语音乐律研究报告

Status Report of Phonetic and Music Research 2012

北京大学中文系语言学实验室

Linguistics Lab
Department of Chinese Language and Literature
Peking University

目　　录

Phonetic Study on Phonations in China

KONG Jiangping

The study on phonation has a long history in China. A story about the production and application of artificial larynx had been documented in an ancient Chinese book titled “Mengxi Bi Tan” (writing in the Mengxi Garden) by Shen Kuo (1031-1095) (Hu, 1957) in Song dynasty. "Devices made by people from materials, such as bamboo, wood, ivory and bone, were called sound generators. A sound generator, which could be put into the throat and produce speech sound by whistling, was called a voice generator. A dumb person, who suffered from injustice, could not argue in court for himself. The judge let people put a voice generator in his throat, and asked him to speak. The speech articulated like puppet talk, but could roughly make sense. His injustice was finally redressed. The case is worthy of being documented." This story vividly described the production and usge of artificial larynx in ancient time of China. Along with the development of speech technology, phonation study has being carried forward in different fields, such as linguistics, speech physiology, speech pathology and speech engineering.

Linguistic study on phonations in the languages of China started in the 50s of last century and the most of phonation types were described by Chinese linguists in the 80s. The phonetic study on phonation types in languages of China started in 1980s by professor Ladefoged and Maddieson in UCLA (Maddieson et al, 1985, 1986) and the most of phonetic studies on phonation types were done by professor Bao Huaiqiao and Kong Jiangping, two Chinese phoneticians in Chinese Academy of Social Sciences and Peking University in 1990s (Bao at al, 1990, 1992; Kong, 1993, 1995, 1996, 1997a, 1997b, 1997c, 1998). In addition, Professor Li Shaoni and Caojianfen studied the phonation types in Bai lanaguage and Chinese Wu dialects (Li et al, 1990, 1997; Cao et al, 1992). In this century, the phonation studies and modeling in physiology, speech production and oral cultures in China were done by professor Kong (Kong, 2001a, 2001b, 2003, 2004a, 2004b, 2005, 2007, 2009, 2011a, 2011b) and the phonetic studies on phonation types in minority languages and Chinese dialects were done by professor Zhu Xiaonong (Zhu, 2006, 2008, 2009, 2010). In the following sections, the phonation types in the languages of China, the voice signals and research methods, the phonetic studies on phonation types and linguistic discussion are briefly introduced.

1. Phonation in languages of China

There are more than 100 languages in China, in which about eighty had been identified academically. Phonation types which are significant in distinguishing meanings appear in many of these languages. Phonation as a phonetic phenomenon was found very early by Chinese linguists in the linguistic study, such as the "tense" and "lax" vowels in Yi language, the voice aspiration in Miao language, the tense and lax vowels and the voiced aspiration in Wa language, the fricative vowel in Western Yugu language, the Yin and Yang vowels in Mongolian, the tense and lax consonants in many language and so on. The phonation differences are mainly between vowels and sometimes between voiced consonants and even tones. The studies show that the phonation types exist in Mongolian, Tibetan, Uygur language, Bai language, Yi language, Hani language, Western Yugu language, Zhuang language, Miao language, Wa Language, Shui language, Zaiwa language, Jingpo language and so on. The different phonations which are significant linguistically are exactly described and used in the historical linguistic study (Ma, 1948; Hu et la, 1964).

Different phonation types mainly exist in the Loloish languages. Huni language has 10 tense vowels and 10 lax vowels which are very good materials for phonetic study on the phonation types of vowels (Li & Wang, 1986). Nu language has 39 vowels, among which there are 12 modal vowels, 8 tense vowels and 19 nasalized vowels, 3 retroflex vowels, 4 nasalized retroflex vowels, 1 tense retroflex vowel, and 1 tense nasalized retroflex vowel (Sun and Liu, 1986). Bai language has 6 to 8 tones among which 3 tones are with creaky voice at the end of syllables which higher pitches comparative with the other tones. Usually there tones are described as tense tones and lax tones linguistically. (Xu and Zhao, 1984). Lisu language has 6 tones which have close relationship with the tense and lax vowels and are usually described as tense tones and lax tones linguistically. (Xu, Mu and Gai, 1986). Lahu language has 5 lax tones and 2 tense tones which are produced by the tense vowels. Usually these tones are described as tense tones and lax tones linguistically. (Chang, 1986).

Zaiwa language which belongs to the Burmese languages, has 5 lax vowels, 5 tense vowels and 68 finals. (Xu, 1984). Miao language which belongs to Miao-Yao languages, has very complex initials and very simple finals. Tones in Miao have are correspondent to Chinese tones (Wang, 1958). Wa language belongs to Austro-Asiatic languages and has 18 vowels among which there are 9 lax vowels and 9 tense vowels, 28 diphthongs among which there 14 tense diphthongs and 14 lax diphthongs and 4 triphthongs among which 2 are tense triphthongs and 2 are lax triphthongs.(Zhou and Yan, 1984). Mongolian, which belongs to Mongolian language groups in Altaic language family, has 12 long vowels which are lax vowels and 11 short vowels among which 7 are tense vowels when they appear in the first syllable of a word. (Daobu, 1983). Western Yugu, which belongs to Turkic language group in Altaic language family, has 6 vowels with short fricative property. (Chen and Lei, 1984). Korean in China is the main language spoken by Korea people who are distributed in Liaoning, Jilin and Heilongjiang provinces. Most scholars regard Korean as a language in Altaic language family. Korea has one set of lax consonants among which 3 are lax stops, 1 lax affricate and 1 lax fricative, and one set of tense consonant among which 3 are tense stops, 1 tense affricate and 1 tense fricative.(Xuan, Jing and Zhao, 1985)

2. Voice signals and research methods used in China

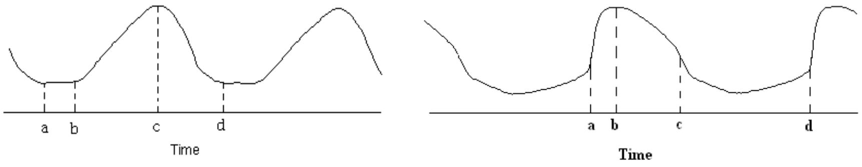
The research methods of phonation depend on the signals. The main voice signals are: 1) speech signal, 2) electroglottography (EGG), 3) signals of air flow and air pressure, 4) laryngeal electromyography (EMG), 5) photoglottography (PGG), 6) fiber stroboscope video, 7) high-speed filming (HSF), 8) high-speed digital imaging (HSDI). In the phonetic study on phonation types in China, two signals are most popularly used which are: 1) speech sound by which acoustical parameters such as ratio of harmonics, voices spectrum tilt, fundamental frequency (F0), open quotient (OQ) and speed quotient (SQ) can be extracted to describe phonations; 2) EGG signal which reflect the contact area of vocal folds in vibration is captured through laryngograph. Since long EGG signal from which physiological parameters such as F0, OQ, SQ, jitter and shimmer can be extracted for phonation study can be easily sampled without invasiveness, it can be used in many research fields. Another signal which is very important in phonation study is high-speed digital images captured by high-speed digital imaging system through endoscope or fiber scope. This kind of signal is not easily collected, but high-speed images from which parameters such as glottal area function, areas of left, right, anterior and posterior glottis, lengths of anterior and posterior glottis, widths of left and right glottis, glottal opening instant, glottal closing instant, F0, OQ and SQ from glottal area function etc. can be extracted is very useful to explain significances of phonations in language communication. With these parameters talked above, different phonation types can be described and defined and glottal geometrical model can be established.

The methods usually used in China are: 1) harmonic analysis, 2) inverse filtering analysis, 3) electroglottography analysis, 4) spectrum tilt analysis, 5) high-speed imaging analysis, 6) multi-dimensional voice processing, 7) pitch range analysis and 8) voice attack time analysis, among which the first three are often used by Chinese phoneticians. In this section，some of these methods are briefly introduces.

Harmonic analysis which is a method often used by phonetician is a simple and easily used method in research of phonation types, because speech sampling and spectrum analysis are all very easy with a personal computer. In the last century, many researches on phonation types including some phonation types in languages of China had been done with this method by the researchers (Ladefoged, 1973, 1988; Laver,1980; Ladefoged et al, 1987, 1988; Kirk et al，1984; Maddieson et al，1985; Bao, 1990; Kong, 1993). Harmonic analysis for phonation main depends on the power or amplitude of different frequencies in a voice. The acoustical principle is that when a voice has large power or amplitude in high frequency, the amplitude of the second harmonic is large than that of the first harmonic.

