Qualitative Linguistics, 1983,
19, 76-102.

VOWEL AND LEXICAL TONE PERCEPTION IN MANDARIN CHINESE: PSYCHOLINGUISTIC AND PSYCHOACOUSTIC CONTRIBUTIONS

Dominic W. Massaro, Chiu-yu Tseng, Michael M. Cohen

Abstract

Three experiments were carried out to assess the evaluation and integration of acoustic information in the perception of vowels and lexical tone in Mandarin Chinese. The experiments were carried out within the framework of a fuzzy logical model of speech perception (ODEN, MASSARO 1978). Using synthetic speech stimuli, it was possible to determine psychoacoustic and psycholinguistic influences by comparing the results of Chinese listeners with those of English listeners. Although formant values were the primary cues in the perception of vowel quality, F_0 pattern and amplitude also influenced vowel judgments along a speech continuum between [i] and [y]. The influence of F_0 pattern appears to be the result of some psychoacoustic variable, whereas the amplitude effect requires a psycholinguistic knowledge of the structure of the relevant language contrasts. In contrast to the influence of F_0 pattern on vowel judgments, vowel quality appeared to have no systematic effect on the identification of lexical tone. These results are consistent with the fuzzy logical model of speech perception, which views perceptual recognition in terms of the integration of a variety of independent cues in the speech signal.

1. *Introduction*

From a pattern recognition perspective, speech perception is an amazing skill. There does not seem to be an exact relationship between the acoustic signal and the perceived patterns in the message. As an example, the units of recognition do not seem to coincide with units of the speech signal. In many cases, we hear word boundaries where there is no or little silence and hear complete words with significant silent periods. In the statement "That you may see," there is usually more silence during the /æt/ portion of "that" than between the words "may" and "see." This example shows that successive segments of sound can produce a coherent or unitary percept and that discrete percepts can result from a relatively continuous sound segment.

The acoustic properties or features characterizing a speech segment seem to vary depending on the placement of that segment in the speech message. As an example, certain properties of stop consonants depend on their position in words. Voice onset time (VOT) is an important property for stops in initial and medial position, whereas this property is less common in word-final position. Accordingly, VOT appears to be an important feature for the identification of stops in initial and medial position, whereas preceding vowel duration is an important feature for voicing of stops in word-final position. Thus, the listener's knowledge characterizing stop consonants must allow for position-dependent features in the identification process.

Another obstacle in the recognition of small speech segments is that the acoustic signal specifying a particular linguistic unit is context sensitive. That is, the acoustic properties of a unit found in one context are significantly modified in another. Consider the classic example of the syllables /di/ and /du/. The acoustic signal corresponding to the initial /d/ sound is significantly different in the two syllables. The following vowel context modifies the properties of the preceding stop. This example shows that perceptual recognition of some speech sounds must take into account the contribution of the surrounding context on the features used in identification.

One reason that speech perception is accomplished even in the presence of the aforementioned difficulties is the contribution of linguistic context. It is generally agreed that the listener normally achieves good recognition by supplementing the information from the acoustic signal with contextual information generated through the utilization of knowledge in long-term memory.

There is considerable debate concerning how informative the acoustic signal actually is (BLUMSTEIN,STEVENS 1979; COLE,SCOTT 1974; LIBERMAN,COOPER,SHANKWEILER,STUDDERT-KENNEDY 1967; MASSARO 1975b). Even if the acoustic signal proved to be sufficient for speech recognition under ideal conditions, however, few researchers would believe that the listener relies on only the acoustic signal.

Our study of pattern recognition in speech has been carried out within a general information-processing model (MASSARO 1975a, 1975b, 1979). A schematic representation of the stages of processing in the model is presented in Figure 1. At each stage of processing, memory

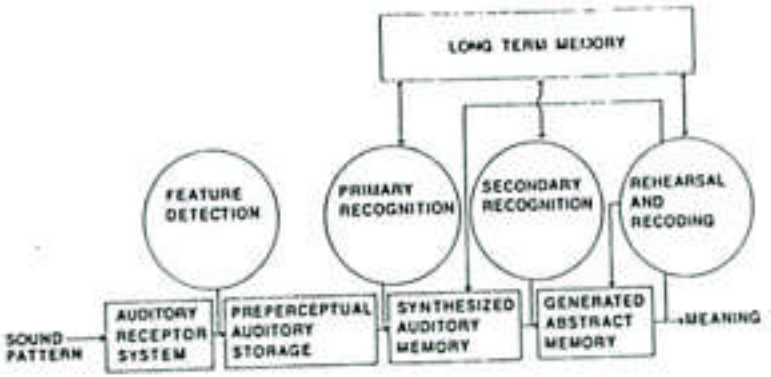

Figure 1. Schematic diagram of the general auditory information processing model.

and process components are represented. A particular memory component (indicated by a rectangle) corresponds to the information available at that stage, whereas the corresponding process component (indicated by a circle) represents the operations applied to the information in the memory component.

The feature detection process transforms the energy pattern created by the language stimulus and transduced by the appropriate receptor system into a set of features held in preperceptual storage. In speech, for example, the changes in sound pressure set the eardrums in motion and these mechanical vibrations are transduced into a set of neural impulses. It is assumed that the signal in the form of continuous changes in vibration pattern is transformed into a set of relatively independent features. Features are not limited to primitive attributes, but can be relatively complex. In speech, for example, the amount of energy in a particular frequency band would be a simple feature, whereas a complex feature might include information about the direction and rate of frequency change of the sound. It would be

possible, for example, to have a feature detector that responds to the rising first formant transition that is characteristic of the class of voiced stop consonants. Primary recognition evaluates and integrates these features into a percept which is held in synthesized memory.

Secondary recognition transforms synthesized percepts into meaningful forms in generated abstract memory. In speech perception it is assumed that the input is analyzed syllable by syllable for meaning. The secondary recognition process makes the transformation from percept to meaning by finding the best match between the perceptual information and the lexicon in long-term memory. Each word in the lexicon contains perceptual and conceptual codes. The concept recognized is a function of at least two independent sources of information: the perceptual information in synthesized memory and the semantic/syntactic context in the message.

In the present model, the same abstract structure stores the meaning of both listening and reading. Generated abstract memory (GAM) in the model corresponds to the working memory of contemporary information-processing theory. Rehearsal and recoding processes operate at this stage to maintain and build semantic/syntactic structures. There is good evidence that this memory has a limited capacity, holding about five, plus or minus two, chunks of information. For a more detailed discussion of processing at this stage, see MACKAY (1975a, CHAPTER 27).

