Experiments on the Perception of Mandarin Vowels and Tones
John M. Howie†
Haskins Laboratories, New York City

I. Introduction

A. Scope of this study.

The principal aim of this study is to make some general statements about the acoustical properties of Standard Peking Mandarin vowels and tones, as they occur in citation forms of monosyllabic morphemes in a controlled environment. The generality of the description is to be attained through the use of speech synthesis, by validating measurements made from a single speaker in perception tests with a number of listeners who are speakers of the language. Generalization of the acoustic cues responsible for the perception of these vowels and tones is also achieved through a simplification of the physical dimensions of the synthesized tones and vowels. For example, pitch curves of each of the four Mandarin tones, as derived from measurements of a number of syllables, are "averaged" by drawing a smooth curve through the superimposed curves for each tone. Vowel qualities likewise are simplified through the use of "averaged" vowel patterns consisting of only two or three formants, where formants 1, 2, and 3 each have constant bandwidths and intensities.

A second purpose of this study is to investigate certain particular questions in the perception of Mandarin vowels and

7.1
tones in citation form. A comparison of listener identification of the same set of synthesized tones heard under two conditions, (a) in a carrier sentence where the following syllable has neutral tone and (b) with the carrier sentence instrumentally deleted, suggests that the pitch of that neutral tone (which is dependent on the preceding tone in a sandhi relationship) acts as a powerful cue for the perception of the tone in question, causing incorrect identifications in some cases where the synthesized tone is different from the tone originally spoken in the sentence. There is also a question of vowel quality in the syllables you-ju, wei-ui, and wen-un where they occur as the raised tone-3 syllable resulting from tone sandhi in disyllabic words or phrases. Both of these matters will be explored in experiments with instrumentally manipulated speech.

B. Instrumentation and Experimental Technique.

In addition to the usual tape recorders and the modified Kay Sonagraph used to produce the sound spectrograms from which fundamental frequency and formant measurements were made, two instruments for the synthesis and manipulation of speech were used in this study. These are (a) the Vocoder with Digital Spectrum Manipulator, used for deleting the carrier sentence and for synthesizing tones, and (b) the Pattern Playback, used for synthesizing vowels. These instruments were made available for this study through the kind cooperation of Dr. Franklin S. Cooper, President of the Hamkins Laboratories, New York City, for which the author wishes to make grateful acknowledgment.

In the pilot tests to be described here, tape recordings of synthesized tones from the Vocoder-DSM and of synthesized vowels from the Pattern Playback were arranged in random order -- connected by four seconds of silence -- and presented to five Mandarin speakers for identification. Because a frequently-occurring syllable may represent well over a dozen different morphemes (due to the small number of possible syllables in
Mandarin), and to avoid difficulties in reading hurriedly-written Chinese characters, multiple-choice answer sheets were prepared, on which the listener simply circled one of several characters corresponding to the syllables among which he was being asked to discriminate.

II. Tones
A. Pilot tests with Vocoder output.

Pilot tests were made including not only synthesized tones but also "real-speech" tones, for which the original tone contours were retained in order to verify the hypothesis that real-speech tones in citation form can be recognized when isolated from their context. Four sets of syllables minimally differentiated by tone, i.e., rao, ma, ke, bao, were processed with the DSM to delete the environmental syllables of the carrier sentence. One subtest of twenty items was made from each of these sets, with five tokens of each of the four tones. Identification by the five listeners in all four subtests of real-speech tones was 96 percent correct.

For the pilot test of synthesized tones, only one syllable was used as a base upon which to construct the four generalized contours. This syllable, bao (high level tone), was processed with the DSM, manipulating the course of the fundamental frequency, and also deleting the carrier sentence, to produce two sets of syllables minimally differentiated by tone, i.e., bao in the carrier sentence and bao in isolation. Again, one subtest of twenty items was made from each set, with five tokens each of the four tones. Except for the randomization, the only difference in these two subtests was the presence of the environmental syllables in one of them. This environment -- probably the neutral tone of the unstressed syllable which immediately follows the tested tone -- must account for any differences in the identification of the tones by the listeners. With the syllables in isolation, it was 100 percent correct; but with the
AVERAGE CURVES FOR MANDARIN TONES

- TONE 1 "High Level"
- TONE 4 "High Rising"
- TONE 2 "High Falling"
- TONE 3 "Low Dipping"
carrier sentence included, it was only 80 percent correct. Of the twenty errors, nineteen consisted of identifying tone 3 as tone 2. In the other six of the twenty-five trials of tone 3, it was recognized successfully.