Inverse filtering analysis is an easily used method in the study of phonation types and also often used by speech engineers, because the speech signal is easily captured and the signal processing of inverse filtering is not difficult to implement by a personal computer. By this way, resonance of the vocal tract can be eliminated from the speech sound and finally obtain the source of speech (Alku P., 1991; Fant G. et al, 1994). The signal processing used in inverse filtering is LPC (linear predictive coding). The steps for inverse filtering are: 1) Calculating LPC coefficients from a period of speech sound (which is pre-emphasized); 2) designing an inverse filter by the LPC coefficients; 3) Filtering the speech sound to get the source of voice.

Electroglottograph analysis is a method by which parameters are extracted from EGG signal which is easy to capture in real-time through larngography simultaneously with speech signal. The parameters often extracted from EGG signal are F0, OQ, and SQ of close phase by which properties of different phonation types can be explained and these parameters also can be used in voice model establishment. From the perspective of physical meaning, fundamental frequency (F0) is the reciprocal of glottal period, open quotient (OQ) is the ratio of open phase over the whole glottal period, and speed quotient (SQ) is the ratio of opening phase over the closing phase. Actually, these three parameters can be extracted not only from the EGG signal, but also from the speech source signal. In Figure 1, the left is the integral of the glottal source waveform extracted from the speech signal, and the right is the EGG signal. The definitions of the three parameters in two cases are also listed in this figure.



$$\mathrm{F0} = 1/\text{glottal preiod} = 1/\mathrm{ad}$$
$$\mathrm{OQ} = \text{open phase}/\text{glottal preiod} = \mathrm{bd}/\mathrm{ad}$$
$$\mathrm{SQ} = \text{opening phase}/\text{closing phase} = \mathrm{bc}/\mathrm{cd}$$

$$\mathrm{F0} = 1/\text{glottal preiod} = 1/\mathrm{ad}$$
$$\mathrm{OQ} = \text{open phase}/\text{glottal preiod} = \mathrm{bd}/\mathrm{ad}$$
$$\mathrm{SQ} = \text{opening phase}/\text{closing phase} = \mathrm{bc}/\mathrm{cd}$$

Figure 1. Definitions of F0, OQ, SQ extracted from speech source signal (left) and EGG signal (right).

3. Phonation types in vowels

There are two sets of vowels one of which is 'lax vowel' and the other is 'tense vowel' so called linguistically in Hani language. The 'tense' and ' lax' are not as the same as those in English and they are only two sets of different vowels. From the view point of physiology, 'The muscles of larynx contract with great strength and short airflow and the vowels are loud and clear in perception, when the tense vowels are pronounced. The muscles of larynx contract with less strength and long smooth airflow and the vowels are so loud and clear.' (Li at al, 1986). Professor P. Ladefoged in UCLA once studied on these two sets of vowels and found that the ratio of airflow over air pressure of lax vowel is larger than that of tense vowel (Ladefged et al, 1987, 1978). They thought that the tense vowels are glottalized at the offset of vowels.

The tense and lax vowels had been studied acoustically by professor Kong from the viewpoint of phonation types, and found that 1) the tense and lax vowels are produced by different mechanism of larynx and they belong to different phonation types, the so called tense vowels are a kind of creaky voice and the lax vowels are modal voice (Kong, 1996); 2) the two phonation types remain unchanged through all the vowels from the onset to the offset; 3) voiced consonants have two different phonation types which are as same as those of the vowels. In addition, the muscles of larynx contract with more strength when pronounce the consonants before the creaky vowels; 4) the articulation of these two sets of vowels are almost the same by analysis the first and second formants in the tense and lax minimum pairs, and the opening of mouth in lax vowels are little bit larger than those of tense vowels; 5) the tone contours in tense and lax vowels are the same and the duration of lax syllables are longer than those of tense vowels.

The method used in this study is 'harmonic analysis' in which the ratio or difference of the power or amplitude of these two harmonics are measured. Usually the difference of the second harmonic and the first harmonic is used to describe the properties of phonations, especially in the phonetic field investigation but it also has weakness. The main problem is that when the first formant is low, such as the first formant in vowel /i/, /u/ and /y/, the frequency of first formants is very low, so it will affect the amplitude of the second harmonic and lead to wrong parameter. So a phonatician with experience will usually choose vowel /a/ or /e/, which has large first formants, to be the experimental samples. In order to solve the problem, the amplitude ratio of the second formant and the first or second harmonic is used to describe phonation types, but the frequencies of the second formants in different voices should be same or very close to each other.

4. Phonation and articulation

Different phonation and articulation can often exist together in a vowel or syllable. So in the phonetic study of such language, research should pay attention to distinguish the function of phonation and articulation. Liangshan Yi language has 5 pairs of vowels, 5 tense vowels and 5 lax vowels. The 5 lax vowels are regarded as modal voice and the 5 tense vowels are regarded as creaky voice. (Maddieson et al, 1985). Meddison also did an experiment on some of vowels in Liangshan Yi language by the method of harmonic analysis and found that the amplitude differences of the second harmonic minus the first harmonic in some tense vowels are smaller than those in lax vowels. This result is not very good to explain the nature of tense and lax vowels in Liangshan Yi language.

As is well known, there are two features in the vowel system of Liangshan Yi language, one is phonation and the other is articulation. According to linguistic study, the articulations are the same and the phonations are different in the high vowel. The articulations and phonations are all different in the middle and low vowels. Based on this, another phonetic study on Liangshan Yi language had been done and found 1) the articulation between the five lax/tense vowels are quite different, the lax vowels have a closer mouth opening than their tense counterparts; 2) in the tense vowels, the phonations at offsets are more tenser than those at onsets; 3) the difference between lax and tense vowels in Liangshan Yi language lies not only in the different tension of vocal folds, but also in the different pharyngeal cavity size and tongue position. As for tense vowels, voices are creaky, tongue root advances, and the pharyngeal cavity enlarges. Therefore the tense vowel in Liangshan Yi language is a comprehensive phonation type produced by a mechanism of vocal folds, tongue root, and laryngeal cavity; 4) in harmonic analysis, the ‘h2-h1’ is not appropriate in this case, while the ‘F1-h1’ and ‘F2-h2’ are applicable (Kong, 1997a).

5. Phonation in initials

It is important for the description of a sound system in language field investigation, if the phonation feature of an initial can be identified. There are 29 consonants in which many are voice consonants, 8 pairs of tense/lax vowels whose articulation quality are the same, and 3 tones which are simple. So they are very good materials to study on the phonation types of voiced consonants. In the study, the method of harmonic analysis is used and found 1) the phonation types are different between the 8 pairs of tense/lax vowels. The tense vowels are a kind of pressed or creaky voice and the lax vowels are modal voice; 2) voiced consonants have two different phonation types which are as same as those of the vowels; 3) the phonation types of Axi Yi language have relationship with the articulation but have no great different. 4) although the 3 tones in Axi Yi language are described as 33, 55 and 21 linguistically, they have close relationship with the phonation types of vowels. The F0 contours in creaky syllable are higher than those in modal syllable (Kong, 1997b).

According to the results, the vowels can be described as two kinds of vowels by different phonation types linguistically, and the voiced consonants also can be described as two kinds of consonants by phonation types, one is creaky voiced consonant and the other is model voice consonants. At present, linguists in China usually regarded the phonation features as the properties of vowels. From the viewpoint of historical linguistics, the initials and final stops in a syllable usually have relationship with the two phonation types.

6. Phonation in finals

The finals in most of languages in Loloish language group consist of single vowels and only few languages have voiced final endings. There are many voiced final endings appear in the language of Burmese language groups. Zaiwa language which belongs to the Burmese languages, has 5 lax vowels, 5 tense vowels, 3 tones which is simple and 68 finals among which many have voiced final endings. There are 3 kinds of final structures, which are single vowel, main vowel+stop ending and main vowel+nasal ending (Xu, 1984).

An experiment on phonation types had been done through the methods of harmonic analysis and the inverse filtering and the results shows 1) the tense vowels belong to a king of pressed voice and the lax vowels are modal voice with a little breathy voice. 2) the phonations types of different final endings are as same as those of the main vowels. 3) according to the results, the nasal endings in Zaiwa language have two different phonation types, one is pressed voice and the other is model voice. 4) finals with stop endings are always pressed voice from which we can see that phonation types in Zaiwa languae have close relationship with stop endings. 5) the phonation types of voiced consonants are as same as those of main vowels. 6) The pressed voice feature not only appear in main vowels and voiced initials, but also appear in voiced final endings and stop endings (Caodao, 1998; Kong, 2001). According to this, the final can be described as a tense final or tense syllable linguistically.

7. Phonation in tones

When phonation types of languages in Sino-Tibetan language family are discovered and studied, linguists China did want to know if phonation features can be a distinguish feature of tones. Jinpo language belonging to Jingpo languages in Burmese branch has 31 initials, 4 tones, and 88 finals. There are 4 final structures which are 1) single vowel, 2) main vowel+vowel ending, 3) main vowel+nasal ending and 4) main+stop ending. If the different phonation types appear in the four kinds of final structures simultaneously, the syllabic nature which can be used to define tone quality can be identified.

