PERCEPTION OF PITCH AND TONAL TIMING: IMPLICATIONS FOR MECHANISMS OF TONOGENESIS

David House
Dept. of Speech, Music and Hearing, KTH, Stockholm, Sweden

ABSTRACT

This paper presents a discussion of tonal categories relating to pitch perception discrimination thresholds and the differences between pitch perception and the perception of tonal timing. Based on the results of a perception experiment using Swedish listeners and tonal timing results from the literature, this paper proposes perceptual mechanisms which may help explain the development of contour tones from level tones. In the perception experiments, listeners were given the task of ranking on a pitch scale /amamam/ stimuli in which rising and falling F0 contours through the second syllable differed in timing and slope. Listeners generally showed much greater tendency to perceive timing differences as pitch differences when the timing differences occurred early in the vowel. It is proposed that level tones may split into contour tones on the basis of their timing related to segments and syllable boundaries.

1. INTRODUCTION

Ever since Pike [23] presented a working definition of the difference between level tones and contour tones in tone languages, researchers have attempted to describe the differences between these two categories in a number of ways [1, 2, 8]. In perceptual terms, House [17, 18] proposes that a level tone is perceived as a pitch level in the vowel or syllable rhyme while a contour tone is perceived as a pitch movement through the vowel or syllable rhyme. Conditions are placed on the tonal movement such that optimal pitch movement perception is possible when the tonal movement occurs through an area of relative spectral stability in the vowel or syllable rhyme. Tonal movements through areas of spectral change are recoded as pitch levels. Thus the perceptual constraints imposed upon contour tones imply a precedence of level tones over contour tones. This is also what we would expect considering the distribution of the different tone types in the languages of the world [20].

Phonetic mechanisms which can explain the development of tones in languages have generally been proposed based on the development of level tones from segmental contrasts and concomitant F0 variation. In this explanation, high tone develops from syllables beginning with a voiceless obstruent and low tone develops from syllables beginning with a voiced obstruent [16]. Ohala [21] offers a perceptual explanation for this type of sound change where listeners perform a “dissociation” parsing error and fail to associate two related events in the speech signal.

Given that the psychophysical pitch discrimination ability of humans is extremely high with difference limens on the order of 1 Hz under 250 Hz for pure tones at 40 dB SL [12] we might expect to find languages making use of this ability and developing tonal systems in which several level tones could differ by slightly more than difference limens. Although this does not seem to be the case as microprosodic variation, intonation effects and cognitive limits probably preclude such rich tonal systems [29], there are tone languages in which a pair of level tones may differ in production by around 10 Hz, e.g. the mid and low tones of Thai [1, 10] and even by less than 10 Hz in perception, e.g. perception of high and low tones in Northern Kammu [9, 26, 27].

The above comparison between pitch sensitivity thresholds (difference limens) and perception of level tones is fairly straightforward. A comparison between pitch sensitivity thresholds for pitch movement and perception of contour tones is more complex. Efforts have been made to compare perception of glissando tones with the perception of static tones [3, 24] where perception of glissando tones involves an integration of the pitch movement. t'Hart [28] presents data which suggest that only differences in the amount of tonal change of more that 3 semitones are relevant for speech communication. However, the perception of tonal movement in contour tones within a given tonal system does not readily lend itself to comparison with psychophysical thresholds. For one thing, contrastive contour tones can demonstrate considerable diversification in terms of both tonal level and tonal configuration (e.g. a high falling tone compared to a low rising tone). Another consideration is that of the timing of the tonal configuration in relationship to the segmental structure of the syllable (e.g. an early fall versus a late rise).

In recent years, much research in the perception of tone and intonation has been directed at the question of tonal timing. This question is inviting in that, as a research paradigm, it allows hypotheses to be tested by the systematic manipulation of the timing of a tonal peak or a tonal rise or fall in relation to a predetermined segment boundary such as the onset of a stressed vowel. Since the stimulus sets consist of a continuum of stimuli, results can be examined and interpreted in terms of perceptual thresholds of timing. The purpose of this paper is to further investigate the relationship between the perception of tonal levels and tonal timing with a view towards explaining the development of contour tones from level tones.

