The Spectral Dynamics of Vowels in Mandarin Chinese

Jiahong Yuan
Linguistic Data Consortium, University of Pennsylvania
jiahong@sas.upenn.edu

Abstract
This study investigated the dynamic spectral patterns of vowels in Mandarin Chinese using a corpus of monosyllabic words spoken in isolation. Mel-frequency cepstral coefficients (MFCCs) were parameterized in different ways to test the nature of the dynamic information in vowels through automatic vowel classification. Compared to the MFCCs extracted at the vowel midpoint, using the MFCCs extracted at two or three points (vowel onset, offset, and midpoint) greatly improved classification accuracies. Legendre polynomials fitted to the MFCCs over the entire vowel duration achieved approximately 30% relative error reductions over the three-point model. Euclidean cepstral distance was employed to measure the magnitude of spectral change. A negative correlation was found between the rate of spectral change and vowel duration. Vowel-dependent spectral changes appear primarily in the first half of a vowel. There is great diversity among the diphthongs and a considerable overlap between the diphthongs and the monophthongs in terms of the spectral dynamics.

Index Terms: vowels, diphthongs, triphthongs, spectral dynamics, Mandarin Chinese

1. Introduction

1.1. The dynamic characteristics of vowels
Speech production is inherently dynamic. There has been an extensive debate in the literature about the nature of dynamic information in vowels. For example, the International Phonetic Alphabet depicts vowels as static targets on a quadrilateral chart, whereas in the framework of articulatory phonology [1], vowels are associated with dynamic gestures that are realized on a time span.

Monophthongal vowels (often simply called vowels) have been traditionally characterized with a static target, a “steady state” that is relatively unaffected by surrounding consonants [2, 3]. However, transitions associated with consonants also provide useful information in vowel identification. For example, [4, 5] found that vowels could be better identified in CVC syllables than in isolated, steady-state form and [6] found that vowels in /bVb/ syllables could be reliably identified even when vowel nuclei were replaced with silence. [7] suggested that the role of consonantal transitions in vowel perception might be due, in part, to vowel-inherent spectral change rather than consonantal context. In their experiment, two brief sections were extracted from “nucleus” and “offglide” portions of naturally produced vowels. The two sections were presented to listeners in three conditions: nucleus followed by offglide (natural order), nucleus followed by itself, and offglide followed by nucleus (reverse order). Their results showed that listeners’ error rates for the natural order condition were comparable to those for unmodified full vowels. However, the error rates for the reverse order and repeated nucleus were significantly higher. These results provided evidence for the importance of vowel inherent spectral change in vowel perception.

It has been long debated whether a diphthong is a sequence of two vowels or simply a vowel with continually changing quality. Many models have been proposed in the literature to characterize the nature of spectral dynamics in diphthongs. The “dual target” model claims that a diphthong contains both an onset target and an offset target [8]. [9] used two exponential functions to approach the two targets. The “onset plus slope” model claims that diphthongs can be classified in terms of formant values at diphthong onset and the rate of F2 change [10]. The “onset plus direction” model claims that it is the direction of spectral change rather than the rate of spectral change over time that is stable [11]. [12] proposed a “truncation model” to explain the acoustic structure of diphthongs and triphthongs in Mandarin Chinese. The model claims that, at the speech-planning level, all of the targets in a syllable are located at the syllable onset position. Each adjacent pair of targets has a transition of a specified rate, and the acoustic realization is achieved by a truncation process between two adjacent transitions.

Another line of research has investigated the nature of vowel spectral dynamics through automatic vowel classification. Most of these studies have used formant frequencies and examined the effects of different parameterizations of formants on classification results. A number of studies have found that monophthongs can be better separated based on the two-point (formants measured at vowel onset and offset) or three-point (formants measured at vowel midpoint, onset and offset) models than on the one-point model (formants measured at the midpoint only) [13-15]. However, [16, 17] reported that using three points or modeling the formant contour with discrete cosine transform (DCT) coefficients did not achieve better results than the use of the midpoint only. [18] demonstrated that vowel classification based on cepstral coefficients was superior to classification based on formants and that DCT-parameterized dynamic information outperformed the one-point model. [19] found that the “dual target” model outperformed the “onset plus rate” and “onset plus direction” models in diphthong classification. [16] showed that vowels, including both monophthongs and diphthongs, were not better classified from a time-delay neural network than from the three-point model in which time is not explicitly represented, suggesting that the temporal order of the three points is not important in vowel classification.