After the acoustical analysis on Jingpo language, it can be found that the tense vowel is a kind of pressed voice and the lax vowel is modal voice. They are two different phonation types. While the results also show that a phonation type which really appears in different syllables is of syllabic and can be used as a suprasegmental feature. Jingpo language is a tonal language. Usually tones are described and defined by levels and contour forms of F0 which produce by frequency of vocal fold vibration. When articulations of segments and F0 contours in one syllable are the same, the different phonation types are the distinguish features which can be defined as tone quality produced by the different vibration methods physiologically (Kong, 2001,2005). In the latter section, we can see that F0 (time domain), OQ and SQ (frequency domain) can be used to define tone quality.

8. Phonation of voiced aspiration

Another phonation type often appears in the languages of China is breathy voice which can be found in Miao and Wa languages. The breathy voice always has relationship with voiced aspiration consonants. In this section, the breathy voice in Shimenkan Miao is taken as an example to explain the phonation feature. The voiced aspiration in Miao has been found many years ago and was described as voiced aspiration consonant, aspiration vowel and tones by different linguists for different purposes (Wang, 1979). Shimenkan Miao has 56 consonants, 21 finals and phonological 8 tone. Usually the voiced consonants are described as voiceless consonants, because they are in a complementary distribution with tone 1, 3, 5 and 7.

The experiment on the voiced aspiration shows: 1) the voiced aspiration is breathy voice in Shimenkan Miao. 2) Breathy voice which is syllabic appears not only in voiced aspiration consonant, but also in the whole finals. 3) The tones in syllable of breathy voice are lower than those in un-breathy voice syllable, which indicate breathy voice can decrease F0 of tones, especially at the onset of tones (Kong, 1993). According to the results, breathy voice in Shimenkan Miao can be described as voice aspiration consonant, voiced aspiration vowel and voiced aspiration tone phonemically and phonologically for different linguistic purposes.

9. Phonations in Mandarin tones

Mandarin is tonal language which has 4 basic tones and 20 diatones including the 4 neutralized diatones. As is well known, F0 in Mandarin tones is significant in distinguish meanings and the tone perception categories have been defined by Professor William S-Y. Wang (Wang, 1983). With the development of speech technology, speech synthesis and recognition of Mandarin have been researched and implement systems have been established. In order to improve speech synthesis system, it is important to study and improve phonation model of Mandarin tones. In this section, 3 basic studies on phonation patterns or models are briefly introduce.

The patterns of tones and diatones of Mandarin were studied by EGG analysis, in which the parameters of F0, OQ and SQ were extracted from EGG signals. The results show: 1) The F0 contour of Tone 1 (Yinping) whose tone value is 55 by the 5 letter tone system by Chao (Chao, 19xx) is 'high-level', the SQ contour is also 'high-level' and the OQ contour is 'rising'. 2) The F0 contour of Tone 2 (Yangping) whose tone value is 35 is 'rising', the SQ contour is 'falling' and the OQ contour is 'rising' which is as same as that of tone 1. 3) The F0 contour of Tone 3 (Shangsheng) whose tone value is 214 is 'low-level', the SQ contour is 'high-level' and the OQ contour is 'falling-rising'. 4) The F0 contour of Tone 4 (Qusheng) whose tone value is 51 is 'falling', the SQ contour is 'rising' and the OQ contour is 'rising-falling-rising'. The same method and parameters were used in the study of the diatones in Mandarin (Kong, 1998).

The phonation model on tones and diatones in Manadrin were studied by the method of inverse filtering. Although LPC inverse filtering analysis is a good method to extract voice source from speech sound, it has weakness in obtaining voice source, because it is an all poles model. As is well know, there are zeros (anti-resonance or anti-formant) in speech sound when nasal is pronounced (Dang, J., K. Honda, et al, 1994), and zeros will also appeared in the coupling of vocal tract and trachea, and the piriformis will also produce anti-resonance (Dang, J., K. Honda, et al, 1997). In the phonetic study on phonations in China's languages, normal vowels, such as /a/ or /e/, are often used as experimental samples and vowels with lower first formants are not used, because the power of F0 will be reduced in the inverse filtering. Since this method cannot extract zeros (anti-formants) for the inverse filtering design, the inverse filtered acoustical source display single peak waveform, double peaks waveform and tri-peaks waveform, when F0 deceases gradually (Kong, 2004b). The phonation model was established through such source waveforms which can be used to improve the speech Intelligibility and naturalness of speech.

With the development of high-speed imaging, the vocal fold vibration of tones in Mandarin can been observed and captured by endo- or fibre-scope through high-speed video camera. So parameters such as F0, OQ, SQ and jitter can be extracted from dynamic glottis of tones for phonation research and modeling. The model of dynamic glottis can (Kong, 2007) can not only synthesize different phonation types such as modal voice, falsetto, vocal fry, creaky and so on, but also synthesize disordered voices such as diplophonia. To compare the 3 studies above, we can see that the research sources are not same.

The first is EGG signal, the second is speech signal and the third is image signal. Although the signals are different, they all revealed the phonation natures in tones of Mandarin.

10. Phonation study by high-speed image

In the 70s last century, the techniques of high speed imaging developed quickly. This technique was also used in the study of vibration of vocal folds. With digital image signals processing, many parameters can be extracted from the high speed image glottis (Kong, 2007). The definitions of parameters display in figure 2.

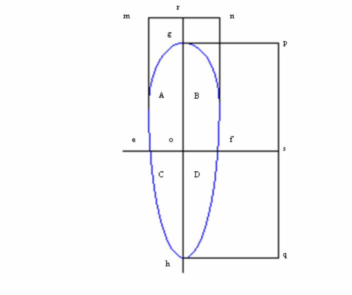


Figure 2. Parameter definition

There are two plots in the following 3 figure. In figure 3., the left displays 24 vocal fold images of modal voice and the right displays 13 parameters extracted from dynamic glottis. In figure 4., the left plot displays 24 vocal fold images of breathy voice and the right displays 13 parameters of this voice. In figure 5., the left displays 24 vocal fold images of creaky and the right displays 13 parameters of this voice.

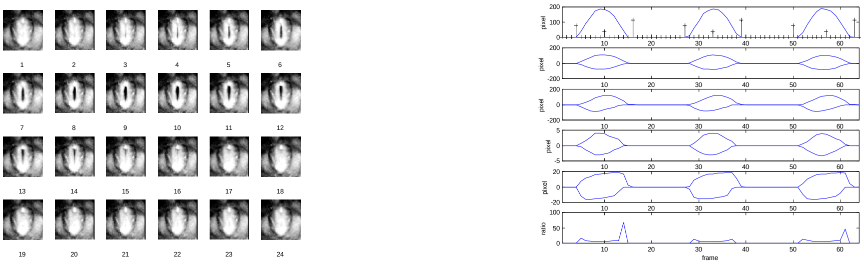


Figure 3. Left plot displays 24 frames of vocal fold images of modal voice. Right plot display 13 parameters of a model voice.

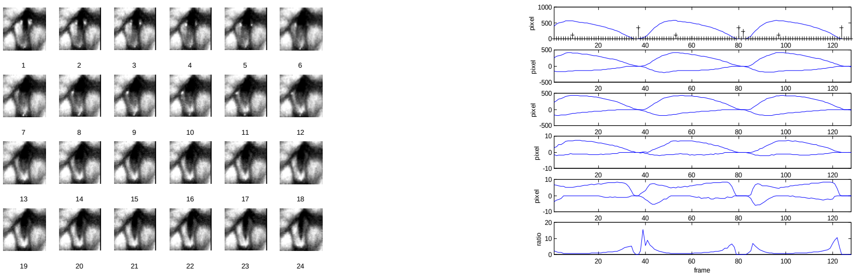


Figure 4. Left plot displays 24 frames of vocal fold images of a breathy voice. Right plot display 13 parameters of a breathy voice

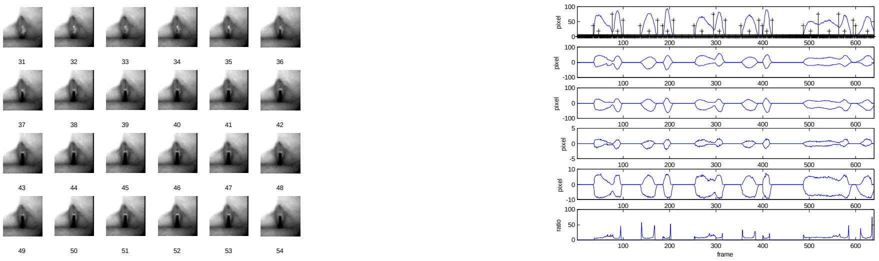


Figure 5. Left plot displays 24 frames of vocal fold images of a creaky voice. Right plot display 13 parameters of a

creaky voice.