The following example may help clarify the differences among these three levels of processing: presented with a tone, the listener can detect or sense the presence of sound, hear and remember a tone of a particular quality, and even identify it as a particular note on the musical scale (cf. Figure 2). Figure 2 illustrates the outcome of these three stages as detection, perception, and conception, respectively. Each of these stages makes information available to a response selection and programming process. Accordingly, some response execution can be initiated by any of the three levels of language processing. Although the boundaries between these stages are sometimes fuzzy, we have found it helpful to maintain these distinctions in our experimental and theoretical research.

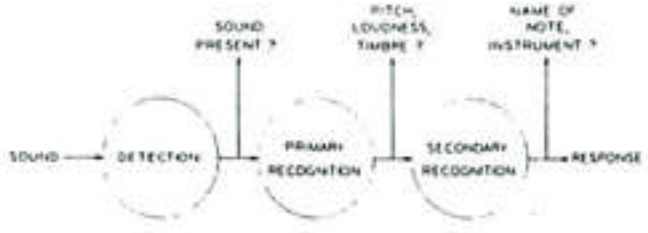

Figure 2. Three stages of processing a sound stimulus in music.

The information-processing model has served as the basis for a fuzzy logical model of pattern recognition in speech perception. Three operations are central in the model of pattern recognition: feature evaluation, prototype matching, and pattern classification. Central to the model is the ideal of prototype descriptions stored in the listener's long-term memory. These descriptions are propositions specifying the featural properties of speech sounds of roughly syllabic size (ODEN,MASSARO 1978).

In order to illustrate how the fuzzy logical model is applied and tested within the domain of pattern recognition, consider an experiment carried out by MASSARO and ODEN (1980). Seven levels of voice onset time (VOT) were crossed with seven levels of the onsets of the F_2-F_3 transitions in the synthesis of stop consonant-vowel syllables. The VOTs ranged from a completely voiced to a completely voiceless sound. The values were 10, 15, 20, 25, 30, 35, and 40 msec. The seven levels of the F_2-F_3 onset frequencies ranged from 1345 to 1796 Hz for F_2 and 2397 to 3200 Hz for F_3 to give a continuum of sounds going from a labial to an alveolar place of articulation. Subjects made repeated identifications of random presentations of the 49 unique syllables from the alternatives /bae/, /dae/, /pae/, and /tae/.

The four panels of Figure 3 present the percentage of /bae/, /pae/, /dae/, and /tae/ identifications, respectively, as a function of the two independent variables. The levels along the abscissa are not equally spaced, but rather have been adjusted to be proportional to the differences between the marginal means across the levels of the F_2-F_3 transitions. These differences were computed separately for each of the four response alternatives and then averaged over response types so that all four of the panels have the same spacing along the abscissa.

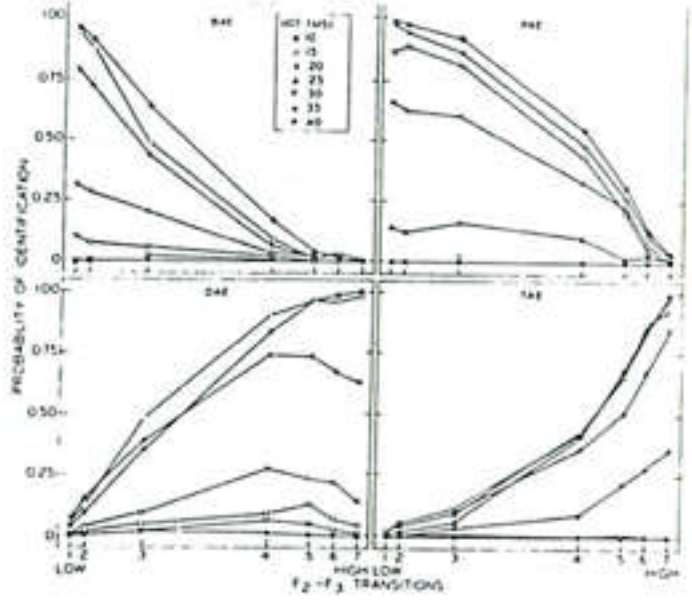

Figure 3. Percentage of /bae/, /pae/, /dae/, and /tae/ identifications as a function of VOT and F_2-F_3 transitions. Note that the spacing along the abscissa is roughly proportional to the spacing of the marginal means across the levels of the F_2-F_3 transitions.

In our fuzzy logical model of these results, we assume the following propositions specifying the prototype descriptions for the four response alternatives in the experiment.

/bae/: (short VOT) and (low F_2-F_3 onsets) (1)

/pae/: (long VOT) and (low F_2-F_3 onsets) (2)

/dae/: (short VOT) and (high F_2-F_3 onsets) (3)

/tae/: (long VOT) and (high F_2-F_3 onsets) (4)

The properties for the prototypes would also include other acoustic features characterizing stop consonants and vowel /ae/. These properties are not included in the propositions, since they are present in all of the four alternatives. As we shall see later, the mathematical form of the fuzzy logical model necessitates consideration of only those properties which differ between prototypes. These propositions specify the ideal values of each of the acoustic features for the particular speech sound.

Upon presentation of a speech sound, the feature detection process evaluates each acoustic feature and produces a fuzzy truth value specifying the degree to which it is true that the sound has the relevant acoustic feature. For example,

$$t[\text{short VOT } (S_{ij})] = .60 \tag{5}$$

represents that it is .60 true that the speech sound S_{ij}, from the ith row and jth column of the factorial stimulus design, has a short VOT. To simplify the notation, let SV_i and LV_i correspond to short and long VOTs, respectively. The subscript i signifies that the values change only with changes in the row variable VOT. Similarly, LO_j and HO_j correspond to low and high F_2-F_3 onsets, respectively. The values change only with changes in the column j variable of the F_2-F_3 onset frequencies. Also, we will henceforth use the expression "short VOT" to represent its truth value t(short VOT). The truth of the negation of a feature can be defined as one minus the truth value of the feature. In this case, if .6 specifies the truth value of a short VOT, then 1 - .6 = .4 would specify the truth value of a long VOT. In general, the value $LV_i = 1 - SV_i$. A similar complementary relationship could be assumed between high and low F_2-F_3 onsets.

$$\text{low } F_2\text{-}F_3 \text{ onsets} = \text{NOT(high } F_2\text{-}F_3 \text{ onsets)} \tag{6}$$

Following this logic, the value LO_j would be equal to $1 - HO_j$. In many cases, the use of complementary features in the prototypes produces models not identifiably different from the apparently more complex model given by Equations 1-4 (MASSARO,COHEN unpublished). Incorporating the assumption of complementary features in the prototype definitions gives Equations 7-10 in place of Equations 1-4.

