COMPUTER MODELLING OF LEXICAL TONE PERCEPTION

Fangxin Chen and Anton J. Rozsypal
Department of Linguistics, University of Alberta

1. Introduction

Theoretical importance and practical implications of research on computer speech recognition generate considerable interest in this field among linguists, psychologists, and communication engineers alike. The long-term aim is to design a model reflecting the speech perception process and eventually simulate it by a computer program able to perform speaker-independent speech recognition. The present study focuses on one aspect relevant for tone languages, the recognition of lexical tones. The language chosen is Mandarin Chinese (MC).

MC is basically a monosyllabic contour-tone language. A syllable encoded with a lexical tone constitutes an independent semantic unit (morpheme). There are four lexical tones in MC citation form: level tone (Tone 1), rising tone (Tone 2), falling-rising tone (Tone 3) and falling tone (Tone 4).

To build our tone-perception model, two aspects of tone were investigated: perceptual dimensions and domain. The perceptual dimensions specify which acoustic cues contribute to the lexical tone perception and how these cues are represented in the auditory system. Tone domain determines where in the syllable the lexical tone is physically located.

For signal editing, analysis and subject testing, the Alligator program, developed in our Department, was used. The stimuli were synthesized on the Wavelet Speech Synthesizer, one of the Alligator signal generating modules (Rozsypal, 1987).

2. Perceptual Dimensions of Tone

To explore the perceptual dimensions of tone, three experiments were conducted. The first experiment examined how pitch contour determines tone perception. A continuum of MC syllables were synthesized with fixed duration and initial fundamental frequency f_i and linear pitch contour slopes ranging from negative to positive slopes. The stimuli were presented to the Mandarin-speaking subjects for tone category identification. The results show that although lexical tone perception relies primarily on auditory detection of linear frequency glides, phonetic tone categorizations do not coincide exactly with the auditory thresholds for non-linguistic tone detection. This suggests that lexical tone is processed not merely by the peripheral auditory system, but that a more central levels of the auditory system must be involved. The asymmetry between the frequency thresholds for rising and falling tone detection also suggests that articulatory constraints on tone production plays an important role in the formation of lexical tone categories. The second experiment explored the interrelationship among pitch contour slope, duration, and initial frequency in lexical tone perception. Two continua of synthesized MC syllables were prepared. In the first continuum, f_i was fixed at 100 Hz; the duration was increased from 40 ms to 200 ms in 20 ms steps; the final fundamental frequency f_f was increased from 100 Hz to 150 Hz in 5 Hz steps, except for syllables with 40 ms duration, in which case f_f was increased from 100 Hz to 200 Hz in 10 Hz steps. In the second continuum, f_i was set to 200 Hz; duration was increased from 40 ms to 200 ms in 20 ms steps; f_f was increased from 200 ms to 250 ms in 5 Hz steps, except for syllables with 40 ms duration, for which f_f was increased from 200 Hz to 300 Hz in 10 Hz steps. These two sets of continua were presented to the subjects for tone identification. The results indicate that the interrelationship among pitch slope, tone duration and the initial frequency can be described by the formula

$$\frac{\Delta f \, T}{\log f_i} = C_j \, ,$$

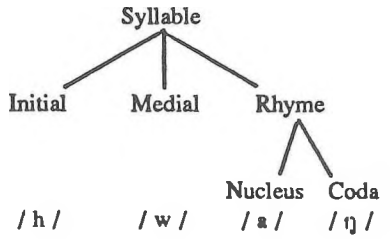
where Δf is the difference between f_f and f_i of the pitch glide and T is the tone duration. The constant C_j assumes a particular value for each tone category denoted by the index j. This equation indicates that the frequency shift threshold Δf for tone detection is inversely proportional to the tone duration T and directly proportional to the logarithm of f_i. The validity of this formula is restricted: the duration of the glide T must be at least about 40 ms long. The minimal duration is required for perceiving a stable tone (Doughty & Garner, 1948). This means that for a male voice range the syllable must be about four periods or longer to have a tonal quality. For a female voice range, the duration threshold of tonality is about seven periods. On the other hand, for tone duration greater than 100 ms, Δf becomes fairly constant and will not change markedly. Similar results were reported in a non-linguistic tone experiment (Dooley & Moore, 1988). This relationship between Δf and T leads to the hypothesis that the human listener's temporal window in lexical tone processing is about 100 ms. The third experiment was conducted to determine whether there is any correlation between intensity contour and lexical tone perception. A list of 28 MC syllables in four tones, read by ten Mandarin-speaking subjects of both sexes, were recorded. The syllables were digitized and their intensity contours extracted. These were classified into nine basic intensity patterns. The matching of intensity contour and tone category indicated that MC syllables with falling-rising and falling-level intensity contours were associated only with Tone 3. Consequently, they can be used as a reliable acoustic cue for Tone 3 perception.