2. METHOD

2.1. Stimuli

A total of twelve stimuli were generated using the KTH rule-based synthesizer [25]. The segmental context /amamam/ was used for six stimuli in which each segment was 100 ms in duration. For the other six stimuli the context /ama:mam/ was used. This was identical to the /amamam/ context except for the

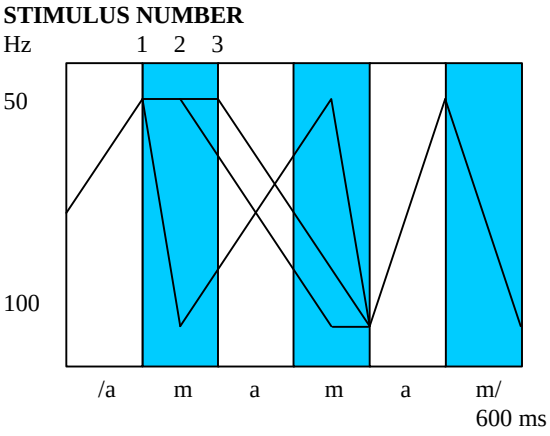

Figure 1. Stylized F0 contours for stimuli 1-3.

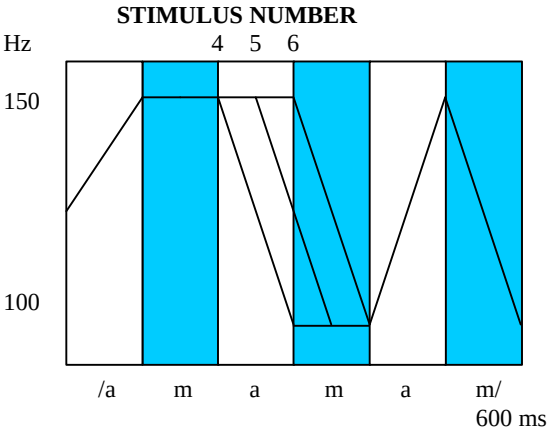

Figure 2. Stylized F0 contours for stimuli 4-6.

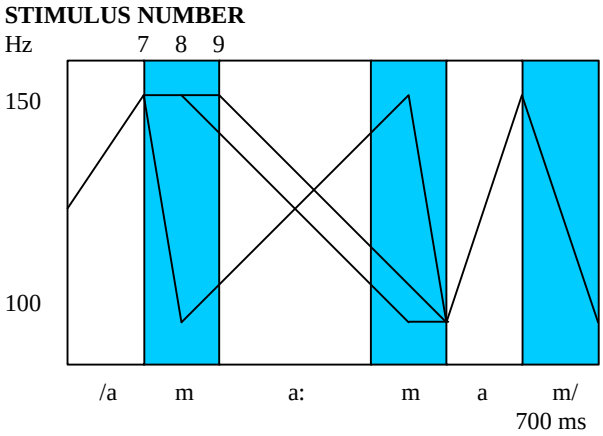

Figure 3. Stylized F0 contours for stimuli 7-10.

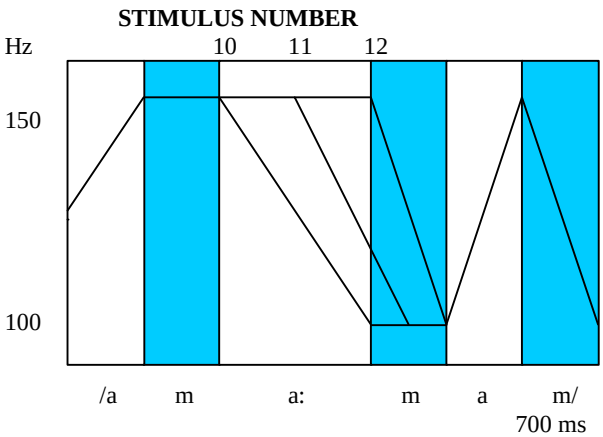

Figure 4. Stylized F0 contours for stimuli 10-12.

middle /a:/ which was doubled in duration to 200 ms. Duration was varied to investigate the effect of vowel duration on perceived pitch level when similar tonal contours vary in terms of their segmental durational frame. Various tonal configurations comprising rises and falls of differing slopes were generated for the stimuli and are presented in Figures 1-4. Stimuli are numbered from 1-6 for the /amamam/ stimuli and 7-12 for the /ama:mam/ stimuli.

2.2. Subjects

Thirteen subjects (nine male, four female), all native speakers of central Swedish, participated in the experiment. Ten subjects were students and three were staff members at KTH; none reported any hearing loss. The students participated in the experiment as part of a course requirement.

2.3. Test procedure and task

The experiment was conducted using an interactive computer-based program implementing a visual sort and rate method (VISOR) [13, 14]. In the program, the stimuli correspond to icons on the computer screen. The subject can click on the icons to listen to the stimuli and move the icons along a visual scale for sorting, rating and/or ranking.