Through the high speed images, vibration of vocal folds of different phonation types can be observed and studied both acoustically and physiologically. The research by high-speed imaging has greatly improved the research on phonation types in China.

11. Linguistic discussion

According to the principle of acoustics, speech production consists 3 parts which are speech source which is correspondent to vibration of vocal folds, resonance which is correspondent to articulation of vocal tract and radiation which does not has linguistic significance in languages. So some phonetics concepts and theoretical frame work should keep consistent with the acoustical principles. In this section, the definitions of model voice, breathy voice, pressed voice and creaky voice are given and discussed, an acoustical chart of phonation is introduced, the concept of tone quality is defined, the concepts of articulation quality and phonation quality are given and discussed and the 3 kinds of vowel qualities are defined.

As is well known, phonation changes in oral speech communication of all languages. From the viewpoint of linguistic phonetics, phonation types in languages of China should be of linguistic significance and those which do not have linguistic significance are not regarded phonation types even though they are real different physiological and acoustical phonations. According to the results and conclusions discussed above, there are at least 3 phonation types which are significant in disguising meanings. They are "modal voice", "breathy voice" and "pressed voice" or "creaky voice". The phonation type which usually has modest F0, modest OQ which is around 55% and modest SQ which is around 200 is regarded as "modal voice". The phonation type which usually has relatively small F0, large OQ and small SQ is regarded as "breathy voice". The phonation type which usually has relatively large or small F0, small OQ and large SQ is regarded as "pressed voice". The phonation type which usually has very small F0, small OQ, very large SQ and irregular period or large jitter is regarded as "creaky voice". There phonation types are defined linguistically based on acoustical parameters.

In linguistic phonetics, acoustical vowel chart is used to display the tongue position and mouth opening of a vowel. In the study on phonation types, we have developed an "acoustical phonation chart" to display status of vocal fold vibration of phonations. The parameters used in this chart which can be 2 dimensions (2D) and 3 dimensions (3D) are F0, OQ and SQ. See figure 6 and 7. Figure 6 is a 2D acoustical chart of phonation whose x axis is OQ and y axis is SQ. It displays 5 phonation types. Figure 7 is a 3D acoustical chart of phonation whose x axis is F0, y axis is SQ and z axis is OQ. It displays the phonation distributions of ordinary speech and chanting of Lama chantings in 3D space.

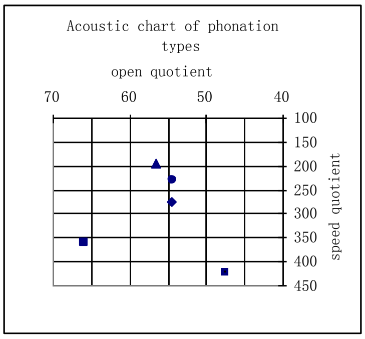


Figure 6. A 2D acoustical chart of phonation.

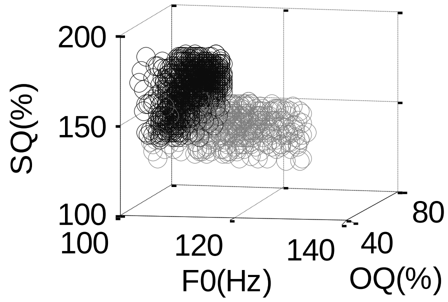


Figure 7. A 3D acoustical chart of phonation.

According to the phonation studies in languages of China, phonation types are often syllabic. If the initials, finals which may has main vowel and vowel or nasal endings and tone contours in two single syllable words which distinguish from the other only by phonation types are the same, the phonation types can be regarded as "tone quality", because the F0 contours reflect the phonation nature of tones in time domain which can be regarded as "tone pattern" and the phonation types reflect the phonation nature of tones in frequency domain.

A vowel is produced by both vibration of vocal folds (speech source) and articulation of vocal tract (resonance) and it is wrong to think that vowel is produced only by articulation. So the vowel quality is regarded as "articulation quality", when articulation is used to distinguish meanings and the vowel quality is regarded as "phonation quality", when phonation type is used to distinguish meanings. These two kinds of qualities are also can be used to describe and define voiced consonants.

Based on the concepts of articulation quality and phonation quality, vowel quality has 3 different forms which are: 1) "articulation quality is same and phonation quality is not same"; 2) "articulation quality is not same and phonation quality is same"; 3) "articulation quality and phonation quality are all not same". These 3 combinations of vowel qualities can all be used to distinguish meanings in a language.

With development of speech technology, many new methods are used in phonation research. For instance, the method of VAT (vocal attach time) is a new method which is developed in the fields of speech physiology (Baken RJ, Orlikoff RF.1998). By this method, the hard and soft voices can be identified. At present, this method is not widely used in the phonetic study for phonations. A according to our preliminary study by this method, it can be found that hard and soft voices indicate some relationship between tones and initials. We hope that it can be used to reveal the nature of tone's origin. In a word, the phonetic and linguistic studies on phonation types in languages of China are the bases which improve the research on phonation types in China's speech science.

12. Reference

[1] Maddieson, I., & Ladefoged, P. (1985). Tense'and'lax'in four minority languages of China. *UCLA Working Papers in Phonetics, 60*, 59-83.

[2] Maddieson, I., & Hess, S. (1986). Tense’and ‘lax’revisited: more on phonation type and pitch in minority languages in China. *UCLA Working Papers in Phonetics, 63*, 103-109.

[3] Cao Jianfen, Maddieson I., “An exploration of phonation types in Wu dialects of Chinese”, Journal of Phonetics, 20, 77-92, 1992.

[4] Kong Jiangping, 1998, EGG model of diatones in Mandarin. Proceedings of 5th International Conference on Spoken Language Processing. Tu5P16, Sydney, Australia.

[5] Caodao Barter and Kong Jiangping., 1998, Acoustic Study of Phonation Types of Tense and Lax Vowels of Zaiwa Language and Spectrum Method Used in Phonation Analysis. Proceedings of Conference on Phonetics of the Languages in Hong Kong, China.

[6] Kong Jiangping, 2004b, Phonation patterns of tone and diatone in Mandarin, From Traditional Phonology to Modern Speech Processing, Foreign Language Teaching and Research Press, edited by G. Fant et al, ISBN 7-5600-4075-6.

[7] Kong, J. (2007). *Dynamic glottal and physical model* Beijing: Peking University Press.

[8] Kong Jiangping, 2009, The basic methods in the study of phonations, Frontiers in Phonetics and Speech Science, edited by G. Font, H. Fujisaki & J. Shen, the Commercial Press

[9] Kong Jiangping and Edwin M.-L.Yiu，2011a, Quantitative Analysis of High-Speed Laryngoscopic Images, in Chapter 12 of Handbook of Voice Assessments, edited by Estella P.-M. Ma and Edwin M.-L Yiu. p.147-164, Plural Publishing Inc., San Diego, Oxford, Brisbane.

[10] Zhu Xiaonong, 2006, Creaky voice and the dialectal boundary between Taizhou and Wuzhou Wu. Journal of Chinese Linguistics, 2006, 34.1: 121-134.

[11] Ladefoged, P. (1973). The Features of the Larynx. *Journal of Phonetics, 1*(1), 73-83.

[12] Ladefoged, P. (1988). Discussion of phonetics: a note on some terms for phonation types *Vocal physiology: Voice Production, Mechanisms and Functions* (pp. 373-375). New York:

[13] Laver, J. (1980). The phonetic description of voice quality.

[14] Ladefoged, P., Maddieson, I., Jackson, M., & Huffman, M. (1987). Characteristics of the voice Source. Paper presented at the ECST-1987, Edinburgh.

[15] Ladefoged, P., Maddieson, I., & Jackson, M. (1988). Investigating phonation types in different languages. Vocal physiology: Voice production, mechanisms and functions, 297-317.

[16] Kirk, P. L., Ladefoged, P., & Ladefoged, J. (1984). Using a spectrograph for measures of phonation types in a natural language. UCLA Working Papers in Phonetics, 59, 102-113.

[17] Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication, 11(2-3), 109-118.

[18] Fant, G., & Gauffin, J. (1994). Speech science and technology (J. Zhang, Trans.). Beijing: the Commercial Press.

[19] Dang, J., K. Honda, et al. (1994). "Morphological and acoustical analysis of the nasal and the paranasal cavities." Journal of the Acoustical Society of America 96(4): 2088-2100.

[20] Dang, J. and K. Honda (1997). "Acoustic characteristics of the piriform fossa in models and humans." Journal of the Acoustical Society of America 101(1): 456-465.

[21] Baken RJ, Orlikoff RF. Estimating vocal fold adduction time from EGG and acoustic records. In: Schutte HK, Dejonckere P, Leezenberg H, Mondelaers B, Peters HF, eds. Programme and abstract book: 24th IALP congress, Amsterdam; 1998:15.

