THE EVALUATION AND INTEGRATION
OF PITCH HEIGHT AND PITCH CONTOUR
IN LEXICAL TONE PERCEPTION IN MANDARIN CHINESE

Dominic W. Massaro, Michael M. Cohen, Chiu-yu Tseng
Program in Experimental Psychology
University of California, Santa Cruz

ABSTRACT:

This study investigated the evaluation and integration of pitch height and pitch contour in the perception of tones in Mandarin Chinese. A fuzzy logical model of perception was used as a framework for the study. An important theoretical assumption was that speech sounds are perceived in terms of continuous acoustic properties. The experiment provided an assessment of the relative importance of the F0 height and the F0 contour during the vowel of a syllable tone perception, both the F0 contour and the F0 height influenced the perceptual recognition of tone. The cues were about equally effective. As in the perception of many other contrasts, lexical tone appears to be recognized on the basis of at least two cues, and the influence of one cue is the largest when the other cue is relatively ambiguous.
INTRODUCTION:

The study of speech perception has been influenced by traditional linguistic theories such as the distinctive feature theory of phonetic differences (Jakobson, Fant and Halle, 1967). According to this theory, a small number of features are assumed to be sufficient to classify all phonemes of a language. In addition, the features are defined in terms of binary distinctions in that a feature is either present or absent in a speech sound. Following a similar logic, much research on speech perception has focused on the all-or-none properties of acoustic features. The assumption is that a listener simply determines whether or not the relevant acoustic features are present or absent in the speech sound.

The study of speech perception has also been influenced by the idea of "categorical perception" (Liberman, Harris, Hoffman, and Griffith, 1957; Liberman, Cooper, Shankweller, and Studdert-Kennedy 1967; Studdert-Kennedy, Liberman, Harris, and Cooper, 1970), which predicts that discrimination of speech is dependent on category assignment. Small steps along an acoustic continuum should produce perceptible differences when they occur between phonetic categories, but not when they occur within a phonetic category. Accordingly, listeners are assumed to hear a sound as /b/ or /v/ but not hear some intermediate sound such as somewhat /β/.

In recent years, the study of speech perception has been influenced by an information processing approach (Mazurek, 1978; Chen and Massaro, 1978; Massaro and Oden, 1980; Thurlow, 1981). Information-processing models assume that speech perception begins with the acoustic stimulus and involves a sequence of internal processing stages before understanding occurs. Each stage of information processing operates on the information available to it and takes the information available to the next stage of processing. The processing stages are logically successive although they can overlap in time. The stages of processing most relevant to speech perception include feature detection and perceptual categorization.

Oden and Massaro (1978) proposed a fuzzy logical model of speech perception within the context of Massaro's (1975) information processing model. Three operations are defined: feature evaluation, prototype matching, and pattern classification. During feature evaluation, listeners evaluate the degree to which each relevant feature is present in the speech sound. In contrast to traditional linguistic description and categorical perception, it is assumed that listeners derive information about the degree to which each feature is present in the speech sound.
Whereas the traditional linguistic description assumed either the presence or absence of a particular feature, the featural information in this model indicates the degree to which the relevant property is present in the speech sound.

Featural information is made available to the prototype matching operation, which matches the features against prototypes of speech sounds stored in long-term memory. Each prototype is a proposition or a logical expression whose terms correspond to the ideal values of the acoustic features representing a particular speech sound. The outcome of prototype matching provides a goodness value indicating the degree to which each prototype matches the speech sound. Pattern classification is a function of the relative goodness values: the probability of identifying a stimulus as a particular speech sound is equal to its goodness value relative to the sum of the goodness values for all prototypes under consideration.

Central to the fuzzy logical model is the manner in which the acoustic features are integrated together in perceptual recognition. Most past research did not address the integration problem because the primary method of study involved experiments in which the speech sound was varied along a single relevant dimension. Very few experiments independently varied more than one variable within a particular experiment. Therefore, little information was available about how these cues were integrated into a synthesized percept. To address this problem, Narasaro and his colleagues (e.g., Narasaro and Cohen, 1976) and workers at Hawkins Laboratories (e.g., Repp, 1978) have carried out a number of experiments aimed more directly at the study of the integration of acoustic features in speech perception. The experiments use factorial designs in creating synthetic speech sounds varying independently along two or more dimensions.

