540
	BA	2.863	4.567	6.167	7.483	9.600
	BB	3.161	3.788	5.009	6.884	11.266

Table 3: The numerical information for Figure 4.

DISCUSSION

Although different phonetic representations can be employed to precisely match the different demands arising from different fields, a phonetic representation which simultaneously meets the different demands coming from a wide variety of fields would be more ideal. As mentioned above, the continuous representation using a polynomial regression line is stylised approximation using a mathematical formula. Therefore, this sort of mathematical information can be easily computed for speech synthesis. At the same time, continuous representation satisfies the demands of linguistic-phonetics.



AA: $y = 2.165 - 0.010x + 7.837E - 6x^2$ AB: $y = 2.462 - 0.013x + 1.152E - 5x^2$

BA: $y = 2.271 - 0.011x + 1.053x - 5x^2$ BB: $y = 2.070 - 0.013x + 1.367E - 5x^2$

Figure 5: The continuously represented F0 contour shape of each phrase.

A comparison based on the mean normalised F0 values [unpaired, two-tail t-test] shows that those phrases having a Type A word as the second component (AA and BA phrases) have significantly higher F0 realisation at their valley point than those phrases having a Type B word as the second component (AB and BB phrases) [AA vs AB: DF (1, 77) = 6.773, p = 0.0001; BA vs BB: DF (1, 76) = 5.032, p 0.0001]. This difference in F0 range identified on the basis of observed normalised F0 values can be visually identified in the continuous descriptions presented in Figure 5, and it was statistically confirmed in terms of the coefficient values (*a* = the coefficient value for the second power item and *b* = that for the first power item) of the two-degree polynomial regression lines [AA vs AB (*a*): DF (1, 75) = -

2.175, p= 0.0328; (*b*): DF (1, 75) = 3.143, p = 0.0024; BA vs BB (*a*): DF (1, 74) = -2.958, p = 0.0042; (*b*): DF (1, 74) = 2.794. p = 0.0066].

If one simply relies on observed normalised F0 values, the points at which one could conduct a statistical comparison are limited to each sampling point. One of the advantages of continuous representation is that comparisons can be statistically performed anywhere along the time scale because these values can be obtained from the mathematical formulae. Statistical comparisons in terms of minimum and maximum values are possible from the mathematical formulae, as well. Therefore, it can be said that continuous representation together with residual related parameters satisfies the requirement for phonetic (tonetic) representation not only in terms of the identification of a particular linguistic feature but also in terms of the magnitude of variations without having the same shortcomings of the discrete-mean representation.

ACKNOWLEDGMENTS

The author would like to thank Dr Phil Rose and two anonymous reviewers for their valuable comments.

REFERENCES

Carlson, R. and Granström, B. (1997). *Speech synthesis.* In Hardcastle, W. J. and Laver, J. (eds.), The Handbook of Phonetic Science. 768-88.
Hirayama, T. (1960). *Zenkoku akusento jiten.* Tokyo: Tokyodo.
Keating, P. A. (1990). *Phonetic representation in a generative grammar.* Journal of Phonetics 18. 321-34.
Ladefoged, P. (1990). *Some reflections on the IPA.* Journal of Phonetics 18. 335-46.
Ladefoged, P. and Maddieson, I. (1986). *Some of the sounds of the world's languages: preliminary version.* UCLA Working Papers in Phonetics 64.
Nearey, T. (1990). *The segment as a unit of perception.* Journal of Phonetics 18. 347-73.
Nolan, F. (1982). *The nature of phonetic representations.* Cambridge Papers in Phonetics and Experimental Linguistics vol. 1.
Nolan, F. (1998). *Phonological representation and phonetic interpretation in intonation analysis.* A paper presented at Sixth Conference on Laboratory Phonology.
Phuong, V. T. (1981). *The Acoustic and Perceptual Nature of Tone in Vietnamese.* Ph.D. thesis, The Australian National University.
Pierrehumbert, J. (1980). *The Phonology and Phonetics of English.* Ph.D. thesis, MIT.
Pierrehumbert, J. (1990). *Phonological and phonetic representation.* Journal of Phonetics 18. 375-94.
Pierrehumbert, J and Beckman, M. (1988). *Japanese Tone Structure.* Cambridge/ Mass.: MIT Press.
Pijper, D. R. de. (1983). *Modelling British English Intonation.* Dordrecht, Cinnaminson: Foris Publications.
Rischel, J. (1990). *What is phonetic representation?.* Journal of Phonetics 18. 395-10.
Rose, P. (1987) *Considerations in the Normalisation of the Fundamental Frequency of Linguistic Tone.* Speech Communication 6, 343-351.
Rose, P. (1993). *A linguistic-phonetic acoustic analysis of Shanghai tones.* Australian Journal of Linguistics 13: 2. 185-20.
Zhu, X. (1999). *Shanghai Tonetics.* Lincom Studies in Asian Linguistics 32. Lincom Europe.