3. Domain of Tone

Before discussing the domain of MC tone, a brief description of MC syllable structure seems appropriate. This structure can be expressed by the following syllable structure rules, where elements within square brackets are optional:

Syllable - [Initial] + [Medial] + Rhyme;
Initial - Consonant;
Medial - Glide (j, w, ɥ);
Rhyme - Nucleus + [Coda];
Nucleus - Vowel;
Coda - Nasal (n, ŋ);

Following is an illustration of how the MC syllable 'huang /hwaŋ/' can be diagrammed by the MC syllable structure rules:



To establish the domain of lexical tones, two approaches were taken. In a perceptual experiment, a list of natural MC syllables in the four tones were recorded and digitized. Either the Initial, Medial or Coda was gated out from the syllables and only the remainders were presented to the subjects. The subjects were then asked to indicate the tone categories of the remainders of the syllables. According to the confusion matrices based on the original tone categories of the syllables and the reported tone categories judged by the subjects, the removal of Initial or Medial had little effect on subjects' tone perception, while the elimination of Coda seriously affected subjects' tone judgements. In the second approach, the pitch contours of a list of natural MC syllables in the four tones were extracted and the characteristic lexical tone patterns were located in the segmental environment. The results of both approaches indicate that the relevant tone contour pattern is contained in the Rhyme part of a syllable, which confirms Howie's (1974) argument. In speech production, the speaker's intended tone might be 'distorted' by the segmental environment, mainly the initial voiced consonant, due to articulatory factors. Intrinsic tone perturbations of this nature appear to have been established as tone variants through language exposure.

4. Computer Tone Recognition

Based on the above findings, we designed a computer program written in Turbo Pascal language, modelling the perceptual aspects of linguistic tone recognition. A 100 ms sliding window was used to simulate the human ear's temporal window in scanning the pitch contour of the input syllable. The pitch slope within this window was computed according to the following formula for the linear regression slope,

$$Slope = \frac{n \Sigma t_j f_j - \Sigma t_j \Sigma f_j}{n \Sigma t_j^2 - (\Sigma t_j)^2},$$

where n is the number of periods within the moving window, t_j are the time values for the end of the j-th period within the moving window, and f_j are the corresponding frequency values. The period index j ranges from 1 to n. In case the pitch glide duration was shorter than 100 ms, the remaining t_j values within the window were increased by constant steps of the average period duration of the segment and the corresponding 'missing' f_j values were set to the initial frequency value f_i. This automatically reduced the slope value within the moving window, reflecting the interrelationship between duration and frequency change in tone detection. The critical slope boundary values for the detection of Tone 1 (level tone), Tone 2 (rising tone), and Tone 4 (falling tone) were determined by analyzing 360 words in citation form spoken by ten Mandarin speakers of both sexes and normalized with respect to the initial fundamental frequency. Recognition of Tone 3 (falling-rising) required detection of both falling and rising pitch slope, where the falling pitch slope must precede the rising one. In tone scanning, the pitch contour within the initial voiced consonant was disregarded as irrelevant to lexical tone recognition.

Besides pitch contour, the tone recognition program also analyzed the intensity contour of the input syllable. The input syllables were divided into three equally long parts and the average values of intensity in each part were calculated. According to the ratios of average intensity values between each two parts, about nine different intensity contours could be identified, in which falling-rising or falling-level intensity contours served as reinforcement for Tone 3 decisions.

To test the reliability of the above procedure, a list of 180 words in citation form with different tones spoken by five new speakers was recorded and submitted to the tone recognition procedure. A correct recognition rate of 98% was obtained. The only confusions encountered were between Tone 2 and Tone 3, the same confusion pattern as found in human listeners (Kiriloff, 1969; Zue, 1976).

References

Doughty, J.M., and Garner, W.R. (1948). "Pitch Characteristics of short Tones. II. Pitch as a function of tonal duration," J. Exp. Psych. **38**, 478-494.

Blicher, D.L., Diehl, R.L., and Cohen, L.B. (1990). "Effects of syllable duration on the perception of the Mandarin Tone 2 / Tone 3 distinction: Evidence of auditory enhancement," Journal of Phonetics, **18**, 37-49.

Diehl, R.L, and Walsh, M. A. (1989). "An auditory basis for the stimulus-length effect in the perception of stops and glides," J. Acoust. Soc. Am., **85**, 2154-2164.

Dooley, G.J., and Moore, B.C.J. (1988). "Detection of linear frequency glides as a function of frequency and duration," J. Acoust. Soc. Am. **84**, 2045-2057.

Howie, J.M. (1974). "On the domain of tone in Mandarin," Phonetica, **30**, 129-148.

Kiriloff, C. (1969). "On the auditory perception of tones in Mandarin," Phonetica, **20**, 63-67.

Kewley-Port, D., Watson. C.S., and Foyle, D.C. (1988). "Auditory temporal acuity in relation to category boundaries; Speech and nonspeech stimuli," J. Acoust. Soc. Am., **83**, 1133-1145.

Rozsypal, A.J. (1987). "Wavelet speech synthesizer," *Proceedings, Acoustics Week '87*, 62-67. Canadian Acoustics Association, Calgary, Alberta.

Zue, V.W. (1976). "Some perceptual experiments on the Mandarin tones," J. Acoust. Soc. Am. **60**, Suppl.1, S45 (Abstract).