Subjects were instructed to listen to the stimuli and determine if the middle syllable was short or long and then sort and rank the stimuli in terms of pitch. Subjects were asked to place each icon in the appropriate horizontal field corresponding to duration (short or long) and then to place the icons in position vertically corresponding to pitch (high=up and low=down). No specific instructions were given in terms of scale. The results thus reflect each subject's individual ranking of the stimuli in terms of pitch and rating in terms of duration. The duration task was included in the test as a separate task to encourage the subjects as an easy task which would counterbalance the more difficult task of ranking the pitch of the stimuli.

3. RESULTS

All subjects were able to separate short from long middle syllables for all stimuli. Individual ranking results were summed in order of lowest (1) to highest (12). Rank numbers obtained from each subject for each stimulus were added together giving a cumulative rank sum for each stimulus. The sum is presented in Figure 5 in rank order from lowest to highest pitch.

Stimulus 12, which maintains a high level throughout the long vowel was ranked as having the highest pitch, while stimulus 4, which falls early and quickly in the short vowel, was ranked as having the lowest pitch. The general trend of the pitch

ranking corresponds roughly to the fundamental frequency value integrated through the middle vowel of the stimuli. However, ranking is not solely a function of the integrated F0 as timing and configuration also play a role.

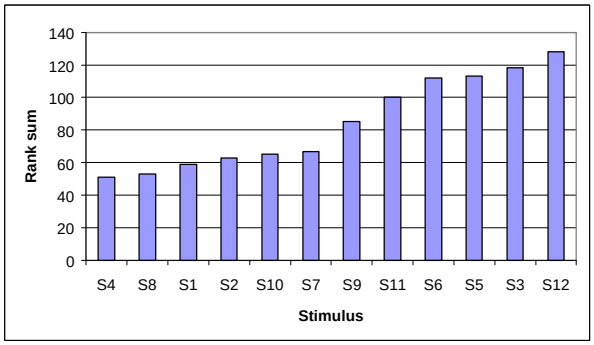

Figure 5. Rank sum for each stimulus presented in order from lowest perceived pitch to highest.

On the basis of the rank sums, the stimuli can be separated into two pitch categories: higher and lower. Stimuli 6, 5, 3 and 12 are all perceived as having higher pitch. These four stimuli have in common the fact that they are all falling and have at least half of their movement in the consonant. Stimuli 4, 8, 1, 2, 10 and 7 are perceived as having lower pitch. These six stimuli have most of their movement in the vowel and all begin their movement early in the syllable. Stimuli 9 and 11 are ranked between the two categories. Both have falling configurations through the final part of the long vowel into the following consonant.

Tonal direction, i.e. rising or falling movement, did not have much influence on the pitch ranking in the short vowel as stimulus 1 and 2 received nearly the same rank sum. In the long vowel, the rise (stimulus 7) was ranked as having higher pitch than the corresponding fall (stimulus 8).

It is also important to note that certain stimuli which are adjacent in their timing configurations are perceived as far from each other in pitch ranking. The stimuli pairs 2/3, 4/5, 8/9 and 10/11 are all separated in ranking. The first stimulus in each of these pairs has a timing configuration beginning at or near the vowel onset.

4. DISCUSSION

4.1. Differential sensitivity to tonal timing

The results presented above demonstrate a differential sensitivity to tonal timing related to the segmental frame. Timing differences of 50 ms occurring near the vowel onset are perceived as greater than differences occurring near the end of the vowel (e.g. stimuli pair 4/5 compared to 5/6). The results of this experiment could be related to the Swedish word accent distinction where low pitch corresponds to accent I and high to accent II. However, timing results in the literature from Swedish and other languages suggest that this differential sensitivity has a more universal nature.

4.2. Peak shift experiments

In a study on Swedish prosodic dialect typology, Bruce [5] tested the importance of the absolute timing of word accents for perception of dialect type. By systematically shifting two F0 peaks in steps of 25 ms through an utterance, he was able to create the percept of passing through four different dialect types. Timing differences between successive peaks representing clear dialect differences are on the order of 100 ms. Bruce also reports, however, that sociolectal differences can be perceived within dialects as the peaks are shifted. This would tend to indicate sensitivity to timing differences of less than 100 ms.