1.2. Vowels in Mandarin Chinese
Syllables in Mandarin Chinese consist of two parts: initials and finals. Vowels appear only in the finals. While researchers have disagreed on the inventories of vowel phonemes in Mandarin Chinese (see discussion in [20]), the inventories of syllables, initials and finals in the language are largely straightforward. A final in Mandarin Chinese may consist of one or more vowels (or vowels and glides, depending on the adopted phonological analysis), with or without a nasal coda. Table 1 lists the finals without a nasal coda, transcribed in
Pinyin (a Roman alphabet system for transcribing Chinese characters) and IPA. The IPA transcriptions were adopted from a textbook compiled by the faculty in Modern Chinese at Beijing University [21].

Table 1. Finals without a nasal coda in Mandarin Chinese, transcribed in Pinyin and IPA (IPA symbols are in square brackets).

Table 2. The distribution of vowels in the corpus (24,528 tokens in total).

Table 3. Analysis

3.1. Automatic classification of vowels

The Mel-frequency cepstral coefficients (MFCC) [28] as well as the tonal category (Tone 1, Tone 2, Tone 3 or Tone 4) and vowel duration were used as features in the vowel classification experiment. Thirteen MFCC coefficients were extracted using a 25ms Hamming window, based on 20 Mel filters ranging from 0Hz to 5000Hz for male speech and from 0Hz to 5500Hz for female speech. No further speaker normalization was applied to the coefficients. The MFCC coefficients were then parameterized in the following ways: 1. MFCCs extracted at 50% of the duration of the vowel; 2. MFCCs extracted at 20% and 80% of the duration of the vowel; 3. MFCCs extracted at 20%, 50% and 80% of the duration of the vowel; 4. the first three coefficients of the discrete cosine transform (DCT) of each MFCC coefficient; 5. the first three coefficients of Legendre polynomials fitted to each MFCC coefficient; 6. contour over the entire vowel duration.

Table 4 lists the classification accuracies from the use of different MFCC parameterizations when the classification was performed over all vowels, including monophthongs, diphthongs and triphthongs. This task is comparable to what listeners and language learners do. From the table, we can see that the two vowels could hardly be separated for example, 40.5% of the tokens were misclassified as /uo/. This result suggests that the acoustic difference between the two vowels is negligible. In Mandarin Chinese, /o/ and /uo/ are in complimentary distribution: /o/ only appears after a labial initial (b, p, m), whereas /uo/ does not appear after a labial initial. Because the classification results indicate little acoustic difference between the two vowels, they are treated as one category, /ou/, in the following discussion.

Table 5 lists the classification accuracies when the classification was performed over all vowels, including monophthongs, diphthongs and triphthongs.

Table 6 lists the classification accuracies when the classification was performed over all classes, including monophthongs, diphthongs and triphthongs, respectively. We can see that using the mid-point only, the monophthongs (accuracy = 96.3%) and the triphthongs (accuracy = 94.1%) could be better separated than the diphthongs (accuracy = 85.2%). Legendre polynomials outperformed the three-point model, with relative error reductions of 34% (from 3.4% to 2.23%) and 38% (from 3.4% to 2.1%), respectively.

Table 7 lists the classification accuracies when the classification was performed over all the three classes, including monophthongs, diphthongs and triphthongs, respectively. We can see that using the mid-point only, the monophthongs (accuracy = 96.3%) and the triphthongs (accuracy = 94.1%) could be better separated than the diphthongs (accuracy = 85.2%). Legendre polynomials outperformed the three-point model for all three classes, with a relative error reduction of 31% for monophthongs (from 1.6% to 1.1%), 32% for diphthongs (from 3.4% to 2.23%), and 38% for triphthongs (from 3.4% to 2.1%).
diphthongs (from 1.9% to 1.3%), and 29% for triphthongs (from 1.4% to 1.0%).

