A Model of Tone Systems

Jean-Marie Hombert

[Department of Linguistics
University of California, Los Angeles]

1. Introduction

1.1. The development of tone and tone space

The process by which linguistically contrastive tones can develop from the influence of neighboring segments is well attested and rather well understood (Brown 1965; Chang 1973, 1975; Haudricourt 1954, 1961; Hombert 1975, in press; Hombert, Ohala and Ewan 1976; Matisoff 1973; Mazaudon, 1977; Sarawit 1973). However, one of the questions which has been left unanswered by researchers working on tonogenesis is the following: do the shapes of the newly phonologized tones remain roughly similar to the fundamental frequency (Fo) shapes from which these tones originated (i.e. preserving the Fo perturbations due to the neighboring consonants which caused the tonal development) or, rather, do they evolve according to some more universal principles?

This paper represents an attempt to demonstrate that the tones are going to be reorganized in the tone space according to universal phonetic principles independently of the Fo shapes prior to the tonal development.

1.2. Articulatory difficulty vs. perceptual distinctiveness

The notion of minimum articulatory difficulty and maximum perceptual distance has been used for a long time to account for diachronic linguistic processes (Passy 1890; de Groot 1931; Martinet 1955). However these notions have rarely been used quantitatively. Lindblom and his colleagues (Lindblom and Sundberg 1969, 1971; Lindblom 1972, 1975; Liljencrants and Lindblom 1972) constructed a model predicting the distribution of vowels in the vowel space. The purpose of this study is to present a comparable attempt with tone systems. The model described here predicts the phonetically optimal tone systems from universal phonetic considerations (both articulatory and perceptual) without reference to the origin and the history of each tone system.

This paper is divided into two parts: first, some general tendencies observed in tone systems concerning the number and the shape of these tones are presented. Second, the model is described and its output is compared with actual data from tone languages.

2. Inventory of tone systems

A rapid survey primarily based on languages presented in Ruhlen's Guide to the Languages of the World (1975) and also on data from the
Phonology Archive Project (PAP), Stanford University, indicates that about 30% of the languages of the world are tonal. This figure would go up to about 50% if we would consider a sample of languages in which each language family would be represented by a number of languages proportional to the actual number it contains. This discrepancy between the two figures comes from the fact that most of the languages spoken in areas of great linguistic diversity (e.g., Chinese and Southeast Asian languages, Papuan New Guinea languages, and Northwestern Bantu languages) are tonal.

An approximate count of the distribution of tone systems as a function of the number of tones in each system is presented in Table 1.

<table>
<thead>
<tr>
<th>Systems with 2 tones:</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot; &quot; 3 &quot; &quot; :</td>
<td>30%</td>
</tr>
<tr>
<td>&quot; &quot; 4 &quot; &quot; :</td>
<td>15%</td>
</tr>
<tr>
<td>&quot; &quot; 5 &quot; &quot; :</td>
<td>10%</td>
</tr>
<tr>
<td>&quot; &quot; 6 or more:</td>
<td>15%</td>
</tr>
</tbody>
</table>

In most cases (90%) two tone systems are constituted by high and low tones. High, mid, and low tones represent the most common three tone system (65%), although high, low and falling systems are found (20%). Two types of four tone systems are found quite commonly: high, low, falling and rising (40%) and high, mid, low and falling (30%). High, mid, low, falling and rising constitute the most common five tone system (60%) and, finally, the most common six tone system \( \text{\textsuperscript{2}} \) (60%) is composed by two level tones (low and mid in most cases), two falling tones (high to mid and mid to low) and two rising tones (low to mid and mid to high). No clear patterns were found for seven and eight tone systems.

3. The model

3.1. Description of the model

The steps which should theoretically be involved are shown in Fig. 1. Considering the relatively small quantity of data available on physiological constraints involved in the production and perception of tones as well as the very limited number of acoustic descriptions of tone languages, it is clear that the model presented here should be regarded as a first attempt at predicting tone shapes from articulatory and perceptual considerations. The input to the model was constituted by the number of tones \( N \) in the system and by 25 tones representing all possible combinations of the two tone numbers (from 1 to 5) used to describe the beginning and the end of each tone (Chao 1930).\( \text{\textsuperscript{3}} \) The model generated all possible systems of \( N \) tones selected among the 25 possible tones. \( N \) was varied from 2 to 8.
FIGURE 1. Flow-chart of a model predicting tone shapes from articulatory and perceptual considerations
3.2. Articulatory data