To address the question of categorical perception of intonation contours, Kohler [19] shifted the F0 peak position in a German utterance creating a change in meaning from 'established' to 'new' to 'emphatic'. In the experiment, the peak was shifted in steps of 30 ms. The results showing one ambiguous stimulus indicated that listeners needed two stimulus steps across the category boundary to perceive a category difference from 'established' to 'new'. Categorical perception was also confirmed but here again, a two-step pairing was needed to produce more than 50% difference responses. These results suggest sensitivity to timing differences in the neighborhood of 50 ms.

Categories of tonal timing have been investigated in English by Pierrehumbert and Steele [22]. They created a series of stimuli by varying the F0 peak in increments of 20 ms through the word 'millionaire' in the phrase 'only a millionaire'. Using an experimental method in which subjects imitated the stimuli, they were able to demonstrate the existence of two categories of peak delays in the responses of the subjects (categories viewed by the authors as assertion versus incredulity). An examination of the timing of the stimuli at the category boundaries for most of the speakers/listeners reveal that three stimulus steps are necessary to elicit a complete category shift, in other words a sensitivity to timing differences of around 60 ms.

Gårding and Eriksson [11] explored the perception of prosodic phrase patterns in Swedish by shifting the F0 peak in steps of 20 ms through a phrase resulting in three different categories: deaccented, accented and a compound. Although there was not a complete separation of the categories, the contours with the highest scores represented peaks differing by between 50 and 60 ms.

Finally, in an investigation of the perception of question and statement intonation in Neapolitan Italian, D'Imperio and House [7] shifted the final F0 peak in a statement in 33 ms steps forward through the vowel. This manipulation created two distinct perceptual categories: statement and question. However, the middle stimulus was ambiguous, thereby suggesting a differential timing sensitivity of about 70 ms.

4.3 Fall and rise shift experiments

To investigate perceptual boundaries for the Swedish word accents, Bruce [4] shifted the F0 fall in 10 ms steps through the 'ma' sequence in the phrase 'inga malmer'. Depending on the timing of the fall, accent I is perceived on the word 'malmer' (woman's name) or accent II (ores). In these experiments timing sensitivity was somewhat greater than in the peak shift experiments reported above. Here sensitivity was around 30 ms. House [17] using an ABX psychophysics test paradigm, shifted a falling F0 contour in 25 ms steps through the vowels and consonants of nonsense /amama/ and /ama/ utterances. In these tests, sensitivity for category shifts was on the order of 50 to 75 ms depending on the spectral environment. Hermes [15] investigated accentuation of nonsense /mamamamama/

utterances by varying the timing of rises and falls in steps of 20 ms. Here, sensitivity was on the order of 80 ms.

4.4 Implications for mechanisms of tonogenesis

In the data presented above, languages use differences in tonal timing for quite a variety of linguistic categories. A common denominator of many of the experiments seems to be a timing sensitivity of around 50 ms at category boundaries. In general terms of speech perception, 50 ms is a relatively long time span when compared to the rapid events of release bursts and formant transitions processed for stop identification. More impressive is pitch sensitivity for level tones on the order of 10 Hz presented in the introduction and the data on auditory sensitivity for pitch.

This difference between sensitivity for pitch and for tonal timing leads to speculation concerning the different perceptual mechanisms involved. The perception of level tones seems to be related to the acute sensitivity for pitch in the human auditory periphery, while sensitivity for tonal timing seems to imply a different kind of perceptual processing.

These different mechanisms may help explain the development of contour tones from level tones. We can consider a scenario whereby a language having level tones develops timing differences between the tones, for example a fall early in the syllable for a low tone and a late fall for a high tone. If this movement is consistently within spectrally unstable portions of the syllable at segmental boundaries, level tone perception will persist. If, however, movement occurs within spectrally stable portions of the syllable, especially within a long vowel (e.g. stimuli 9 and 11 in the current experiment), tonal movement perception could lead to a tone split from a level tone into a level tone and a contour tone.

Level tones can thus rely on pitch perception alone or on a combination of pitch and timing perception as illustrated by the current experiment. Contour tones, however, must rely on both pitch perception and timing perception related to segmental boundaries. Contour tones, therefore, may involve a more complex perceptual mechanism implying higher order cognitive processing and short-term memory such as the precategorical acoustic storage presented in Crowder and Morton [6]. This could help account for the predominance of level tones over contour tones in the languages of the world.

REFERENCES

[1] Abramson A.S. 1962. The vowels and tones of Standard Thai: Acoustical measurements and experiments. *International Journal of American Linguistics* 28 (No. 2 Part III).