[22] 胡道静，1957，新校正梦溪笔谈，中华书局出版，11 月，第 1 版。

[23] 鲍怀翘、吕士楠，1992，蒙古语察哈尔话元音松紧的声学分析，民族语文，1 期。

[24] 鲍怀翘、周植志，1990，佤语浊送气声学特征分析，民族语文，2 期。

[25] 李绍尼、艾杰瑞，1990，云南剑川白语音质和音调类型-电脑语音实验报告，中央民族学院学报，第 5 期，70-74 页。

[26] 艾杰瑞、李绍尼，1997，云南剑川白语的音质和滤音实验，彝缅语研究，四川人民出版。

[27] 孔江平，1993，苗语浊送气的声学研究，民族语文，第 1 期。

[28] 孔江平，1995，汉语普通话嗓音特征相关分析，中国声学学会 1995 年青年学术会议论文集，西北工业大学出版社。

[29] Kong, J., 1996, Study on phonation types and timbre of Hani language. *Minority Languages of China*, *1*(1).

[30] 孔江平，1997a，凉山彝语松紧元音的声学研究，彝缅语研究，四川人民出版社。

[31] 孔江平，1997b，阿细彝语嗓音声学研究，中国民族语言论丛，云南民族出版社。

[32] 孔江平，1997c，中国少数民族语言发声类型研究，中国民族年鉴，Vol.1996-1997。

[33] 朱晓农，2009，弛声化：佤语中的松元音 (合作). 《民族语文》2:69-81.

[34] 朱晓农，2010，嘎裂声作为低调特征：河北省方言的声调考察（合作）. 复旦《语言研究集刊》第 7 辑 134-147.

[35] 朱晓农，2008，嘎裂化：哈尼语紧元音（合作）. 《民族语文》4:9-18

[36] 马学良，1948，倮文作祭献药供牲经译注，中央研究院历史语言研究所集刊第 20 本 579 页。

[37] 胡坦、戴庆夏，1964，哈尼语元音的松紧，中国语文，第一期，76-87 页。

[38] Kong, J. (2001a). *On Language Phonation*. Beijing: Central University For Nationalities Press. 孔江平，《论语言发声》，中央民族大学出版社，2001.

[39] Kong J.P., (2001b). Correlation and Classification Study on EGG Parameters of Mandarin *Modern Phonetics in New Century*. Beijing: Qinghua University Press.

[40] Kong Jiangping, 2003, Extracting voice source though inverse-filtering, *Language and Law*, Law Press China

[41] Kong Jiangping, 2004a, A study on the acoustic properties of phonation types and parameter synthesis, *Modern Phonetics and Phonology*, Tianjin Social Science Academic Press.

[42] Kong Jiangping, 2005, A study on tense and lax vowels in Jingpo language and the acoustical research methods of phonation types, A Research on the Sino-Tibetan Languages, Yunnan Minzu Chubanshe, Yunnan Minority Press.

[43] Li, Y., & Wang, E. (1986). Overview of Hani Language. Beijing: The Ethnic Publishing House 李永燧,王尔松(1986). 哈尼语简志. 北京: 民族出版社. .

[44] Sun, H., & Liu, L. (1986). Overview of Nu language. Beijing: The Ethnic Publishing House 孙宏开，刘璐. (1986). 怒族语言简志. 北京: 民族出版社.

[45] Xu, L., & Zhao, Y. (1984). Overview of Bai Language. Beijing: The Ethnic Publishing House 徐琳,赵衍荪(1984) 白语简志 北京: 民族出版社.

[46] Xu, L., Mu, Y., & Gai, X. (1986). Overview of Lisu Language. Beijing: The Ethnic Publishing House 徐琳,木玉璋,盖兴之 (1986) 傈僳简志 北京:民族出版社.

[47] Chang, H. (1984). Overview of Lahu Language. Beijing: The Ethnic Publishing House 常竑恩(1986) 拉祜语简志 北京: 民族出版社

[48] Wang, F. (1985). Overview of Miao Language. Beijing: The Ethnic Publishing House 王辅世. (1985). 苗语简志. 北京: 民族出版社.

[49] Xu, X., & Xu, G. (1984). Overview of Zaiwa Language. Beijing: The Ethnic Publishing House 徐悉艰，徐桂珍. (1984). 景颇族语言简志（载瓦语）. 北京: 民族出版社.

[50] Zhou, Z., & Yan, Q. (1984). Overview of Wa Language. Beijing: The Ethnic Publishing House 周植志，颜其香. (1984). 佤语简志. 北京: 民族出版社.

[51] Dao, B. (1983). Overview of Mongolian Language. Beijing: The Ethnic Publishing House 道布(1983). 蒙古语简志. 北京: 民族出版社.

[52] Chen, Z., & Lei, X. (1984). Overview of Western Yugur Language. Beijing: The Ethnic Publishing House 陈宗正，雷选春 (1984). 西部裕固语简志. 北京: 民族出版社.

[53] Xuan, D., Jin, X., & Zhao, X. (1985). Overview of Korean Language. Beijing: The Ethnic Publishing House 宣德五，金祥元，赵习. (1985). 朝鲜语简志. 北京: 民族出版社.

A study on multi-speech models of Mandarin and multi-media learning system

KONG Jiangping

Abstract

This paper introduces the multi-speech modality research of Mandarin in the department of Chinese language and literature, Peking University and discusses the role and possibility in establishing a multi-media learning system of Mandarin. In the study on speech production of Mandarin, five basic models which are 1) the model of vocal fold vibration established by high-speed digital imaging; 2) the model of dynamic vocal tract establishes through X-ray and MRI, 3) the model of lip motion set through video and motion capture, 4) the model of palatal contact studied through electropalatography and 5) the model of speech aspiration studied through an respiration belt are introduced. The advantages and possibility of these models used for Mandarin learning are discussed including the teaching and learning for people with hearing and pronouncing problems. Finally the application prospects in multi-media teaching and learning system of Mandarin are talked.

Keywords：

Mandarin, multi models of speech production, multi-media learning system1. Introduction

With the development of China, the activities of economics and cultures have been increasing rapidly. So people of the world have been paying more and more attention to the learning of Mandarin which is also an official language used in the United Nations. Mandarin belonging to the Sino-Tibetan language family is a typical tonal language and has many speech characteristics in speech physiology, acoustics and psychology. As a standard Chinese spoken language, studying on the Mandarin speech production physiologically, acoustically and psychologically is very important for its teaching and learning.

In the traditional education, people tends to regard the education as a kind of "art" but not "science" or a kind of scientific methods. At present, since the speech technology and internet develop fast, many techniques and methods are used in establishing e-teaching system which leads language teaching and learning into a new field. As to the spoken language teaching and learning, the study on multi models in speech production scientifically and technologically can improve the way of teaching and learning and set up new educational system.

The spoken language acquisition of second language has its own property which has close relationship with speech production and the modeling is well benefit from the new techniques used in recent ten years. In Peking University, the multi models of Mandarin had been studied for almost ten years and the models of vocal fold vibration, vocal tract, lip motion, palatal contact and speech aspiration were established. In this paper, the multi-speech models of Mandarin are briefly introduced and the application prospects in Mandarin teaching and learning system are discussed.

2. Model of dynamic vocal tract

X-ray was used in studying on vocal tract when it was just invented. Through X-ray, phoneticians got to know the activities of human's speech organs and defined vowels by mouth openings and tongue positions. In China, only one set of X-ray materials including 250 single syllables and disyllables of 3 persons, 1 male subject and 2 female subjects, was captured by Professor Bao

Huaiqiao in 1970s. Since X-ray is invasive, these materials are very valuable in studying the movement of articulation in Mandarin.

With the development of image signal processing, it is relative easy to process these videos and detect the edges of vocal tracts and tongues of different vowels. A database of Mandarin vocal tract was established by 32000 frames of images by which a 2D model of dynamic vocal tract in Mandarin was set up. In this model, the speech organs were divided into 6 parts which are: 1) lips; 2) mouth; 3) soft palate; 4) tongue, 5) tongue tip and 6) vocal folds and 12 parameters were used to drive the model. See Figure 1. There are 4 plots in figure 1. which displays the vocal tract of initial /b/, the vocal tract of vowel /a/, the 3D speech organs by MRI and the separated 3D speech organs.