In Hombert (1977) a first attempt at quantifying the articulatory difficulty associated with certain tone shapes was made. It was found that contour tones covering a small $F_0$ range (one step according to Chao's tone numbers) were probably articulatorily more complex than level tones or contour tones ending at one extremity of the $F_0$ range (i.e. 1 or 5). Considering the precariousness of these data, it was decided to simply discard the contour tones having a change of only one step (e.g. 12, 54, etc.) from the input set composed of 25 tones. However, contour tones ending at the extremity of the $F_0$ range were retained even when their $F_0$ change was only one step (i.e. 21 or 45). As a result, only 19 tones were considered as input to the perceptual component of the model:

5 level tones: 11, 22, 33, 44, 55
7 rising tones: 13, 14, 15, 24, 25, 35, 45
7 falling tones: 53, 52, 51, 42, 41, 31, 21

3.3. Tone perception

Data from experiments using pure tones (Brady et al 1961, Heinz et al 1968, Pollack 1968, Nabelek and Hirsh 1969, Nabelek et al 1970, Tsumura et al 1973) indicate that the end point of the tone plays a more important perceptual role than the beginning of the tone. In a recent study of tone perception involving English, Thai and Yoruba subjects, Gandour and Harshman (1977) used multidimensional scaling techniques to tease out the perceptual dimensions used by their subjects in making their similarity judgments on synthesized stimuli. Using PARAFAC analysis (Harshman 1970) they extracted five dimensions interpretively labeled (1) Average Pitch, (2) Direction, (3) Length, (4) Extreme endpoint, and (5) Slope. Of these five dimensions, Gandour and Harshman claim that average pitch and extreme endpoint are nonlinguistic dimensions. Hombert (1976) used the same PARAFAC technique to investigate the perception of tone patterns associated with bisyllabic (V1CV2) meaningful nouns in Yoruba. Two dimensions were extracted, the first one correlated clearly with the direction of $F_0$ change (falling, level, rising). The interpretation of the second dimension is not clear; the magnitude of the slope of $V_2$ seems to be the more important factor. However, results of lineal regression analysis indicate that the averaged fundamental frequency of $V_2$ and the difference between fundamental frequency offset of $V_1$ and the onset of $V_2$ are also included in dimension 2.

It is possible that with a larger number of subjects these factor would have been extracted as separate dimensions.

3.4. Computation of perceptual distance

Considering these experimental results, it was decided to use four perceptual factors: the average $F_0$, the onset, the offset, and the slope of the tone. Since the sign of the slope was considered, this last factor includes direction of change as well. Distinction based on duration was not considered here.
The perceptual distance between any two tones (taken from the set of 19 tones presented in 3.2) was then computed in two ways:

a. city-block distance:
\[ D_{cb} (T_{ij}, T_{kl}) = \sum_{m=1}^{4} \alpha_m \Delta f_m (i,j,k,l) \]

b. euclidian distance:
\[ D_e (T_{ijk}, T_{kl}) = \sqrt{\sum_{m=1}^{4} \alpha_m \Delta f_m^2 (i,j,k,l)} \]

with
- \( \Delta f_1 \) = difference in average \( F_0 \) between \( T_{ij} \) and \( T_{kl} = \frac{|i+k|}{2} - \frac{k+l}{2} \)
- \( \Delta f_2 \) = difference at onset between \( T_{ij} \) and \( T_{kl} = |i-k| \)
- \( \Delta f_3 \) = difference at offset between \( T_{ij} \) and \( T_{kl} = |j-1| \)
- \( \Delta f_4 \) = difference in slope between \( T_{ij} \) and \( T_{kl} = |(i-1) - (1-k)| \)

Theoretically, weighting factors obtained from multidimensional scaling analyses could have been used instead of arbitrary values. But since the available data are very restricted, it was decided to test several sets of weighting factors. These eight sets are presented in the left hand column of Table 2.