Table 3. Confusion matrix between /o/ and /uo/ (17 /o/ tokens were, for example, classified as other vowels).

<table>
<thead>
<tr>
<th></th>
<th>/o/</th>
<th>/uo/</th>
<th>other vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td>/o/</td>
<td>277 (56.1%)</td>
<td>200 (40.5%)</td>
<td>17 (3.4%)</td>
</tr>
<tr>
<td>/uo/</td>
<td>13 (0.7%)</td>
<td>1713 (98.4%)</td>
<td>14 (0.8%)</td>
</tr>
</tbody>
</table>

Table 4. Percentage accuracies of classification over all vowels from 10-fold cross validation.

<table>
<thead>
<tr>
<th>MFCC parameterizations</th>
<th>Number of features</th>
<th>Average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>One point</td>
<td>15</td>
<td>77.74%</td>
</tr>
<tr>
<td>Two points</td>
<td>28</td>
<td>95.62%</td>
</tr>
<tr>
<td>Three points</td>
<td>41</td>
<td>96.60%</td>
</tr>
<tr>
<td>DCT coefficients</td>
<td>41</td>
<td>97.77%</td>
</tr>
<tr>
<td>Legendre Polynomials</td>
<td>41</td>
<td>97.90%</td>
</tr>
</tbody>
</table>

Table 5. Percentage accuracies of classification for monophthongs, diphthongs and triphthongs, respectively, from 10-fold cross validation.

<table>
<thead>
<tr>
<th>MFCC parameterizations</th>
<th>Mono-</th>
<th>Diph-</th>
<th>Triph-</th>
</tr>
</thead>
<tbody>
<tr>
<td>One point</td>
<td>96.3%</td>
<td>85.2%</td>
<td>94.1%</td>
</tr>
<tr>
<td>Two points</td>
<td>97.8%</td>
<td>97.8%</td>
<td>96.5%</td>
</tr>
<tr>
<td>Three points</td>
<td>98.4%</td>
<td>98.1%</td>
<td>98.6%</td>
</tr>
<tr>
<td>Legendre Polynomials</td>
<td>98.9%</td>
<td>98.7%</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

3.2. Patterns of spectral change

Euclidean cepstral distance was employed to measure the magnitude of spectral change: The Euclidean distance between two MFCC vectors extracted at different points in a vowel measures the magnitude of spectral change between the two points. The distances were computed using 12-dimensional MFCC vectors (the zeroth MFCC coefficient was excluded to remove the effect of the vowel amplitude). The MFCC coefficients were normalized by speaker and by MFCC dimension using z-scores before calculating the Euclidean distances.

Figure 1 plots the cepstral distance (cepstral change) for every 10% of vowel duration from the onset to the offset. For example, the 0-10% distance was the distance between the MFCC vector extracted at 10% of the vowel duration and that extracted at the onset of the vowel. From Figure 1, we can see that all vowels start with a relatively high distance, indicating rapid spectral changes near the vowel onset. These changes near the vowel onset can be explained by coarticulatory transitions from syllable-initial consonants to vowels. We can also see in Figure 1 a vowel-independent pattern of spectral change toward the vowel offset: the distance becomes larger and increases at a similar rate for all vowels. The increasing spectral change toward the vowel offset cannot be explained by coarticulation because the vowels are in open syllables spoken in isolation, and it cannot be explained by the amplitude change toward the vowel offset because the zeroth MFCC coefficient was excluded. This change may represent the characteristics of the source, the vocal fold vibration, in the concluding phase of vowel production in isolation (For example, [32] found that many utterances appear to end in a “breathy-laryngealized” type of phonation). Vowel-dependent spectral changes mainly appear in the first half of vowel duration. For monophthongs, the magnitude of spectral change quickly decreases after the onset and remains level. For triphthongs the magnitude of spectral change increases to reach a peak before 30% of the vowel duration and then decreases. Diphthongs show a great diversity – for some diphthongs (e.g., /ia/) the magnitude of spectral change reaches a peak before 20% of the vowel duration, whereas for others (e.g., /ie/) a peak is not reached until 40%-50% of the vowel duration (/üe/ is different from the other diphthongs, but /üe/ has only 50 tokens; hence, its result may not be robust).