Narasaro and Cohen (1976) tested various models of how the acoustic features are integrated together to arrive at a perceptual decision. Additive and multiplicative integration rules were tested within the framework of the fuzzy logical model. The difference between these two rules involves the relative contribution of the acoustic features as a function of their information value. According to the additive rule, a given feature value makes the same contribution regardless of the information value of the other features. The multiplicative rule has the consequence that the contribution of a particular feature value is directly related to the ambiguity of the other available features. The results from a wide variety of experiments were best described by the multiplicative combination rule (Narasaro and Cohen, 1980). The use of factorial designs and the fuzzy logical model appears to offer a promising approach to understanding the psychological processes involved in the perceptual recognition of speech sounds.
The present study extends the same paradigm to tones in Mandarin Chinese, to answer which acoustic features are functional and how the cues are integrated in perception. It has been assumed in linguistic theory that the principal phonetic features of tone are found in the domain of pitch, whose primary acoustic correlate is the fundamental frequency (F0). The term "tone" refers to a particular way in which pitch is utilized in language. A tone language is a language that utilizes pitch to contrast individual lexical items or words (McCawley, 1978). This definition includes the traditional tone languages of Africa, the Americas, and Asia as well as the marginal tone (or "pitch accent") languages of Europe and Japan; and excludes intonation languages, like English, in which pitch is used to signal syntactic and/or semantic distinctions at the phrase or sentence level (Gandour, 1978).

There are four lexical tones in Mandarin Chinese. According to the overall F0 patterns, these tones can be described as high-level, mid-rising, mid-falling-rising, and high-falling (Chao, 1968). They are traditionally called Tones 1, 2, 3, and 4 respectively. For example, the syllable ma with Tones 1, 2, 3, and 4 would mean 'mother', 'bear', 'horse' and 'refresh'. Although linguistic theory has stressed the feature of pitch in the description of tone, it is possible that other characteristics are used in perception. In addition to the F0, some other possible acoustic characteristics specifying differences in tones are vowel duration, amplitude, and vowel quality.

Studies of tone production (Chung, Ito, Sano, and Rimura 1971; Bowie, 1976; Tseng, 1981) have found that certain acoustic characteristics vary with changes in tonal variants for Mandarin words produced in isolation. The F0 pattern differed for the four different tones, roughly following the descriptive labels high-level, mid-rising, mid-falling-rising, and high-falling. In addition, the intrinsic duration differed for different tones. The falling-rising tone (Tone 3) was the longest in vowel duration, while the falling tone (Tone 4) was the shortest. The intrinsic amplitude also seemed to vary. The falling-rising tone (Tone 3) was produced with the lowest amplitude, while the falling tone (Tone 4) was the highest. Also, the F0 pattern for a particular tone appears to change in different vowels (Bowie, 1976). If these observations reflect reliable differences in the acoustic characteristics of Chinese tone, then listeners might use the information in the perceptual recognition of tone.

Using synthetic vowels, Tseng, Massaro, and Cohen (unpublished) manipulated two or more acoustic variables to assess their role in the perception of tones in Mandarin Chinese. The most important variable was the F0 pattern. Vowel duration was used as a cue by only two of the six subjects. Vowel quality did not influence the tone
judgments in that the F0 pattern had the same influence across a wide range of vowels (/a/, /e/, /i/, /y/). The amplitude of the vowel produced only a small influence on tone perception, and this influence was in the opposite direction from the amplitude differences observed in natural speech. Thus, the F0 pattern appears to be the important property relevant to the perception of tones in Mandarin Chinese.

Tseng et al (unpublished) observed that the F0 pattern did not necessarily correspond to a single perceptual feature. The F0 pattern has both overall frequency and contour information. To assess whether both of these properties were relevant to tone perception, the authors created a continuum of sounds between Tones 3 and 4. Seven F0 patterns were created by interpolating between falling-rising and falling tone patterns in equal steps. These patterns differed in both F0 height and F0 contour. The F0 height was varied by increasing or decreasing the F0 pattern by some absolute frequency throughout the F0 pattern. Each of the seven patterns was increased by -30, -10, -10, -10, -20, or 30 Hz, giving a total of 7 x 7 = 49 test stimuli. The results indicated strong effects of both F0 height and F0 pattern.