/bae/: (short VOT) and NOT(high F_2-F_3 onsets) (7)

/pae/: NOT(short VOT) and NOT(high F_2-F_3 onsets) (8)

/dae/: (short VOT) and (high F_2-F_3 onsets) (9)

/tae/: NOT(short VOT) and (high F_2-F_3 onsets) (10)

At the prototype matching stage, a determination is made regarding the degree to which the conjunction of features in each prototype definition has been realized in the speech signal. On the basis of much empirical work, the use of the multiplication operator for conjunction has been found to give the best fit to the data in speech perception (MASSARO,COHEN 1976, 1977). The multiplicative rule gives the matching function

$$bae(S_{ij}) = SV_i \times LO_j. \tag{11}$$

Given the matching functions for each of the alternative prototypes, the speech sound is identified on the basis of the relative degree of match. Following the rationale of LUCE's (1959) choice model, it is assumed that the probability of identifying a stimulus to be a particular syllable is equal to the relative degree to which that syllable matches the stimulus compared to the degree of match of the other syllables under consideration. In our example, the person must identify the speech sound as either /bae/, /pae/, /dae/, or /tae/.

The probability of a bae identification will, therefore, be given by

$$P(\text{bae}|S_{ij}) = \frac{bae(S_{ij})}{bae(S_{ij}) + pae(S_{ij}) + dae(S_{ij}) + tae(S_{ij})} \tag{12}$$

where the variables in the ratio represent the matching functions for the four alternative speech sounds.

The fuzzy logical model has been developed and utilized to provide a framework for research in the identification of speech sounds varying on two or more dimensions. In this paper, the model will be extended to include other theoretical and empirical issues in speech perception research. The goals of the present experiments were threefold. First, it is important to test the fuzzy logical model in the domain of another language besides English. Second, we evaluated the perception and discrimination of tone and vowel in Chinese. Third, we were able to evaluate the role of a language user's knowledge in the perception of distinctions in their language by comparing their performance to that of American subjects unfamiliar with tone languages.

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 and Asia as well as the marginal tone (or "pitch accent") languages of Europe; 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; TSENG,MASSARO,COHEN unpublished).

In Mandarin Chinese, there are four lexical tones. These tones can be described according to the overall fundamental frequency patterns as high-level, mid-rising, mid-falling-rising and high-falling (CHAO 1978). These tones are traditionally called Tones 1, 2, 3, and 4 respectively by Chinese speakers. As an example of the role of lexical tone use, the syllable ma with Tones 1, 2, 3, and 4 would mean /mother/, /hemp/, /horse/, and /reproach/. Although linguistic theory has stressed the feature of pitch in the description tone, it is possible that other acoustic characteristics are used in perception. Besides the onset frequency and direction of movement, other possible acoustic characteristics specifying tones are vowel duration, amplitude, and vowel quality.

2. *Experiment* 1

One goal of Experiment 1 was to investigate whether tone distinctions and vowel quality influences are based on different acoustic features. That is, do different acoustic features cue

these different phonetic distinctions or is a given acoustic feature relevant to both phonetic differences? Specifically, we evaluated whether fundamental frequency (F_0) pattern influences the perception of vowel quality in addition to its influence on lexical tone. Analogously, we asked whether the formant frequencies influence the perception of lexical tone in addition to their influence on vowel quality. The experiment also provides evidence concerning how these two acoustic characteristics are integrated to achieve perceptual recognition of both vowel and lexical tone in Mandarin Chinese.

A tone continuum was created between falling-rising and high-falling tone by varying the F_0 pattern. A vowel continuum was created between [i] and [y] by varying the formant frequencies. These two continua were combined in a factorial arrangement so that every level of F_0 pattern was paired with every level of the formant frequencies. Given seven levels along each continuum, a total set of 49 speech sounds was employed in the identification task. Each of the four combinations of tone (falling-rising or high-falling) and vowel quality ([i] or [y]) made a word, and Chinese listeners were given the four words as response alternatives. Four keys on a computer terminal keyboard were labeled with the appropriate Chinese characters. On each trial, the subjects identified the test sound as one of the four alternatives.

Method

Subjects

Six Chinese subjects, four males and two females from University of California, Santa Cruz, participated in the experiment. All of the subjects selected were native speakers of Mandarin Chinese who had resided in the Peking area for most of their lives prior to their recent arrival at the U.S. Although their native and most dominant language is Mandarin, four subjects had experience with Chinese dialects other than Mandarin; S2 could speak and comprehend Shanghai dialect, S3 could speak and comprehend Tientsin dialect, and comprehend Shanghai dialect. S5 could speak and comprehend Shanxi dialect, and S6 could comprehend Honan and Shanghai dialects. The subjects also had studied foreign languages. S1 studied English and Japanese; S2 English and Russian; S3 English, Russian, and French; S4 English and Russian; S5 Russian and almost no English; and S6 English, Russian,

and Japanese. Prior to their recent trip to the U.S., these subjects had limited or no contact with foreigners. Subjects were paid at $5.00 per hour for their service. The subjects had participated in a previous study of lexical tone identification.

Speech sounds

The two variables manipulated were the F_0 pattern and vowel quality. The F_0 pattern ranged between falling-rising tone (Tone 3) and falling tone (Tone 4) in Mandarin Chinese. The continuum was divided in six equal steps to generate a set of seven patterns, ranging from falling-rising tone to high-falling tone. (See Table 1 for the endpoint values of the F_0 pattern continuum.) The values of F_0 pattern were specified at 11 points during the 350 msec vowel

TABLE 1

The values of F_0 frequency for the endpoint stimuli at eleven points during the vowel presentation

Percentage of Vowel Duration	falling-rising (Tone 3)	falling (Tone 4)
Onset	120	150
10	115	145
20	111	140
30	106	135
40	101	129
50	97	124
60	100	120
70	103	117
80	106	113
90	109	110
End	112	106

presentation. For vowel quality, a continuum of seven vowels from [i] to [y] was generated by varying the formant frequencies. Both vowels are front high vowels in Mandarin Chinese and differ in manner of articulation only, i.e., [y] is a rounded vowel. The formant frequencies were initially based on the values given by BOMIE (1978) and were later revised in order to make the continuum of vowel quality relatively fuzzy and the boundary between the vowel categories at the center of the continuum. The seven stimuli along the vowel continuum were equally spaced in terms of their formant frequencies. The values of F_1, F_2, F_3, and F_4 for vowel [y] were 262, 1973, 2542, and 3443 Hz respectively. The formant frequency values for vowel [i] were 252, 2082, 2864, and 3586. The values for F_5 remained constant at 4500 Hz for both vowels.