Figure 2 plots the rate of spectral change in the portion of vowel “target”, i.e., from 20% to 70% of the vowel duration (determined from Figure 1). The rate was calculated by dividing the total spectral change in the “target” portion by the duration (in seconds) of the portion, and the total spectral change was calculated by adding the cepstral distances for every 10% of duration within the “target” portion (i.e., 20%-30%, 30%-40%, ..., 60%-70%).

Figure 2. The rate of spectral change between 20% and 70% of vowel duration (ii = ɨ; iii = ɨ; ve = üe).
From Figure 2, we can see that there is a negative correlation between the rate of spectral change and vowel duration \( (r = -0.58; p < 0.001) \). Longer vowels have a slower spectral change rate. The vowels are in monosyllabic words with different tones spoken in isolation. Because Mandarin tones differ with respect to duration - the average durations of Tone 1 to Tone 4 in the dataset are 0.373s, 0.394s, 0.452s, and 0.320s, it may be possible that the correlation between vowel duration and the rate of spectral change shown in Figure 2 is an artifact of tonal differences. To test this hypothesis, a full-factorial ANOVA on the rate of spectral change was performed, using tonal category, vowel duration, and vowel identity as the main factors. The result does not support the hypothesis. It shows that both vowel duration and tonal category are significant factors for predicting the rate of spectral change (\( p < 0.01 \)). Moreover, vowel duration as a main factor explains 39% of the total variance in the rate of spectral change, whereas tones explain only 9% of the variance as a main factor. The negative correlation between vowel duration and the rate of spectral change suggests that the total spectral change in a vowel is more stable than the rate of spectral change when vowel duration varies.

Finally, both Figure 1 and Figure 2 show that the triphthongs are categorically different from the monophthongs, and there is great diversity among the diphthongs. Many diphthongs behave like monophthongs in terms of the pattern of spectral change.

4. Discussion

Legendre polynomials fitted to the MFCC coefficients over the entire vowel duration achieved about 30% relative error reductions over the three-point model in automatic vowel classification for all vowel classes, including monophthongs, diphthongs, and triphthongs. This result demonstrates that the dynamic information in vowels can be better modeled using a time-varying function than a sequence of static targets and supports the claim that vowels are inherently dynamic. The vowel-inherent spectral change is realized on the first half of a vowel (as shown in Figure 1). This is true even for the diphthongs and triphthongs. This result further demonstrates that the diphthongs and triphthongs are not a sequence of simple vowels that are realized one after another. Instead, all vowels, including monophthongs, diphthongs, and triphthongs, have a unitary and dynamic target.

The triphthongs in Mandarin Chinese are categorically different from the monophthongs in terms of the trajectory and the rate of spectral change (as shown in Figures 1 and 2). There is great diversity among the diphthongs and a considerable overlap between the diphthongs and the monophthongs. Therefore, the distinction between diphthongs and monophthongs is not categorical, but gradual. The vowels /uo/ and /o/, conventionally treated as a diphthong and a monophthong, can hardly be separated by the classifiers. Whether the two vowels merged or are “incorrectly” transcribed in the pinyin system, the non-separability between these vowels illustrates that there is no categorical distinction between diphthongs and monophthongs in Mandarin Chinese.

Both tone and vowel duration affect the rate of spectral change in vowels, but duration accounts for significantly more variance than tone. Longer vowels have a lower rate of spectral change, suggesting that the magnitude of spectral change of a vowel (in the citation form) is more stable than the rate of spectral change when the duration of the vowel varies. This result does not favor the “formant undershoot” model of vowel reduction [33, 34] or the “onset plus slope” model of diphthongs [10]. In both of these models, the rate of formant frequency change is not dependent on vowel duration. The effect of tone on the rate of spectral change requires further investigation. On one hand, tone production is the vibration of the vocal folds (i.e., the “source”); on the other hand, tone production may affect the position of the articulators (i.e. the “filter”). Both of these may play a role in the spectral change measured in Euclidean cepstral distance.
5. References