B. Discussion of results and tentative conclusions.

Two points are worthy of note in the results of the pilot tests of synthesized tones. One concerns the attempt to make the most general statements about the fundamental-frequency/time patterns of the four distinctive tones by synthesizing the tones according to a set of average pitch curves. From the results of the subtest where the tones were heard in isolation, at least, it is reasonable to assume that these average curves represent a valid acoustic description of the Mandarin tones. (It is to be hoped, of course, that results obtained with further listeners will substantiate this preliminary validation.) The interesting point, however, is that these average curves were generalized from curves which exhibit several variations in shape within each tone, associated with different syllable structures, i.e., the type of -- or absence of -- phonetic segment preceding the main vowel. By suppressing these variations in the average curves, the intention was to test the assumption that only the most essential features of a tone -- those shared by syllables of all structures -- are necessary for its perception. This assumption is upheld in the pilot test where all listeners recognized all tokens of the four synthesized tones in isolation.

The other point concerns the effect of the carrier sentence's neutral tone on the recognition of the preceding citation tone. Since the utterance from which all the synthesized tones were produced contained the citation tone 1, it will be understood that the pitch of the neutral tone following it, when the carrier sentence was heard, was "wrong" with the other three synthesized tones. Normally, tone 1, which is high and level, is followed by a neutral tone with a relatively low pitch; and
tone 3, which is low and dipping, is followed by a neutral tone with a relatively high pitch. The sentences containing the synthesized tone 3, then, had a sequence which is never heard in real speech: a stressed syllable with the low tone 3 followed by an unstressed syllable also with low pitch. But tone 2, which is rising, normally is followed by a neutral tone with a pitch in the mid range, which would be a sequence somewhat similar to the unnatural sequence obtained with the synthesized tone 3 -- except for the normal difference in the average pitch ranges between tones 2 and 3. Apparently the only way the listeners could interpret and accept the sequence of synthesized tone 3 plus low neutral tone was as a deviant sequence of tone 2 plus neutral tone, the deviations being a rather low pitch range in the whole sequence and less than a normal rise for the tone 2. A full series of perception tests is planned, using each of the four tones in turn as the syllable base for synthesizing the tones, to see whether comparable interpretations will be made of other unnatural sequences.

III. Vowels

A. Pilot tests with monosyllables from the Pattern Playback.

One pilot test was originally made of synthesized vowels, consisting of two subtests of different syllable types. In one of the subtests, the listeners were asked to discriminate among seven representative vowels, as found in the syllables xi, xu, shi, hu, si, she, and ha, with three tokens of each, totaling twenty-one test items. Beyond being simply a check on the distinctness of the synthesized vowels, however, this subtest was intended to explore the feasibility of including an ambiguous friction noise before each vowel, so that the two vowels which occur only in syllables with an initial consonant (shi, si) could be tested in a series of consonant-vowel syllables. The ambiguous noise, rather than a specific fricative, is needed --
Intervals of vowel duration between the connected points for a given syllable are equal. For syllables with final /n/, the last point is 100 percent of vowel duration; for ye and wo, it is 80 percent; for she, it is 66 percent. For xi, xu, and shi, the two connected points are the beginning and mid-point in the duration; for vowels with only one point, it is the mid-point.

Average frequencies for formant 3:

<table>
<thead>
<tr>
<th></th>
<th>beginning</th>
<th>mid-point</th>
</tr>
</thead>
<tbody>
<tr>
<td>xi</td>
<td>3825</td>
<td>3550</td>
</tr>
<tr>
<td>xu</td>
<td>3075</td>
<td>2750</td>
</tr>
<tr>
<td>shi</td>
<td>3800</td>
<td>3500</td>
</tr>
<tr>
<td>si</td>
<td>3200</td>
<td></td>
</tr>
<tr>
<td>she</td>
<td>3000</td>
<td></td>
</tr>
<tr>
<td>ha</td>
<td>2675</td>
<td></td>
</tr>
<tr>
<td>hu</td>
<td>none</td>
<td></td>
</tr>
</tbody>
</table>
if discrimination is to depend entirely on the vowel -- because no one consonant can occur with all seven vowels, due to the constraints on the distribution of segmental units in the Mandarin syllable. Identification by three listeners in this subtest was only 50 percent correct, but was perfect for hu -- while shi and xu were recognized in only three trials each and xi only twice, out of the nine trials of each vowel. In the other subtest, the listeners were asked to identify ten syllables, making a choice for each test item from only two possibilities. There were twenty items, two tokens each of the following syllables (arranged here in the pairs from which the choices were to be made): ye - ya, wo - wa, yin - yan, yun - yuan, en - an. The point of this set of syllables was to distinguish mid vowels from low vowels, within the small number of permitted sequences in Mandarin. Identification by five listeners in this subtest was 96 percent correct.