3.5. Optimization

Given the number of tones \( [N] \) in the system and the set of weighting factors \( [\omega] \), the distance \( D (T_{ij}, T_{kl}) \) was computed for each pair of tones for each possible combination of \( N \) tones selected among the 19 tones mentioned earlier. For each tone system considered, the smallest distance (DMIN) between two tones was stored. A system was considered perceptually optimum for a given set of input parameters (i.e. number of tones and set of weighting factors) when it was found to have the greatest DMIN after all possible tone systems with \( N \) tones were compared. In other words, the chosen perceptual criteria was to keep the two closest tones of a given system maximally apart. A simple example will help clarify this procedure. Let us consider two three tone systems \( S_1 \) (11, 33, 55) and \( S_2 \) (11, 51, 55) and the set of weighting factors \( \alpha(1,1,1,1) \), i.e. \( \alpha_1 \) (average \( F_0 \)) = 1, \( \alpha_2 \) (offset) = 1, \( \alpha_3 \) (onset) = 1, \( \alpha_4 \) (slope) = 1.

The DMIN for each tone system can be computed as follows:

\[ S_1: \quad D(11,33) = \alpha_1 \Delta f_1 + \alpha_2 \Delta f_2 + \alpha_3 \Delta f_3 + \alpha_4 \Delta f_4 \]
\[ = 1 \times |1-3| + 1 \times |1-3| + 1 \times |1-3| + 1 \times |0-0| \]
\[ = 6 \]
\[ D(11,55) = 12 \]
\[ D(33,55) = 6 \]

thus \( \text{DMIN} (S_1) = 6 \)
\[ S_2: \quad D(11,51) = 2 + 4 + 0 + 4 = 10 \]
\[ D(11,55) = 12 \]
\[ D(51,55) = 10 \]

thus \[ \text{DMIN}(S_2) = 10 \]

Consequently, with this set of weighting factors (i.e. equal weight given to average F₀, onset of the tone, offset of the tone, and slope of the tone) and this type of perceptual distance, \( S_2 \) (11,51,55) would be perceptually preferable to \( S_1 \) (11,33,55).

2.6. Relative weight of perceptual and articulatory data

The total lack of data in this area forces us to disregard this step. It should be noted however that some implicit weighting decision was made when some of the tones of the complete set of 25 tones were discarded for articulatory reasons. At this stage, articulatory factors were given complete priority enabling these tones to be "saved" on perceptual grounds.

2.7. Output of the model

For each condition (i.e. N and o) the output was constituted by three types of data: (i) the number of systems which were perceptually equivalent under this condition; (ii) the shapes of the tones of these systems; and (iii) the minimum distance between any two tones for these systems (since these systems are perceptually equivalent, they have the same distance between their two closest tones).

The number of systems perceptually equivalent is presented in Table 2 when city-block distance is used, and in Table 3 when Euclidian distance is used.\(^7\) The predicted systems are in very good agreement with the most commonly found tone systems presented in section 2, that is, 55, 11, 15, 51 for the four tone system; 55, 33, 11, 15, 51 for the five tone system, and 2 level, 2 rising and 2 falling for the six tone system. The predictions were generally bad for the three tone system (i.e. we were not able to predict 55, 33, 11), and for the two tone system the predictions were split between 11, 55 (i.e. the correct prediction) and 15, 51. We will present a possible explanation for this difference in accuracy of prediction for two vs. three tone systems in a moment. It should be emphasized that these predictions are relatively independent of the set of weighting factors used and consequently can be considered as a perceptually stable solution. Let us notice however that the predictions are improved when a relatively heavier weight is attached to the average F₀ parameter for systems with small numbers of tones and when a relatively heavier weight is attached to the slope parameter for systems with larger numbers of tones.

Figure 2 shows the normalized distance between the two closest tones of all systems perceptually optimum as a function of the number of tones. The dots represent distances computed with the city-block
### TABLE 2. Number of systems perceptually equivalent - city block distance

<table>
<thead>
<tr>
<th>average F_o</th>
<th>offset</th>
<th>onset</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### TABLE 3. Number of systems perceptually equivalent (Euclidian distance)

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
Fig 2. Normalized perceptual distance between the two closest tones of perceptually optimum systems as a function of the number of tones in the system.
distance, the asterisks represent Euclidian distances. Each point is
the average of eight distances obtained from the eight different sets
of weighting factors. Two important facts should be noted:

1. The two types of distances (city-block and Euclidian) give
extremely similar results as far as minimum perceptual distances are
concerned.