The current experiment extends the previous study in two ways. First, the F0 height manipulation in the Tseng et al. study was not completely independent of F0 contour, the

LEXICAL TONE PERCEPTION

F0 patterns of Tones 1 and 2 differ in F0 height initially and F0 height was varied by simply increasing or decreasing the F0 patterns by some absolute frequency throughout the F0 pattern. Thus, the relative importance of F0 height and F0 contour could not be evaluated exactly, and it was not possible to provide a quantitative test of the fuzzy logical model of speech perception. The present experiment manipulates F0 height and F0 contour independently of one another and uses the results to provide a quantitative test of the model. The second extension of the current experiment is the study of the distinction between Tones 1 and 2, called high falling and mid-rising, respectively (Chao, 1968).

Method

Stimuli—The goal of this experiment was to evaluate the contribution of F0 height and F0 contour to tone judgments. The tones used in this experiment were high-level tone and mid-rising tone. The variables manipulated were the F0 height and the F0 contour. The vowel /i/ was the carrier vowel with the first five formants set at 255, 2150, 1150, 3710, and 4500 Hz. The vowel duration was 250 msec. The F0 patterns for the level tone (Tone 1) and rising tone (Tone 2) in Mandarin Chinese were the average values across 3
Figure 1. Endpoint FO patterns used in the experiment.

The speech sounds were synthesized using the Klatt (1980) software speech synthesis program implemented on a Digital PDP-11/34A computer. FO values were specified and parameters were calculated in 5-msec increments. The output of the speech synthesizer was recorded with a sampling rate of 10 KHz and stored on files on a disc. During the experiment, the stimulus was played back at 10 KHz by a 12-bit D-to-A converter (Data Translation Module 1711). The output of the D-A converter was filtered at 20-4000 Hz by a Krohn Hite Model 3560B bandpass filter. The stimuli were then amplified (McIntosh model MC-50) and presented to subjects at approximately 65 dB SPL over Ross Pro-4AA headphones.

Procedure Each trial began with the presentation of one of the 49 vowels selected randomly without replacement
in blocks of 49 trials. The subjects were told to identify the total pattern of the stimuli, i.e., level tone or rising, traditionally called Tone 1 and Tone 2 in Mandarin Chinese. Each subject responded by pressing either the key "1" or "2" on a computer terminal keyboard (TekiVision Model 205). Subjects had 3 sec to make their response. The next trial began immediately after the response interval. There were 25 practice trials followed by 254 experimental trials in each experimental session. The subjects did not know that the first 25 trials were practice and would not be analyzed. The subjects participated in two sessions on a single day. The subjects took about a 10 min break between sessions. A total of 12 observations were collected for each subject for each of the 49 test sounds.

Subjects—Six Chinese subjects, visiting scholars at University of California Santa Cruz, participated in the experiment. All of the subjects selected were native speakers of Mandarin Chinese. Five of the subjects had participated in the Tseng et al. (unpublished) experiments. The subjects were paid $5.00 per hour for their service. Up to four subjects could be tested simultaneously in separate sound-attenuated rooms.

Results

Figure 2 gives the proportion of Tone 1 (level) identifications as a function of F0 contour going from rising to level; F0 height is the curve parameter. The proportion of level responses increased as a function of a flattening of the F0 contour, F(5, 38) = 17.3, p < .001, and with an increase in F0 height, F(6, 10) = 11.4, p < .001. In addition, the effect of each variable was largest at the more ambiguous levels of the other variables, F(36, 150) = 2.41, p < .001. These results indicate that the high-level and mid-rising tones are perceived on the basis of both F0 height and contour. However, two of the six subjects showed no systematic effect of either F0 contour or F0 height. It would have been interesting to test what, if any, cues these two subjects would have used to distinguish the high-level and mid-rising tones. These two subjects had utilized both F0 contour and F0 height in distinguishing between high-falling and falling-rising tones (Tones 3 and 4) in Experiment 5 of Tseng et al., however, and it could have simply been the case that these subjects became bored with the series of experiments and responded randomly.