A second pilot test was subsequently made of the consonant-vowel syllables, simply by supplying new stimuli for three of the syllables in the first subtest. Attempts were made to improve the synthesized vowels in the syllables which fared the worst in the first perception test; xi, xu, and shi. This time, identification by three listeners was about 60 percent -- only a slight overall improvement -- and was perfect this time for three syllables: hu, she, and ha. For xu, recognition improved by two trials compared with the first test, but for shi and xi it was worse by one trial each. Because of these results, the vowel in xu was tentatively considered acceptable, and effort was concentrated on making xi and shi distinct. As a part of this effort, a series of vowel patterns was prepared in which the first, second, and third formants were varied systematically for the vowel in xi, and the second formant only was varied for the vowel in shi. This series of stimuli will be presented to
listeners to learn which xi and which shi they prefer. (For this experiment, the bandwidth of the friction noise for the retroflex sh was made to correspond more closely to that of real speech.) A third test will then be made, in which new stimuli for these two syllables will replace those of the original subtest.

B. Discussion of results and tentative conclusions.

From the results of the pilot tests so far completed of synthesized vowels, it is possible to assume that the averaged vowel formant patterns represent a valid acoustic description of three or four out of the seven vowels of the CV syllables, and of all of the mid and low vowels of the VV, VN, and VVN syllables. As might be expected, the synthesized diphthongs were strikingly more recognizable as speech than the steady-state vowels, the moving formants giving rise more readily to the perception of a linguistic event -- rather than of a musical or mechanical event. Further, the device of including the ambiguous friction noise to produce the series of CV syllables may be considered successful, since nearly half of the syllables in the second pilot test were recognized in all trials, in spite of the ambiguous and unnatural-sounding consonant. If the new stimuli for xi and shi succeed in making these syllables distinct (the responses in the second pilot test showed both of them to be confused mainly with xu), it seems reasonable to expect that the preliminary validation will be extended to all seven of these vowels.

The decision to use three-formant vowel patterns -- except for a two-formant u -- was based on the primary aim of making the most valid description of the language, rather than of investigating the perception of the more simplified vowel-like sounds which would be synthesized from two-formant patterns. Another factor in this decision was the nature of some of the vowels in the language. Two areas of vowel distinction in
Mandarin seem to depend on the presence of $F_3$ and its relation to $F_2$: (a) the r-colored vowel of the syllable er (not included in the pilot tests), and (b) the distinctions among the vowels of xi, xu, and shi. From the formant measurements of the informant's speech, there appears to be little difference in the frequency of $F_2$ between the vowels of xi and xu, and a relatively greater difference in the frequency of $F_3$. In the attempts to make xi more distinct from xu, while retaining the uniform dimensions of the formant bandwidths and intensities, the only fruitful course seemed to be to raise $F_2$ of xi to a frequency above 2700 Hz. (The frequency adopted for $F_2$ of xu is 2340 Hz.) In the experimental series of patterns made for the vowel of xi, this vowel sounds convincing only when $F_2$ is 2700 Hz. or higher. The vowel of shi exhibits downward shifts in $F_2$ and $F_3$ and an upward shift in $F_1$ (as do the formants of the vowels of xi and xu to a much lesser extent). In order to adopt a steady-state set of formant values for shi, comparable to those for the other high vowels, a frequency range close to 2000 Hz. was chosen for $F_2$, together with a rather high $F_3$. This optimum range for $F_2$ may explain the auditory impression that the vowel of shi is in some way similar to the r-colored vowel of er, whose $F_3$ has a distinctively low frequency near 2000 Hz. (It is interesting to note in this connection that the Yale romanization uses the letter r for this vowel, where the Pinyin system uses i, resulting in the spelling shr for shi.)

Other patterns were prepared for both steady-state vowels (without initial consonant) and diphthongs, and stimuli were recorded from the Pattern Playback for use in later perception tests with a larger number of listeners. A number of real-speech vowels were also processed with the Vocoder-DSM, to produce monotone syllables which will allow the perception of the syllables you-ju, wei-ui, and wen-un to be studied, in tests.
where the identification of morphemes normally distinguished mainly by tone will depend on variation in vowel quality alone.

†Also, the University of Indiana.