2. The two extremities of the graph deviate noticeably from the
almost linear relationship obtained for 3, 4, 5, and 6 tone systems.
a. The flattening of the curve obtained for 7 and 8 tone
systems indicates some sort of saturation in the tone space. A huge
number of tone systems are then found to be perceptually equivalent.
This implies high instability. These data are in agreement with the
fact that it is extremely rare to find tone languages with 7 or 8 tones
using only F_0 to distinguish these tones.
b. The fact that the minimum distance is much larger in the
case of two tone systems than for other systems suggests that we were
wrong in assuming that when a system has two level tones, these two
tones are 11 and 55. These data seem to indicate that these two tones
should be closer to each other. This may explain why we got more
accurate predictions for the two tone systems than for the three tone
systems. This discrepancy is only an artifact of the conventions we
used to code high and low level tones.

4. Conclusion

Although it is clear that articulatory, perceptual and acoustic data
are badly needed to improve and validate the model, the results presented
here are quite encouraging. They suggest that tone shapes can be predicted
with reasonable success knowing the number of tones in the system, only
on the basis of articulatory and perceptual data, and independently of
the origin and history of the particular systems considered. These re-
results suggest that (i) the perceptual parameter slope acquires more
importance when the number of tones increases; (ii) fundamental frequency
alone is not sufficient to code 7 or 8 tones (in these cases secondary
cues are used, e.g. phonation types, syllable structure); and (iii) the
spacing between the top and the bottom of the fundamental frequency range
is smaller for two tone systems than for systems with greater numbers
of tones.

Footnotes

1. In fact they used a "perceptual" representation of the acoustic space
by transforming the linear frequency scale into mel scale.

2. Since languages with more than five tones were poorly represented in
Ruhlen's and PAP's surveys, I used data from Tai (Brown 1965), Sarawit
(1973) and Miao-Yao languages (Chang 1973).
3. Relatively smaller numbers correspond to relatively lower \( F_0 \) (e.g. "51" would represent a tone falling from high to low).

4. We obviously do not imply that the voiced portion of the tone bearing unit has a perfect steady-state \( F_0 \).

5. It should be noted that if these two dimensions are considered non-linguistic, Candour and Harshman's dimensions do not allow measuring differences between level tones in a tone language.

6. Obviously these three factors are not independent: the slope can be derived from the onset and the offset of the tone for instance. Choosing slope as a separate factor implies that information can be provided by the slope itself.

7. Since the number of systems for 7 and 8 tone systems was extremely high, only the distance was computed in most of these cases.

8. This normalized distance was obtained by dividing the computed distance by the distance obtained for a two tone system.

9. Systems with more than 6 tones generally use secondary cues such as phonation types of syllable structure to distinguish some of the tones.

10. In a recent paper, Maddieson (1977) suggested that the frequency range used by speakers of tone language was a function of the number of level tones in the system (the greater number of level tones, the greater the frequency range). These data partially support and partially refute Maddieson's claim since they indicate that the distance between high and low is smaller in a two tone system than in a system with a greater number of tones but they do not indicate that the same increase in \( F_0 \) range is found between 3 and 4 or 4 and 5 tone systems.
References


Hombert, J.M. 1977. Difficulty of producing different Fo in speech. 
Journal of the Acoustical Society of America 60, 1, 544.


and queries. Report of the Phonology Laboratory 1, Berkeley, 48-77.

Proceedings of the 7th International Congress of Phonetic Sciences. 

the 8th International Congress of Phonetic Sciences, Leeds.

production and the distinctive features of Swedish vowels. STL- 
QF8R [KTH] Stockholm 1, 14-30.

Lindblom, B.E.F. and J. Sundberg. 1971. Acoustical consequences of 
ilp, tongue, jaw and larynx movement. JASA 50, 1166-79.

Maddieson, I. 1977. Tone loans, a question concerning tone-spacing 

Martinet, A. 1955. Economie des changements phonétiques. Berne: 
Francke.

Consonant types and tone. Southern Calif. Occasional Papers in 

Mazaudon, M. 1977. Tibeto-Burman tonogenetics. To appear in: Linguis-
tics of the Tibeto-Burman Area.

Nabelek, I.V. and Hirsh, I.J. 1969. On the discrimination of frequency 
transitions. JASA 45, 1510-1519.

changing frequency. JASA 48, 536-533.

Passy, P. 1890. Etude sur les changements phonétiques et leur 