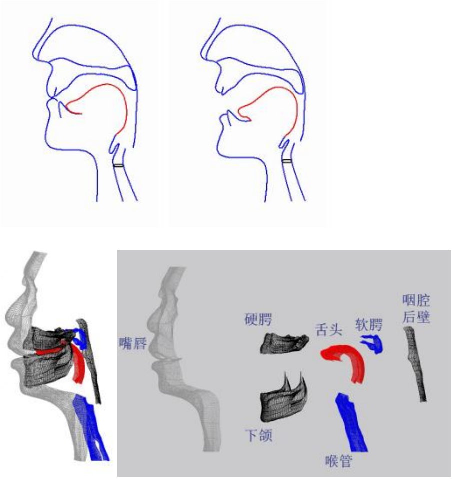


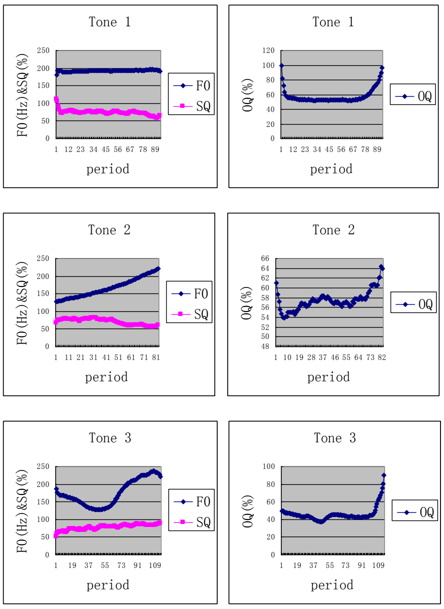
Figure 1: Vocal tracts of Mandarin captured by X-ray and MRI

In the teaching of Mandarin pronunciation, pictures of speech organs are often used in class and the disadvantage is that they are not able to show the movement of articulations. Based on the model of dynamic vocal tract, a teaching system for Mandarin pronunciation was established, in which the Mandarin articulation was carefully described by images of real vocal tracts and videos. It is a very useful and convenient tool and has plenty of materials for both teachers and second language learners. In addition, the system is also useful for Chinese children who have partially hearing or hearing problems to learn the pronunciation of Mandarin in the first language acquisition.

3. Model of vocal fold vibration

Since the vocal folds are in the throat and can't be easily observed directly, the phonetic study on phonation types developed much late than that of articulation. Since the technique of high-speed imaging was used in studying on phonation types and the vibration procedure of vocal folds can be observed by eyes, phonation types of different languages were greatly understood. The fundamental frequency (F0) and its contours were usually regarded as the only distinguish feature in Mandarin tones, and now people found that the phonations, especially the phonation in the third tone (shang sheng) was a kind of creaky voice which was significant to intelligibility and naturalness of Mandarin.

The high-speed image samples of 4 basic Mandarin tones of 3 males were captured by fiberscope and the high-speed image samples of 8 persons, 4 males and 4 females, were captured by endoscope, while the speech sound and electroglottography signal were sampled simultaneously. After signal processing, three kinds of parameters can be extracted from these 3 signals. Based on these parameters, phonation models of Mandarin tones can be established. See figure 2.. There are 2 plots for each tone in figure 2 in which the upper display the parameters of F0 and speed quotient (SQ) which is defined as opening phase over closing phase and the lower displays the parameter of open quotient (OQ) which is defined as open phase over period.



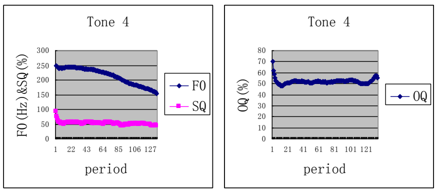


Figure 2: Phonation parameters of 4 basic tones in Mandarin.

From the phonation patterns of tones, we found that OQ and SQ changed along with F0. In the learning of Mandarin tones, all attention had been paid to the change of pitch and the phonation had been disregardful. Now we know that phonations are very important in learning Mandarin tones, especially the third tone in which creaky voice is the most important thing that a learner has to imitate. By taking advantage of modern techniques, visual phonation feedback should be designed in a Mandarin learning system to help the learners. With the visual feedback of F0, OQ and SQ, phonations, it is not only helpful for the normal Mandarin learner to learn tones, but also very useful and helpful for the partially hearing people to learn tones. In addition, phonation type visual feedback in a learning system will also be helpful for singers to imitate different phonation types in different operas or original folk songs.

4. Model of lip motion

From the viewpoint of speech communication, lip reading or speech reading is very important in learning of spoken language and lip reading is obliged to the language teaching of deaf mute child. From the viewpoint of linguistics, viseme is defined as a unit which is significant in distinguish meanings. The results of present research on viseme show that static viseme was not very useful in spoken language learning and the dynamic viseme was more useful and significant in language learning. So dynamic lip reading should be designed in the spoken language teaching system which needs multi-media technique.

Mandarin lip motion has been studied through video materials and image signal processing. In this study, video of 4000 basic Chinese words were captured by video camera and motion capture and the contours of inner and outer lips were detected for setting up a database. With this parameter database, a model of lip motion was established which could be used in a Mandarin teaching and learning system. See Figure 3.. There are 3 plots in figure 3. in which the first plot displays a lip image with detected contours, the second plot displays the definition of the model, and the third plot displays the 3D parameters sampled by motion capture.

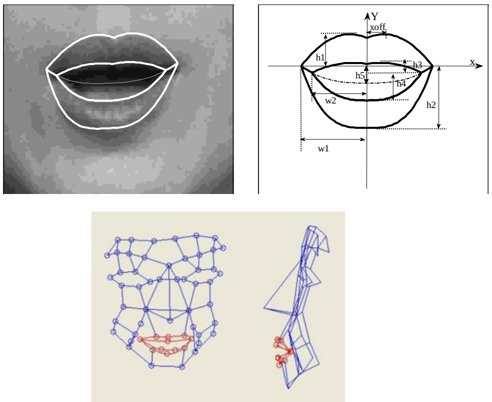
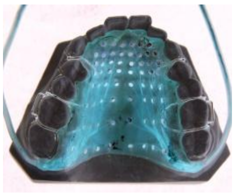


Figure 3: Lip parameters and definition of Mandarin.

According to our study, the dynamic viseme produced by the model of lip contours would lose information of speech in communication. So a 3D model of lips should be studied and established for a multi-media Mandarin teaching and learning system which is not only necessary for normal language learners, but also for the learners who need lip reading in their work.

5. Palatal contact by EPG

One of the methods in the study of Mandarin articulation is to capture the signal of palatal contact area by EPG, an instrument often used in the research of physiology. According our research, we found that it was a very good method to study on the articulation of Mandarin, especially the co-articulation between consonants and vowels in one syllable and the co-articulations in running speech. However, this instrument was originally designed for the training of child with cleft palate after medical operation and an artificial palate whose cost was not low should be made for each user beforehand. All this impedes the application of EGP in normal language training.



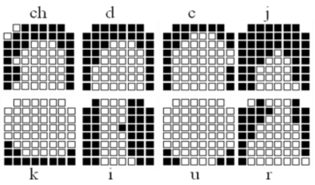


Figure 4: An artificial palate and the parameters extracted by EPG in Mandarin.

A database of EPG in Mandarin was established in the phonetic lab in Peking university and the co-articulations within syllables and between syllables were studied. According to the research results, the language training on cleft child is necessary after medical operations and the method of EPG can also be used in the training of normal learners of Mandarin, if the language training system with EPG is low.

6. Model of speech respiration

Respiration is the power of speech and was rarely studied phonetically because there was no appropriate instrument. In recent 10 years, an respiration belt which was an instrument in the study of human physiology, was used in the phonetic study on reading aloud, sutra chanting and oral cultures in China. Some results show: 1) people usually would use both costal respiration and abdominal respiration in different kinds of speech communication and oral culture performance; the abdominal respiration would mainly provide power of speech and the costal respiration had more close relationship with articulation; 3) different patterns of respiration were used in different languages because of their different syntax structures.

In the study of respiration in Mandarin, two respiration belts were used to capture the signals of costal respiration and abdominal respiration. We found that there were at least 3 kinds of respiration resets in reading Chinese seven quatrains (七言绝句，七言律诗) and at least 4 respiration resets in broadcasting news. See figure 5.

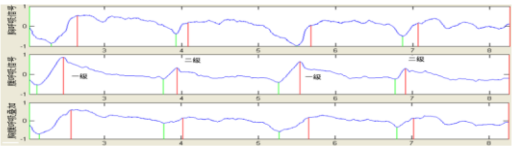


Figure 5: Signals of costal respiration and abdominal respiration in reading a Chinese seven quatrains.

The resets of 3 respiration signals were labeled in figure 5. The upper plot displays the signal of costal respiration and abdominal respiration, the middle plot displays the signal of abdominal respiration and the lower plot displays the signal of costal respiration. Based on a database of Mandarin respiration, a respiration model of Mandarin was preliminary set up to imitate the activities of speech communication. From the viewpoint of language learning, the respiration belt can be used in a Mandarin teaching and learning system for the training of reading aloud, broadcasting news, narrating story, sutra chanting, oral culture performance and singing opera and songs.