Theoretical Description

The fuzzy logical model was applied to the individual results of each of the six subjects. Both F0 height and F0 contour are assumed to be evaluated and integrated together for perceptual recognition. This model specifies how F0 pattern and duration cues are evaluated and integrated together to arrive at the perceptual recognition.
Figure 2: proportion of level identifications as a function of F0 contour; onset frequency in Hz is the curve parameter.

LEXICAL TONE PERCEPTION

of the lexical tone. The model does not predict what cues will be used in speech perception; it only predicts how multiple cues are evaluated and integrated.

In the model, the lexical tones are stored as prototype descriptions in memory and might contain information about both F0 contour and height.

Tone 1: (level F0) and (high F0) = L & H
Tone 2: (rising F0) and (mid F0) = R & M

where "and" and "&" represent logical conjunction.

For the discrimination of these two tones, it is assumed that the value of a feature for Tone 2 is just one minus the value for Tone 1. The additional assumption of a multiplicative combination of features defines the prototypes as

Tone 1: L1 x H1
Tone 2: (1 - Li) x (1 - Hi)

Given a speech event, truth values are determined for each feature indicating the degree to which the feature has the property in question. The truth values vary between zero and one to span the range from completely false to completely true. For example, the F0 contour may be assigned the value
At the pattern classification stage, the listener computes the relative goodness of match of the stimulus to each of the two prototypes. The likelihood of a high-level tone (Tone 1) classification is, therefore, equal to the goodness of match of the stimulus with Tone 2 relative to the sum of the goodness of match values for Tone 1 and Tone 2.

\[ P(\text{Tone 1}) = \frac{L_1 x H_1}{(L_1 x H_1) + (1 - L_1) x (1 - H_1)} \]

In order to fit the fuzzy logical model to the results, fourteen parameters must be estimated to describe the 49 data points. Seven parameter values are needed for the seven levels of FO contour and seven parameters for FO height. Parameters must be estimated from the obtained results. The parameters are necessary because the fuzzy-logical model does not predict the exact relationship between some stimulus property and its feature value. Thus, the experiment tests how the various sources of information are combined rather than how the feature values are determined in relation to the physical stimulus. We allow one parameter value for each hypothetical output value of each feature in our analysis of theoretical models. The parameter values can be used to determine the relative contribution of each source and to ascertain the psychophysical relationship between the stimulus source and the perceptual consequence.

To determine the quantitative predictions of the model, the observed proportions of a "high-level" response for each subject for each of the 49 experimental conditions were used to estimate the parameters with the program STERPIT (Chandler, 1968). A model is represented to this program as a set of prediction equations and a set of unknown parameters. Initially, all parameters are set to .5. By iteratively adjusting the parameters of the model, STERPIT minimizes the squared deviations between the 49 observed and predicted points. Thus, what STERPIT does is to find a set of parameter values which, when put in the model, come closest to predicting the observed data. Figure 3 shows the proportion of level judgments for the four subjects as a function of the FO contour; FO height is the curve parameter. Table 1 gives the parameter estimates which change in a meaningful way for the four subjects.

Considering the small number of observations, the fuzzy logical model provided a good description of the observed results, with an average NSE of .086 for the four subjects.

It should be noted that the model was not applied to the results of the other two subjects, simply because their results do not provide an informative test of the model. The model predicts the integration of acoustic cues in speech perception and can only be tested if two or more cues are
Figure 3. Predicted and observed proportion of level identifications for 4 subjects as a function of the FO contour; onset frequency in Hz is the curve parameter.

Table 1. Parameter values for the fit of the fuzzy logical model for four subjects