The 49 vowels representing the factorial combination of F_0 pattern and formant frequencies were synthesized using the KLATT (1980) software program implemented on a Digital PDP-11/34A computer. Each stimulus was specified as a series of lists of parameter vectors. Time values for the synthesis were specified and parameters were changed in 5-msec increments. The amplitude was set at 0 dB at the onset of the vowel, then moved linearly to 55 dB at 30 msec after the onset. The amplitude then decreased linearly to 50 dB at 30 msec before the offset, and to 0 dB at the offset. The amplitude values given here are the Klatt software synthesizer parameter values. The output of the speech synthesizer was recorded at a sampling rate of 10 KHz and stored in files on disc. During the experiment, stimuli were played back at 10 KHz by a 12 bit D to A converter (Data Translation Model 1711). The output of D-A converter was filtered within 20-4800 Hz by a Krohn-Hite Model 3500R bandpass filter. Stimuli were then amplified (McIntosh Model MC-30) and presented to each subject at roughly 55 dB SPL A over Koss Pro-4AA headphones. The subjects sat in individual sound-deadened rooms.

Procedure

Each trial began with the presentation of one of the 49 vowels selected randomly without replacement within a block of 49 trials. The subjects were given the four response choices [i] and [y] in both the falling-rising tone (Tone 3) and falling tone (Tone 4). Four keys on the computer terminal keyboard (v, b, n, m) were labeled with

Chinese characters corresponding to vowel [i] in falling-rising tone (Tone 3), vowel [i] in falling tone (Tone 4), vowel [y] in falling-rising tone (Tone 3) and vowel [y] in falling tone (Tone 4) respectively. Subjects had three seconds to make their response. The next trial began immediately after this response interval. There were 25 practice trials followed by 294 experimental trials in each experimental session. The subjects did not know that the first 25 trials were practice and would not be analyzed. Before the first session of the first day, a practice session of 150 trials was also administered. The Chinese subjects participated in two experimental sessions on four different days, giving a total of 48 observations for each speech sound.

Results and discussion

An analysis of variance was performed on the percentage of identification with subjects, vowel quality, F_0 pattern, and responses as factors. There were significant effects of each of the independent variables along with significant interactions. Figure 4 presents the observed proportion of each of the four judgments as a function of vowel quality and F_0 pattern. The two independent variables influenced performance in the expected direction. The identification of vowel quality was primarily determined by the formant frequencies, $F(6,30)=145$, $p < .001$, and the identification of lexical tone was primarily determined by the F_0 pattern, $F(6,30)=88$, $p < .001$.

The fuzzy logical model can be applied to the perception of vowel quality and lexical tone. It is assumed that prototypes corresponding to the four alternatives are stored in memory. The prototypes specify the ideal acoustic features for each alternative. For the simple representation, the prototypes for the four alternatives would be defined as

[i]-FR: high vowel formants and falling-rising F_0 (13)

[i]-HF: high vowel formants and high-falling F_0 (14)

[y]-FR: low vowel formants and falling-rising F_0 (15)

[y]-HF: low vowel formants and high-falling F_0. (16)

As applied to the present task, the four alternatives are distinguished on the basis of the acoustic features representing the formant

frequencies and the F_0 pattern. Following the logic of the model, it can be assumed that the two values of each feature are additive complements of one another. That is, having low vowel formants can be specified as having not [high vowel formants]. Therefore, the prototypes can be represented as

[i]-FR: HVF $\wedge$ FR (17)

[i]-HF: HVF $\wedge$ (1-FR) (18)

[y]-FR: (1-HVF) $\wedge$ FR (19)

[y]-HF: (1-HVF) $\wedge$ (1-FR) (20)

where HVF and FR correspond to high vowel formants and falling-rising F_0 pattern, respectively, and $\wedge$ represents the conjunction and.

The two features are assumed to be combined multiplicatively at the prototype matching stage. Pattern classification is based on the goodness of each alternative relative to the sum of goodnesses for the four alternatives. In this case, the probability of a [i]-FR response, P([i]-FR) is equal to

$$P([i]\text{-}FR)=\frac{HVF \times FR}{(HVF \times FR)+(HVF \times (1-FR))+((1-HVF) \times FR)+((1-HVF)\times(1-FR))} = HVF \times FR \quad (21)$$

since the denominator sums to one.

To fit the model to the results, it is necessary to estimate a unique parameter for each of the seven levels of the vowel formant frequencies and a unique parameter for each of the seven levels of the F_0 pattern. Therefore, 14 parameters are necessary to predict the responses. Since a response is required on each trial, the four response probabilities sum to one for each stimulus, giving (4-1) independent observations for each of the 49 stimuli. Thus, there are $49 \times 3=147$ independent observations being predicted by 14 parameters.

This model was fit to the results of each of the six subjects. The average predictions are shown in Figure 4 along with the average observed results. The model did a fairly good job of describing the results, considering that 147 independent observations are being predicted with only 14 free parameters. The root mean squared deviation (rmsd) between the predicted and observed results ranged between .030 and .072 across the six subjects and averaged .044.