7. Application prospects of multi-media teaching system

It is a new task to put basic phonetic results and models into use of language teaching and learning system. From the viewpoint of basic phonetic study, there are two aspects which we should consider in a multi-media spoken Mandarin system, one is demonstration and the other is visual feedback. As to the first aspect, acoustical and physiological parameters or models can display speech sound in different forms. The acoustical parameters of tones and diatones can be displayed by F0, dynamic phonation types by SQ and OQ, vowels by F1 and F2 in an acoustical chart, consonants by the parameters of palate contact area and respiration rhythm by signals of respiration. The physiological demonstrations can be implemented by the models of vocal tract driven by parameters of X-ray and MRI videos, by the model of vocal folds driven by parameters of high-speed images, by the model of dynamic lip driven by parameters of videos and the by the model of respiration driven by parameters of the signals of respiration. As to the second aspect, the visual feedback of speech sound, especially the visual feedback of speech physiology is very indispensable in a multi-media language system. Physiological visual feedback of a learner's speech sound can be regarded a real-time test by which learners can judge himself in speech sound learning. At present, the implementation of speech sound visual feedback still has many technical problems in speech signal processing and physiological signal processing. However, we can display the acoustic speech feedback of tones by F0 or the feedback of vowels by F1 and F2 in acoustic vowel chart. But it is still difficult to display dynamic vocal tract or vibration of vocal folds of speech sounds pronounced by Mandarin learners.

8. References

[1] Kong Jiangping. (2004). Phonation Patterns of Tone and Diatone in Mandarin, From Traditional Phonology to Modern Speech Processing. Foreign Language Teaching and Research Press. edited by G. Fant et al, ISBN 7-5600-4075-6.

[2] Kong jiangping. (2007). Laryngeal Dynamics and Physiological Model. Publishing house of Peking University

[3] Gaowu WANG. Tatsuya KITAMURA. Xugang LU. Jianwu DANG. Jiangping KONG. (2008). MRI-based Study on Morphological and Acoustic Properties of Mandarin Sustained Vowels. Signal Processing

[4] Pan Xiaosheng. Kong Jiangping. (2008). Lip contour extraction based on support vector machine. Proceedings of the 2008 International Congress on Image and Signal Processing. Sanya. China

[5] Li Yonghong. Kong Jiangping. Wang Gaowu. Ding Lijuan. Based on X-ray Mandarin Speech Physiological-Learning System. (2011). International Conference on Computer, Electrical, and Systems Sciences, and Engineering. 2011.4.ISBN:978-0-615-42292-3/pp.412-415

[6] 李英浩，孔江平，2011，汉语普通话 V1#C2V2 音节间逆向协同发音，《清华大学学报》自然科学版，No.9, Vol.51, p.1220-1225

[7] 孔江平，2010，普通话语音多模态研究与多媒体教学，第四届全国普通话培训测试学术研讨会论文集，国家语言文字工作委员会培训中心编，2010 年 12 月，语文出版社。

[8] 汪高武. 鲍怀翘. 孔江平. (2008). 从声道截面积推导普通话元音共振峰. 中国语音学报. 第一辑, 商务印书馆

[9] 谭晶晶. 李永宏. 孔江平. (2008). 不同文体朗读时的呼吸重置特点. 清华大学学报. 第四期. 自然科学版.

Prosodic Boundary Effects on Segment Articulation and V-TO-V Coarticulation in Standard Chinese

LI Yinghao, KONG Jiangping

ABSTRACT

This paper investigates the prosodic conditioning of the segment production and vowel-to-vowel coarticulation in the Standard Chinese through electropalatographic and acoustic analysis. The articulatory and acoustic measures were obtained for un-aspirated alveolar consonant /t/ and vowels /i/ and /a/. Results show that the domain-initial consonant is strengthened in a hierarchical manner in higher prosodic domains. The vocalic gesture after the strengthened consonant tends to be reduced, and the domain-final vocalic gesture shows more linguapalatal contact and longer duration. The vocalic anticipatory effect is shown to appear up to intermediate phrase boundary, but is constrained by the articulatory constraint for vocalic gesture. The vocalic carryover effect is likely constrained by foot domain.

Keywords:
electropalatography, prosodic structure, Standard Chinese, vowel-to-vowel coarticulation

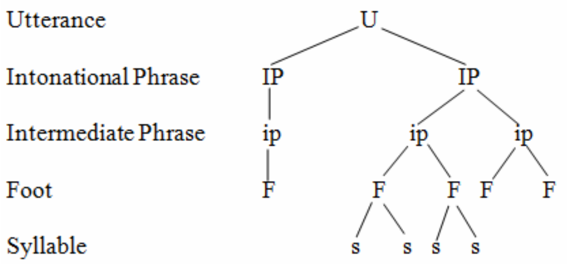
1. INTRODUCTION

Previous articulatory and acoustic studies have shown that segment articulation is modulated by prosodic structure in a hierarchical fashion in languages [1, 2, 3, 4, 9]. Segment articulation at the prosodic boundary edges is different from that at the medial position of prosodic domain, and the boundary effect on segment articulation is cumu-lative along the prosodic hierarchy. The strength of prosodic boundary is also shown to harnesses the vowel-to-vowel coarticulation in that the vocalic effect is more salient across lower vs. higher prosodic boundaries [3, 10].

In the Standard Chinese (SC), the pitch contour and pause are the main apparatus for marking the prosodic boundaries, but the foot boundary rely less on the two acoustic cues. In [13] it is hypothesized that the bisyllabic foot is phonetically characterized by the closer gestural overlap and tonal coarticulation within the foot than across foot domain. But the tone sandhi does not strictly comply with prosodic domain in the SC [11]. It is thus hypothesized that the articulatory and acoustic correlates for foot boundary might be related to the consonantality of initial consonant in that it is more consonant-like at the foot-initial than at syllable-initial position within foot domain [13].

The paper aims at finding out the prosodic signatures in articulatory and acoustic properties of segmental articulation in the SC with a special interest in comparing the foot-internal and foot-boundary segment articulation. The boundary constraint on vocalic anticipatory and carryover effects is also discussed.

Figure 1: A five-level prosodic hierarchy for the SC.



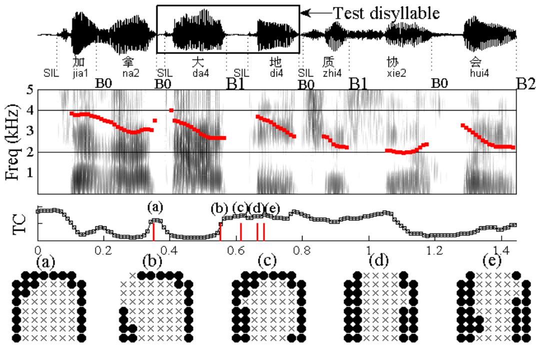
2. METHOD

2.1 Speech material

A five-level prosodic hierarchy in [8] with the prosodic word level being replaced by foot was tentatively adopted and shown in Figure 1. The test segments for articulatory strengthening experiment were unaspirated alveolar stop /t/ and the vowel /i/. In the experiment of boundary constraint for vocalic coarticulation, additional low vowel /a/ was used to construct four bisyllable sequences, /ta#ti/, /ta#ta/, /ti#ta/, and /ti#ti/ (# stands for morpheme boundary).

Five boundary conditions were placed either before the first or the second syllable, and a total of 32 utterances were designed. The syllable was always ended with /a/ preceding the test disyllables while the segment following the disyllables was not controlled. The tonal condition for the disyllable was not controlled either. A part of example utterance was shown in Figure 2.

Figure 2: A part of an annotated utterance (From top to bottom are speech signal with three tiers of prosodic annotation, spectrogram superimposed with pitch contour derived from EGG signal, total contact profile, and five EPG frames).



2.2 Recording

The recording was taken in a sound-attenuated booth. Each sentence was repeated for three times, and the 96 sentences were divided into six blocks in random with two dummy sentences placed in the first and last position in each block. One female speaker participated in the experiment and was instructed to read the sentences at normal speed. The electropalatographic signal was recorded by 62-electron electropalatography (100Hz), and the speech signal was recorded at a sampling rate of 22 kHz.

2.3 Measurements

The maximum linguapalatal contact (MaxC) of the alveolar closure was measured for /t/ (Frames (a) and (c) in Figure 2). The alveolar seal duration (ASD) was defined as the interval from the first to last frame that showed alveolar closure. The acoustic silent duration (AD) was the interval from the last pitch pulse to the energy burst for stop. The maximum linguapalatal contact (MaxV) for /i/ was taken between the one third and one half intervals into the vowel (Frame (e) in Figure 2). The center of gravity (CoG) of the linguapalatal contact was defined following [7], and the centrality index (CC) was defined following [5]. The vocalic duration (VD) and F1/F2 near the MaxV-matched time point were measured for both vowels.