<table>
<thead>
<tr>
<th>Stimulus Level</th>
<th>FO Height</th>
<th>FO Contour</th>
<th>FO Height</th>
<th>FO Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.08</td>
<td>.12</td>
<td>.14</td>
<td>.21</td>
</tr>
<tr>
<td>2</td>
<td>.17</td>
<td>.19</td>
<td>.17</td>
<td>.36</td>
</tr>
<tr>
<td>3</td>
<td>.38</td>
<td>.31</td>
<td>.25</td>
<td>.33</td>
</tr>
<tr>
<td>4</td>
<td>.76</td>
<td>.85</td>
<td>.67</td>
<td>.69</td>
</tr>
<tr>
<td>5</td>
<td>.22</td>
<td>.25</td>
<td>.84</td>
<td>.83</td>
</tr>
<tr>
<td>6</td>
<td>.97</td>
<td>.89</td>
<td>.90</td>
<td>.84</td>
</tr>
<tr>
<td>7</td>
<td>.98</td>
<td>.99</td>
<td>.96</td>
<td>.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stimulus Level</th>
<th>FO Height</th>
<th>FO Contour</th>
<th>FO Height</th>
<th>FO Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.13</td>
<td>.05</td>
<td>.09</td>
<td>.11</td>
</tr>
<tr>
<td>2</td>
<td>.08</td>
<td>.07</td>
<td>.07</td>
<td>.11</td>
</tr>
<tr>
<td>3</td>
<td>.29</td>
<td>.12</td>
<td>.22</td>
<td>.31</td>
</tr>
<tr>
<td>4</td>
<td>.63</td>
<td>.47</td>
<td>.83</td>
<td>.73</td>
</tr>
<tr>
<td>5</td>
<td>.86</td>
<td>.70</td>
<td>.80</td>
<td>.86</td>
</tr>
<tr>
<td>6</td>
<td>.97</td>
<td>.99</td>
<td>.96</td>
<td>.87</td>
</tr>
<tr>
<td>7</td>
<td>.98</td>
<td>.99</td>
<td>.98</td>
<td>.97</td>
</tr>
</tbody>
</table>

Note: The levels 1 to 7 go from mid to high for FO Height and from rising to level for FO Contour.
used by subjects in a given experiment. Since these two
two subjects did not use the cues, no test is available
concerning how the cues are combined. The model is also
silent with respect to the individual differences that were
observed. The question of why these two subjects did not use
the cues is not answered.

Discussion

The results of this experiment demonstrated that a
significant effect of both the F0-contour and F0 height on
the perception of high-level tone and mid-rising tones.
Recent findings support this interpretation. After the
completion of this research, we learned that Chang and
Sherwood (1987) found similar results in the machine
recognition of Chinese tones. Accurate recognition required
information about both F0 height and F0 contour; one
dimension alone was not sufficient for accurate recognition
performance. Hence, it appears that both overall F0
frequency and contour are important for the perception of
all four lexical tones in Mandarin. It has been well
documented that various tone languages, for example, Thai,
Cantonese, Yoruba, etc., possess register tones, that is,
tones that are distinguished on the basis of F0 frequency
(for pitch height). As an example, Abraham (1978) found that
the height of flat F0 patterns cued the distinction among
the low, mid and high tones in Thai, but not between the

falling and rising tones. What is of special interest is
that both the F0 contour and F0 frequency are utilized by
the listeners as cues to perceive tones in Mandarin which
does not possess true register tones. The interaction
between these two effective cues is such that the
contribution of one cue is increased when the other cue is
more ambiguous.

ACKNOWLEDGMENT

This research was supported by a grant from the
National Institute of Mental Health (MH 38324) to Dominic W.
Massaro. Requests for reprints should be sent to Dominic W.
Massaro, Program in Experiential Psychology, University of
California, Santa Cruz, California 95064.
LEXICAL TONE PERCEPTION

Society of America, 1976, 68, 704-717.

Nasaworo, G. N. and Odem, G. C. Speech perception: a
framework for research and theory, in W. A. Lemb (Ed.), Speech... and Language: advances... in basic research... and practice, Vol. 3. New York: Academic

McCawley, J. C. What is a tone language? in Fromkin, V. A.,
Tone: A Linguistic Survey. New York: Academic Press,
1978, 112-121.

Odem, G. C. and Nasaworo, U. M. Integration of semantic
information in speech perception. Psychological

Fision, V. A. Speech Perception. In W. A. Lemb (Ed.)
Handbook of Learning and Cognitive Processes, William

Hepp, E. H. Perceptual integration and differentiation of
spatial cues for nonpitches stop consonants. Reception and

Studdert-Kennedy, M., Liberman, A. M., Ranks, K. S., and
Cooper, F. S. Motor theory of speech perception: A

Tseng, C. An acoustic phonetic study on tones in Mandarin

Tseng, C., Nasaworo, K. W., and Cohen, W. N. Lexical tone
perception in Mandarin Chinese: evaluation and integration of acoustic features, unpublished
manuscript. 1983. Available from A. W. Nasaworo,