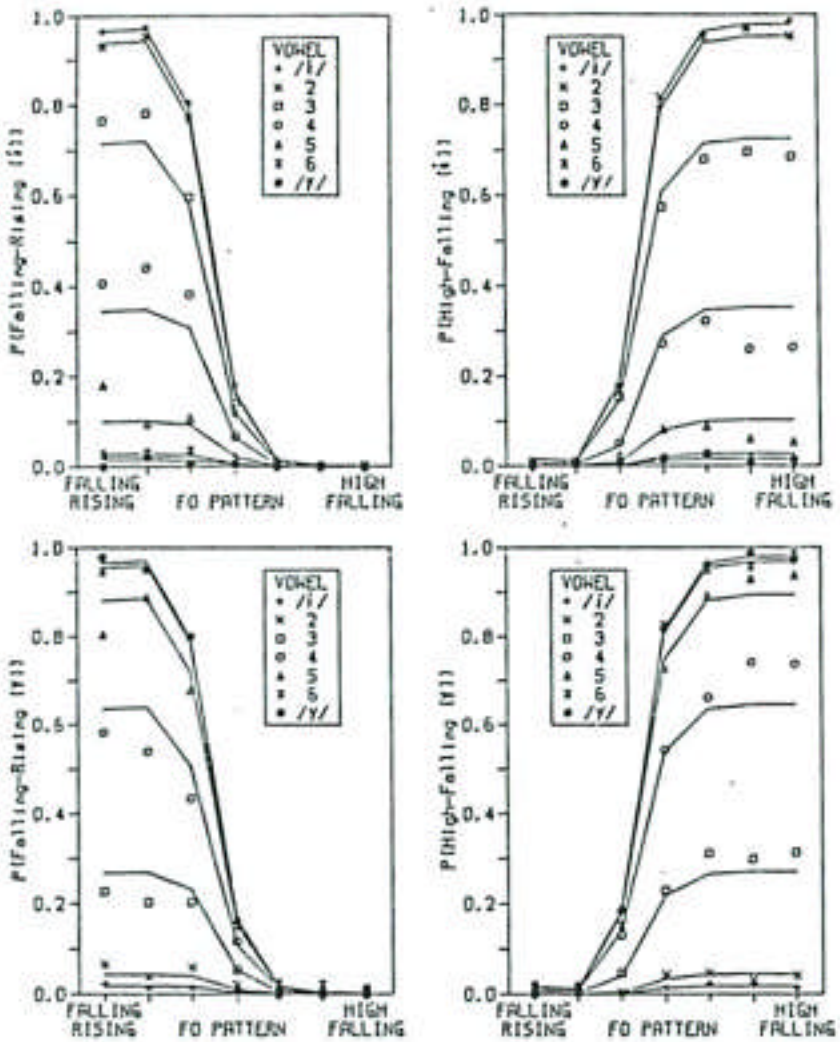

Figure 4. Percentage of identifications as a function of F_0 pattern and vowel quality (Experiment 1).

The parameter values changed systematically across the levels of the independent variables. The parameter for HVF was .982, .955, .720, .353, .103, .028, and .017 for the seven levels of the formant frequencies going from [i] to [y]. The parameter for FR was .983, .988, .816, .174, .019, .004, and .003 for the seven levels of the F_0 pattern going from falling-rising to high-falling.

Although the simple model does a fairly good job describing the results, there are some systematic differences between the predicted and observed results. These differences are more easily seen in separate analyses of the vowel quality and tone judgments. Figure 5 plots the proportion of [y] identification in the left panel and the proportion of high-falling tone (Tone 4) identification in the right panel. The [y] identifications are plotted as a function of vowel quality with F_0 pattern as the curve parameter. The high-falling identifications are plotted as a function of F_0 pattern with vowel quality as the curve parameter. As can be seen in the figure the

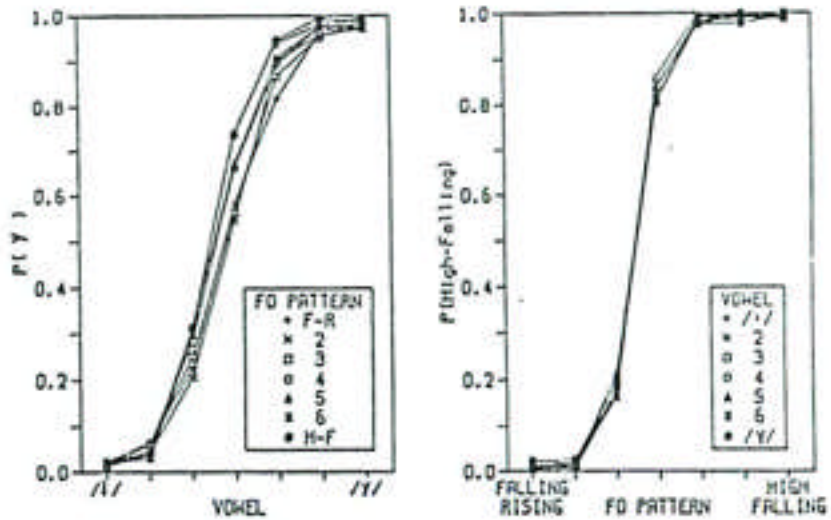

Figure 5. Left panel: percentage of y identification as a function of vowel quality; F_0 pattern is the curve parameter. Right panel: percentage of high-falling identifications as a function of the F_0 pattern; vowel quality is the curve parameter (Experiment 1).

vowel judgments are mainly dependent on the vowel quality, $F(6,30) = 145$, $p < .001$. However, the vowel was more likely to be heard as [y] as the F_0 pattern changed from falling-rising to high-falling, $F(6,30) = 3.27$, $p < .001$, especially at the more ambiguous levels of vowel quality, $F(36,180) = 2.76$, $p < .001$. The tone judgments changed systematically with F_0 pattern, $F(6,30) = 88$, $p < .001$, and were not influenced by vowel quality.

The results for the Chinese listeners indicated that the vowels were heard as more [i]-like for the falling-rising than for the high-falling tone. This result is called a boundary shift in that the boundary between [i] and [y] shifts with changes in F_0 pattern. To recognize [i], these listeners are satisfied with a less good [i] for a falling-rising than a high-falling tone. Subjects tend to hear an ambiguous level of formant structure as [i] to the extent the vowel has a falling-rising tone. For the fourth level of vowel quality, the most extreme falling-rising tone was identified as [i] about 15% more often than the most extreme high-falling tone.

3. *Experiment 2*

One important result in Experiment 1 was the influence of F_0 pattern on the vowel judgments. There are two major classes of explanation for this result. These explanations might be called psycholinguistic and psychoacoustic accounts of the results. According to the psycholinguistic account, Chinese listeners are utilizing some knowledge of the normal relationship between F_0 pattern and vowel quality in spoken Mandarin. According to the psychoacoustic account, the F_0 pattern modifies the auditory processing of the vowel formants. For the first account, it is the listener's experience with Mandarin Chinese that is critical, whereas the second account is independent of the language experience of the listener.

Experiment 2 provides a direct test of whether the boundary shift resulting from the influence of tone on the identification of vowel quality for the Chinese subjects is due to psycholinguistic or to psychoacoustic factors. In Chinese, F_0 pattern and vowel quality may be systematically related in some way and Chinese listeners may

utilize this relationship in their perceptual recognition of the vowel quality of the speech sounds. As an example, Chinese speakers might produce a more extreme [i] for a high-falling tone than for a falling-rising tone. Therefore, Chinese listeners might expect a more extreme [i] for a high-falling than for a falling-rising tone. The expectation would lead to fewer [i] identifications for a relatively ambiguous set of formant frequencies when the tone is high-falling as contrasted to falling-rising.

The second possible explanation of the boundary shift is that some psychoacoustic factor is responsible. For some reason, the vowel formants might be heard as more like [i] for the falling-rising than for the high-falling tone. That is, the influence of F_0 pattern shown in the left panel of Figure 2 is simply the indirect result of F_0 pattern modifying the acoustic featural evaluation of the vowel formants. The frequencies of F_2, F_3, and F_4 are higher for [i] than for [y]. In terms of auditory processing, going from a high-falling to a falling-rising F_0 pattern might be equivalent to raising these formant frequencies.