[2] Abramson A.S. 1976. Static and dynamic acoustic cues in distinctive tones. *Haskins Laboratories Status Report on Speech Research*. New Haven, Conn., SR-47, 121-127.

[3] d'Alessandro C., Tosset S. and Piot O. 1995. Measurement of pitch perception for F0 glides. *Proceedings of Eurospeech '95, 4th European Conference on Speech Communication and Technology. Madrid*, 957-960.

[4] Bruce G. 1977. *Swedish word accents in sentence perspective*. Lund: Gleerup.

[5] Bruce G. 1983. Accentuation and timing in Swedish. *Folia Linguistica* 17, 221-238.

[6] Crowder R.G. and Morton J. 1969. Precategorical acoustic storage (PAS). *Perception and Psychophysics* 5, 365-373.

[7] D'Imperio M. and House D. 1997. Perception of questions and statements in Neapolitan Italian. *Proceedings of Eurospeech 97, 5th European Conference on Speech Communication and Technology*, 251-254, Rhodes, Greece.

[8] Gandour J.T. 1978. The perception of tone. In Fromkin V.A. (ed) *Tone: a linguistic survey*. 41-76. New York: Academic Press.

[9] Gandour J.T., Gårding E. and Lindell K. 1977. tones in Northern Kammu: a perceptual investigation. *Acta Orientalia* 39: 181-189.

[10] Gandour J.T., Potisuk S., Ponglorpisit S., and Dechongkit, S. 1991. Inter- and intraspeaker variability in fundamental frequency of Thai tones. *Speech Communication* 10, 355-372.

[11] Gårding E. and Eriksson L. 1991. On the perception of prosodic phrase patterns. *Working Papers, Department of Linguistics, Lund University* 38, 45-70.

[12] Gelfand S.A. 1981. *Hearing, an introduction to psychological and physiological acoustics*. New York: Marcel Dekker, Inc.

[13] Granqvist S. 1996. Enhancements to the visual analogue scale, VAS, for listening tests. *Speech, Music and Hearing. TMH QPSR 4/1996*: 61-65 KTH Stockholm.

[14] Granqvist S. 1999. Spruce signal workstation add-on package. Stockholm, Sweden. AB Nyvalla DSP http://www.nyvalla-dsp.se/

[15] Hermes, D.J. 1997. Timing of pitch movements and accentuation of syllables in Dutch. *Journal of the Acoustical Society of America* 102, 2390-2402.

[16] Hombert, J.-M., Ohala, J.J. and Ewan, W.G. 1979. Phonetic explanations for the development of tones. *Language*, 55, 37-58.

[17] House D. 1990. *Tonal perception in speech*. Lund: Lund University Press.

[18] House D. 1996. Differential perception of tonal contours through the syllable. *Proceedings of the International Conference on Spoken Language Processing. ICSLP 96*, Philadelphia, 2048-2051.

[19] Kohler K.J. 1987. Categorical pitch perception. *Proceedings of the Eleventh International Congress of Phonetic Sciences*, Tallinn, 5:331-333.

[20] Maddieson I. 1978. Universals of tone. In Greenberg (ed) *Universals of Human Language, Volume 2, Phonology*, 335-365. Stanford: Stanford University Press.

[21] Ohala, J.J. 1993. Sound change as nature's speech perception experiment. *Speech Communication*, 13, 155-161.

[22] Pierrehumbert J.B and Steele S.A. 1989. Categories of tonal alignment in English. *Phonetica* 46, 181-196.

[23] Pike K.L. 1948. *Tone Languages*. Ann Arbor: University of Michigan Press.

[24] Rossi, M. 1971. Le seuil de glissando ou seuil de perception des variations tonales pour les sons del la parole. *Phonetica* 23, 1-33.

[25] Sjölander, K., Beskow, J., Gustafson, J., Levin, E., Carlson, R., and Granström, B. 1998. Web-based educational tools for speech technology, In *Proceedings of ICSLP'98*, Sydney, Australia.

[26] Svantesson J.O. and House D. 1996. Tones and non-tones in Kammu dialects. *Proceedings of Fonetik 96, Swedish Phonetics Conference, TMH-QPSR* 2/1996, 85-87.

[27] Svantesson J.O. and House D. Forthcoming. Tone production, tone perception and Kammu tonogenesis.

[28] 't Hart J. 1981. Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical society of America* 69, 811-821.'

[29] 't Hart, J., Collier, R., and Cohen, A. 1990. *A perceptual study of intonation. An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.