The present test between the psycholinguistic and psychoacoustic explanations involves a comparison between Chinese and American subjects. Since these English speakers are naive about any effect of tone on vowel quality, tone should have no psycholinguistic influence in the perception of vowel quality. The Americans should be influenced by any psychoacoustic variable, however, to the same degree as Chinese listeners.

For the American subjects, tone was unfamiliar and would probably have been difficult to discriminate. Tone is a very difficult distinction when foreign language learners are faced with their first tone language. Therefore, the American subjects could not be expected to classify the speech sounds on the basis of tone. Vowel quality differences between [i] and [y] are easy to hear, however, even for subjects who have not experienced [y] in any language. Thus, the American subjects were asked to identify the speech sounds as [i] or [y] by hitting keys labeled as EE or YU. The same set of 49 speech sounds was presented in the same manner as for the Chinese listeners.

Subjects

Eight American subjects, four males and four females, from University of California, Santa Cruz, participated in the experiment. All of the American subjects were native speakers of American English, and were undergraduate students majoring in psychology. Although all of these subjects had studied foreign languages, none of them had studied any foreign language for more than two years. Subjects 1, 2, 3, and 8 had studied Spanish, S4 Polish and Spanish, S5 French, S6 Hebrew and German, and S7 German. These subjects either fulfilled a psychology course exercise or received $5.00 per hour for service.

Stimuli and procedure

The same test stimuli and procedures of Experiment 1 were used for the American subjects. The only difference was that unlike the Chinese subjects who were given four response choices of both the vowels and tones, these American subjects were given the two response choices of the vowels [i] and [y]. Two keys on the computer terminal keyboard (Z,I) were labeled EE and YU respectively.

The American subjects participated in two sessions on each of two days, giving a total of 24 observations for each speech sound.

Results

To provide a direct comparison with the six Chinese subjects, the performance of six of the eight American subjects was analyzed. The six subjects that gave the best vowel discrimination were chosen. The right panel of Figure 6 shows the proportion of vowel [y] judgments by the American subjects as a function of vowel quality; F_0 pattern is the curve parameter. The likelihood of a vowel [y] judgment increased systematically with changes in the formant frequencies from vowel [i] to [y], $F(6,30) = 311$, $p < .001$. The effect of F_0 pattern was also significant, $F(6,30) = 41.8$, $p < .001$. The interaction between vowel quality and F_0 pattern was also significant, $F(36,180) = 7.83$, $p < .001$.

The American subjects also tended to hear the falling tone as more [y]-like, similar to the performance of the Chinese subjects shown in the left panel. American subjects show a boundary shift in the same direction as do the Chinese subjects, supporting a psychoacoustic explanation of the boundary shift for the Chinese listeners.

A closer comparison between the two groups of listeners also provides evidence for a psycholinguistic influence. Surprisingly, the American subjects demonstrated an even larger boundary shift than did our Chinese listeners. This difference between the two groups might be due to the contribution of vowel loudness to perceptual recognition. In natural speech, the vowel [i] is louder than the vowel [y]. FANT (1973) presents loudness levels for the first three formants for the Swedish vowels [i] and [y]. The formants are louder in [i] than in [y], F_1 and F_2 are each about 1 dB louder and F_3 is about 6 dB louder in the unrounded vowel. If Chinese listeners utilize this knowledge in perceptual recognition, then a given set of formant frequencies will have different consequences as a function of vowel loudness. A louder vowel will require a better match of the formants with those prototypical for [y] than will a softer vowel. Analogously, a relatively ambiguous vowel will be heard as [i] if it is loud and as [y] if it is soft. This analysis is based on the idea that both formant frequencies and loudness provide independent cues to vowel identity in Mandarin Chinese. If Chinese listeners use vowel loudness as a cue to vowel identity, they

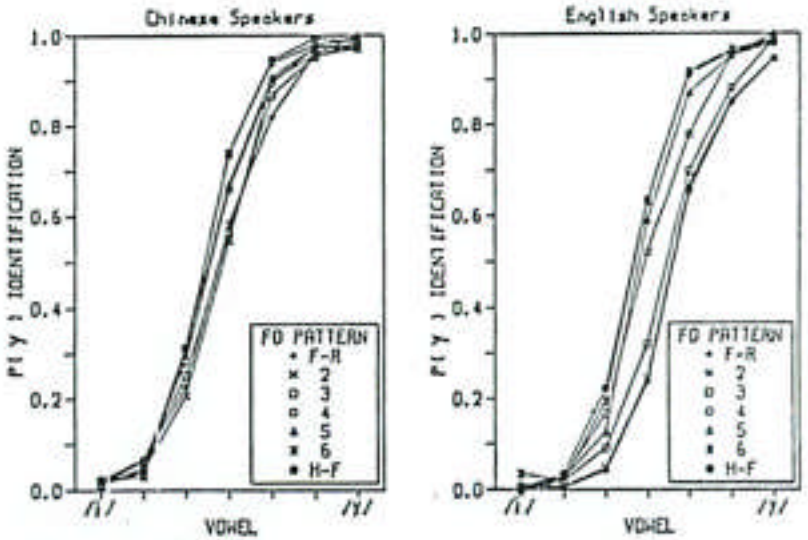

Figure 6. Percentage of y identifications for Chinese (left panel) and English (right panel) speakers as a function of vowel quality; F_0 pattern is the curve parameter (Experiments 1 and 2).

would tend to identify louder vowels as [i] rather than [y], especially for ambiguous formant frequencies. Americans, on the other hand, would not use this information since they have had no experience with the [y] alternative in natural speech.

Unlike natural speech, the loudness of the vowels did not decrease across the [i] - [y] continuum of experimental stimuli. However, the vowels decreased from 57 to 53 dB across the continuum of high-falling to falling-rising F_0 pattern. This difference might have been due to the higher onset frequency of the high-falling F_0 pattern or some other variable in the synthesis. Therefore, the louder high-falling tones will be heard as more [i]-like for the Chinese. The American subjects do not have the necessary psycholinguistic knowledge to use loudness as a cue to vowel identity. Thus, the varying loudness of the F_0 pattern should have no influence on their judgments. The Chinese are faced with two opposing influences of the F_0 pattern. The falling-rising F_0 pattern is heard as more like [i] for some as yet to be identified psychoacoustic reason. The same falling-rising F_0 pattern is softer and is heard as less like [i] because Chinese listeners know that softer vowels are less likely to be [i]. These two factors tend to cancel each other out, although the results indicate that the psychoacoustic influence must have been the stronger factor in the experiment. For the American subjects, the potential psycholinguistic influence is not present, and the psychoacoustic influence is revealed in pure form. Therefore, the American subjects show a larger boundary shift due to the F_0 pattern because their perceptual recognition of vowel quality is uncontaminated by the concomitant change in loudness of the F_0 pattern.

4. *Experiment 3*

A critical assumption in the explanation of the differences between Chinese and American listeners is that loudness provides an independent cue to vowel identity for Chinese, but not for American speakers. Experiment 3 provides a strong test of this assumption, by independently varying the loudness of the speech sounds. Three variables were factorially combined in the synthesis of the test sounds: five levels of formant frequencies going from [i] to [y], three levels of F_0 pattern going from falling-rising to high-falling tone, and three levels of vowel amplitude. The combination of these three independent variables gives a total of 45 sounds. Chinese and American subjects were tested following the same procedure as in the earlier studies. Chinese subjects identified the sounds as one of the four words, whereas the American subjects identified the sounds as one of the two vowels.

Method

Stimuli and procedure

Three variables were manipulated in this experiment: namely, vowel quality, F_0 pattern, and amplitude. Forty-five stimuli were generated by combining five levels of vowel quality, three levels of F_0 pattern, and three levels of amplitude. All stimuli were 250 msec. The five levels of vowel quality were equivalent to the levels 2 through 6 in Experiment 1. The values of the formant frequencies, F_1, F_2, F_3, and F_4 for the endpoint vowel [i] were 254, 2064, 2795, and 3562 Hz respectively. The corresponding formant frequencies for the endpoint [y] were 261, 1991, 2593, and 3466 Hz. F_5 remained constant at 4500 Hz. The F_0 pattern continuum was created by choosing steps 3 and 4 from the falling-rising to falling continuum in Experiment 1 as the range. This continuum was then divided in two equal steps, yielding three different F_0 patterns.

The three levels of amplitude in two equal steps were 0-55-54-0 dB, 0-55-50-0 dB and 0-51-46-0 dB, respectively. These sets of values were set at the onset of the vowel, 30 msec after the onset, 30 msec before the offset, and the offset. The mid-level of amplitude, i.e., 0-55-50-0 dB, corresponds to the amplitude used in the other experiments in the present study. As mentioned before, these amplitude values are specifications for the Klatt software synthesizer parameter values, not actual measurements at the subjects' ears. The corresponding amplitudes at the subjects' ears were 51, 55, and 59 dB SPL A, respectively.

The same group of six Chinese subjects participated along with another group of eight American subjects chosen from the same pool described in Experiment 2. All of the American subjects had

backgrounds in foreign languages in high school. S1 had studied French; S2, 3, 4, and 8 Spanish; S5 German and French; S6 German; and S7 Latin and Spanish. All of the subjects participated on two different days for one hour on each day following the same procedures of the previous experiments.

The Chinese subjects were given the same four response choices as in Experiment 1. The American subjects were given the same two choices as in Experiment 2. A total 24 observations on each of the 45 stimuli were collected for both the Chinese and American subjects during four sessions of the experiment.

Results

The left panel of Figure 7 shows the proportion of vowel [y] judgments by the Chinese subjects as a function of the vowel quality, amplitude is the curve parameter. For the Chinese subjects, the likelihood of a vowel [y] judgment increased systematically with changes in vowel quality from [i] to [y], $F(4,20) = 99.2$, $p < .001$. Vowel [y] judgments also increased systematically with decreases in

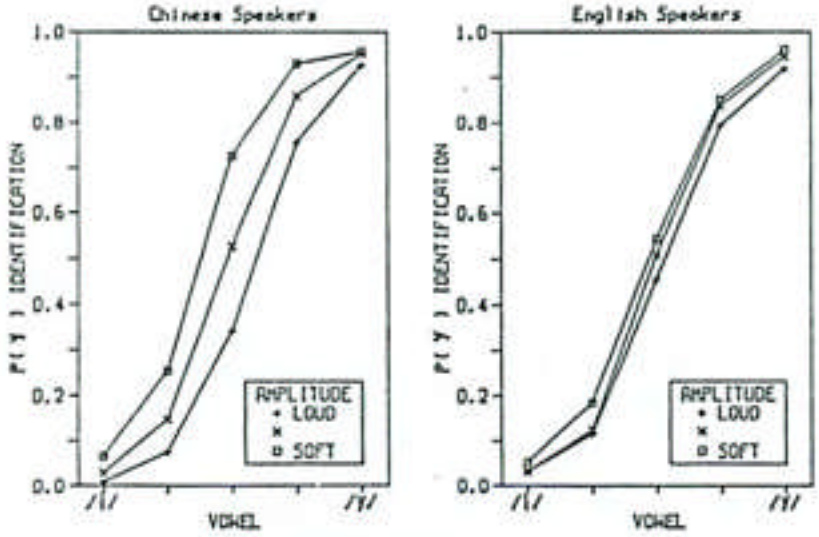

Figure 7. Proportion of y identifications for Chinese (left panel) and English (right panel) speakers as a function of the F_0 pattern; amplitude is the curve parameter (Experiment 3).

amplitude. The main effect of amplitude was significant, $F(2,10) = 73.22$, $p < .001$, as was the interaction between vowel quality and amplitude, $F(8,40) = 7.17$, $p < .001$. Although the effect of F_0 pattern was insignificant, $F < 1$, the significant interaction between vowel quality and F_0 pattern, $F(8,40) = 2.66$, $p < .05$, indicated that the vowel tended to be heard as [y] with a falling F_0 pattern.

The right panel of Figure 7 shows the proportion of vowel [y] judgments by the American subjects as a function of vowel quality; amplitude is the curve parameter. For the American subjects, the likelihood of a vowel [y] judgment increased systematically with changes in vowel quality from [i] to [y], $F(4,20) = 138.7$, $p < .001$. The main effect of amplitude was insignificant, $F(2,10) = 1.89$, $p > .2$, as was the interaction between vowel quality and amplitude, $F > 1$. Although the effect of F_0 pattern was insignificant, $F < 1$, the interaction between vowel quality and F_0 pattern was significant, $F(8,40) = 2.95$, $p < .05$. This result indicated that the American listeners, like the Chinese listeners, tended to hear the vowel as [y] when ambiguous formant values were presented with a falling F_0 pattern.

Discussion

The results of Experiment 3 demonstrated the expected effect of vowel quality and F_0 pattern on the subjects' judgments of vowels for both the Chinese and American subjects. However, a significant effect of amplitude on vowel quality judgment was found only for the Chinese subjects. Amplitude did not affect the American subjects' judgment of vowel quality. The Chinese subjects appeared to utilize amplitude as a cue for vowel quality, and tended to perceive louder vowels as more [i]-like. At the middle level of the vowel continuum, the loudest vowel was identified as [i] 39% more often than the softest vowel for the Chinese. This result is consistent with the idea that Chinese listeners utilize their knowledge of the loudness of naturally occurring vowels in perceptual recognition. American subjects do not have experience with the differences in the loudness of [i] and [y] since the latter is completely unfamiliar. Therefore,

they can not utilize loudness in the perceptual recognition of vowel quality. Consistent with this explanation, the right panel of Figure 7 shows no significant effect of amplitude on the identification of vowel quality for the American subjects.

The results of Experiment 3 confirm the explanation of the dilution of the psychoacoustic effect of F_0 pattern on vowel identification by Chinese listeners. The boundary shift indicated that the vowels were perceived as more [y]-like as they went from falling-rising to high-falling. This effect was smaller for the Chinese speakers supposedly because of the corresponding changes in the loudness of the vowels with changes in F_0 pattern. In the synthesized sounds used in these experiments, as well as in citation form of natural Mandarin Chinese, the amplitude of the falling-rising tone is less than the amplitude of the high-falling tone (LI 1964; TSENG 1981). The vowels increased 4 dB as the F_0 pattern changed from falling-rising to high-falling in Experiments 1 and 2. In Experiment 3, we found that for Chinese speakers, but not for English, louder vowels were perceived as more [i]-like. This provides an explanation of the smaller boundary shift for Chinese than for English speakers. The explanation is that for Chinese speakers the higher amplitude of the high-falling tones makes the vowel more [i]-like and thus partly counteracts the direct psychoacoustic effect of the high-falling tone, which makes the vowel more [y]-like.

5. *General Discussion*

Future experiments should be carried out to resolve some of the remaining questions concerning th recognition of vowel quality. First, it will be of interest to determine whether F_0 pattern influences the identification of other vowel contrasts. Given the putative psychoacoustic explanation, we might expect the influence to be unique to the particular formant values used in the [i]-[y] contrast. The high-falling and falling-rising F_0 patterns differed in terms of both F_0 height and contour. Accordingly, it is of interest to assess the relative contribution of each of these properties of the F_0 pattern.

In summary, we believe that the present framework offers a productive approach to the study of acoustic features in speech perception. Factorial designs have shown that a simple identification task of sounds differing on just a few dimensions produces relatively complex results. The results have revealed both psychoacoustic and psycholinguistic contributions to speech recognition. The psycholinguistic influences show that a speech distinction is conveyed by multiple acoustic features which are integrated together in a multiplicative manner. Unfortunately, we were not able to explain the psychoacoustic effect of F_0 contour on the perception of vowel quality. Although this important issue remains unresolved, the present framework appears to offer a formal theoretical and empirical approach to the study of these issues.

Acknowledgement

This research was supported by a grant from the National Institute of Mental Health (MH-35134) to Dominic W. Massaro. Requests for reprints should be sent to Dominic W. Massaro, Program in Experimental Psychology, University of California, Santa Cruz, California 95064.

References

BLUMSTEIN, S.E., STEVENS, K.N., Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America 66, 1979, 1001-1017.

CHAO, Y.R., A Grammar of Spoken Chinese. Berkeley, University of California Press 1968.

COLE, R.A., SCOTT, B., The phantom in the phoneme: Invariant cues for stop consonants. Perception and Psychophysics 15, 1974, 101-107.

FANT, G., Speech sounds and features. Cambridge, MIT Press 1973.

GANDOUR, J., The perception of tone. In V. FROMKIN (Ed.), Tone: A linguistic survey. New York, Academic Press 1978, 41-76.

KLATT, D.H., Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America 67, 3, 1980, 971-995.

LI, K-P., Machine Recognition of Mandarin Monosyllable. In Project on Linguistic Analyis Report No. 9, Columbus, Ohio, The Ohio State University 1964.

LIBERMAN, A.M., COOPER, F.S., SHANKWEILER, D.P., STUDDERT-KENNEDY, M., Perception of the speech code 74, 1967, 431-461.

LUCE, R.D., Individual choice behavior. New York, Wiley 1959.

MASSARO, D.W., Experimental Psychology and Information Processing. Chicago, Rand-McNally 1975. (a)

MASSARO, D.W. (Ed.), Understanding language: An information processing analysis of speech perception, reading, and psycholinguistics. New York, Academic Press 1975. (b)

MASSARO, D.W., Reading and listening (Tutorial paper). In KOLERS, P.A., WROLSTAD, M., BOUMA, H. (Eds.), Processing of Visible Language, 1. New York, Plenum 1979, 331-354.

MASSARO, D.W., COHEN, M.M., The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. Journal of the Acoustical Society of America 60, 1976, 704-717.

MASSARO, D.W., COHEN, M.M., The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction. Perception and Psychophysics 22, 1977, 373-382.

MASSARO, D.W., ODEN, G.C., Evaluation and integration of acoustic features in speech perception. Journal of the Acoustical Society of America 67, 1980, 996-1013.

MCCAWLEY, J.C., What is a tone language: In FROMKIN, V. (Ed.), Tone: A Linguistic Survey. New York, Academic Press 1978, 113-131.

ODEN, G.C., MASSARO, D.W., Integration of featural information in speech perception. Psychological Review 85, 1978, 172-191.

TSENG, C., An acoustic phonetic study on tones in Mandarin Chinese. Ph.D. dissertation, Brown University, Providence, R.I. 1981.

TSENG, C., MASSARO, D.W., COHEN, M.M., Lexical tone perception: Evaluation and integration of acoustic features in Mandarin Chinese, unpublished.

Oral gesturing in two unrelated Languages

I. Fónagy, Paris
M.-N. Han, Kangam-Ku Bango
P. Simon, Strasbourg

1. *Aims and methods*

In an earlier paper emotive articulatory distortions, as reflected in Hungarian cinematographic recordings, were interpreted in terms of para-linguistic oral mimicry (FÓNAGY 1976). If this interpretation was correct, similar attitudes should produce similar distortions in other unrelated languages, in French for intance.

French consonants and vowels have been analysed previously by means of cineradiographic recordings by Pêlagie SIMON (1967) and Catherine BRICHLER-LABAEYE (1970), using sentences pronounced in a neutral tone.

A corresponding study was carried out in the frame of a complex research work sponsored by the Centre National de Recherche Scientifique (Communication verbale et non verbale, contract A.T.P.no 1950). Five young artistes of the School of Dramatic Arts of Strasbourg had to recite seven sentences, suggesting different attitudes: neutral statement, surprise, joy, sadness, anger, hatred, tenderness, admiration, contempt and irony. Five judges, teachers of linguistics (of both sexes) were asked to evaluate the performances of the five actors sentence