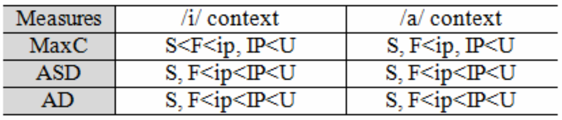
For the second experiment, the linguapalatal contact of the posterior four rows was measured respectively in the final frame over V1 interval (POS_E) and the first frame over V2 interval (POS_S)(Frame (b) and (d) in Figure 2). The V1-end F2 (F2_E) and V2-start F2 (F2_S) were first derived from 20-order covariance LPC in PRAAT, and manually adjusted in the EPG analysis platform developed on Matlab.

3. RESULTS

3.1 Domain-initial consonant strengthening

A series of one-way ANOVA were conducted for three measures of the test consonant, and significant differences were yielded at 0.0001 level. Table 1 showed the summary of the LSD post hoc multiple comparisons for the three measures for /t/.

Table 1: Summary of LSD post hoc comparison results for /t/ (p<0.05).



Measures	/i/ context	/a/ context
MaxC	S<F<ip, IP<U	S, F<ip, IP<U
ASD	S, F<ip<IP<U	S, F<ip<IP<U
AD	S, F<ip<IP<U	S, F<ip<IP<U

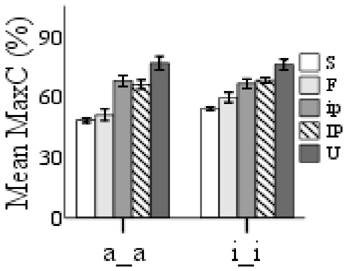
Figure 3 shows the MaxC for /t/ at five domain-initial positions in two symmetrical flanking vowel contexts. In both contexts the speaker successfully distinguished the utterance from the immediate lower prosodic domain, namely intonational phrase. However, the difference between the IP and the ip was not significant across vocalic contexts, with ip-initial position appears to have more contact than IP-initial position in the /a/ context. This might be attributed to the punctuation effect, for the speaker made purposeful pauses when encountering comma. The MaxC for lowest two prosodic domains, syllable and foot, were sig-nificantly lower than that for higher domains, but it only differed in the symmetrical /i/ vs. /a/ context. A careful examination looking into the frame tokens for maximum linguapalatal contact frames in the /i/ context found that 3 out of 12 tokens lacking alveolar closure at syllable-initial position, which was reflective of increased vocalic effect on the production of tongue tip closure gesture in a short interval of time.

The post hoc results indicated that the alveolar closure duration (ASD) and acoustic silent interval (AD) were progressively longer above foot domains across both vocalic contexts, with those measures not distinguished below foot domains. However, the result of additional independent *t*-tests showed that the syllable-initial ASD was significant less than foot-initial ASD in both vocalic contexts (p<0.05), showing a genuine cumulative effect of prosodic boundaries.

The contextual effect on two articulatory measures was also compared on domain-matched basis through independent *t*-tests. The results showed that syllable-initial /t/ had more contact in the /i/ vs. /a/ context, and the seal duration was shorter in the /i/ vs. /a/ context at both syllable and foot boundaries. This may explain the distinction between syllable and foot domains in the /i/ context in that the consonant gesture is reduced in lower domain because of the vocalic gesture influence as the articulatory interval for consonant gesture becomes shorter.

The post hoc results for the AD patterned with those for the ASD, but no difference was found between syllable and foot domains across vocalic contexts through independent *t*-tests. On the one hand, this result indicates that acoustic silent interval is an effective device to mark higher prosodic domain, on the other hand it fails to distinguish lower domains.

Figure 3: Mean MaxC of /t/ at five domain-initial positions. Error bars refer to one standard error.



The above results indicate that the prosodic structure modulates the domain-initial consonant production in a hierarchically cumulative manner. The prosodic domains above foot are well distinguished acoustically, but the distinction between syllable and foot boundaries relies more on the articulatory measures. The articulatory reduction for consonant at the syllable-initial position tends to show a combined effect of flanking vowel and time interval.

3.2 Boundary effect on high front vowel

The articulatory and acoustic measures for /i/ in the syllable /ti/ marked with both left and right boun-daries were submitted to the one-way ANOVA analysis, and the LSD post hoc results were shown in Table 2.

A systematic articulatory variation for the pro-duction of /i/ was found depending on the boundary strength for either left or right boundary. When the domain-initial consonant was prog-ressively strengthened, the gestural magnitude for /i/ tended to be reduced accordingly, with reduced vocalic duration and centralized tendency in F1/F2 space. However, when the boundary condition immediately after /i/ ascended along the prosodic hierarchy, the production of /i/ tended to be more target-like. But the result for vocalic duration only partially supported the possible strengthening of the vocalic gesture, for U-final vocalic duration was significant shorter than IP- or ip-final one. This might the result of the segmentation procedure when trailing portion of /i/ at U-final position was truncated because of irregularity of vocalic pulses. No significant difference was found for F1/F2, but the average F1 was lower and average F2 higher in higher prosodic domain.

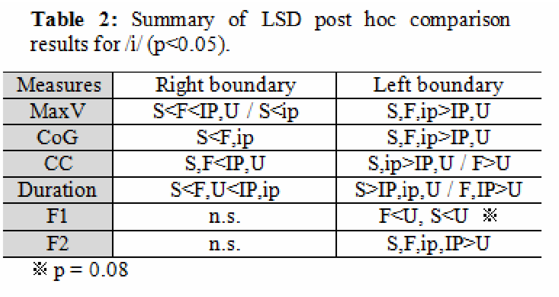


Table 2: Summary of LSD post hoc comparison results for /i/ (p<0.05).

Measures	Right boundary	Left boundary
MaxV	S<F<IP,U / S<ip	S,F,ip>IP,U
CoG	S<F,ip	S,F,ip>IP,U
CC	S,F<IP,U	S,ip>IP,U / F>U
Duration	S<F,U<IP,ip	S>IP,ip,U / F,IP>U
F1	n.s.	F<U, S<U ※
F2	n.s.	S,F,ip,IP>U

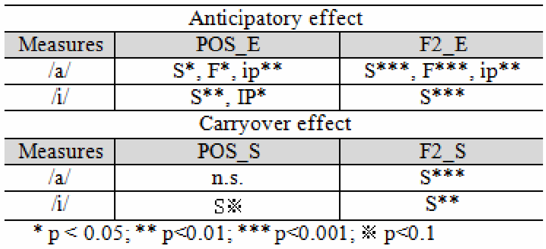
※ p = 0.08

The above results show an interesting interaction of domain-initial consonant streng-thening and domain-final vocalic strengthening in a syllable. On the one hand the strengthened consonantal articulation suppresses the vocalic influence in higher domain, on the other hand vocalic gesture immediately after the consonant tends to be reduced as the preceding consonant is strengthened. The vocalic strengthening is con-ditioned by the boundary condition imme-diately after the vowel.

3.3 Vowel-to-vowel coarticulation

The hypothesized closer gestural overlap in lower prosodic domain implies that vocalic coarticulation is more salient than in higher domains. The vocalic anticipatory effect was investigated by comparing the boundary-matched POS_E and F2_E when V1 was held constant while V2 varied. The vocalic carryover effect was investigated by comparing the boundary-matched POS_S and F2_S when V2 was held constant while V1 varied. Table 3 showed the results of the independent *t*-tests. The V1-end linguapalatal contact and F2 showed significant vocalic anticipatory effect on domains below ip for V1 /a/. The same effect was found for V1 /i/ at Syllable domain, but a significant difference also propped up at IP domain. The latter may be attributed to the tone condition, for the third tone (T3) was applied for the first syllable in /ti (B3) ta/ whereas the first tone (T1) for that in /ti (B3) ti/, with the longer duration under T3 vs. T1 condition facilitating the larger linguapalatal contact.

Table 3: Independent *t*-test results for vocalic anticipatory and carryover effects.



	Anticipatory effect	
Measures	POS_E	F2_E
/a/	S*, F*, ip**	S***, F***, ip**
/i/	S**, IP*	S***
	Carryover effect	
Measures	POS_S	F2_S
/a/	n.s.	S***
/i/	S※	S**

* $p < 0.05$; ** $p<0.01$; *** $p<0.001$; ※ $p<0.1$

The V2-start linguapalatal contact and F2 showed significant vocalic carryover effect for V2 /i/, but significant difference was only found for F2 for V2 /a/. Considering that carryover effect is generally not salient in the SC, it is safe to say here that carryover effect is constrained by the foot domain.

The above results indicate that the vocalic anticipatory effect is conditioned by the boundary strength on the one hand, and the articulatory constraint for vowels on the other. By careful examination of all tokens, it is found that timing of the initiation of the V2 gesture relative of the V1 gesture is responsible for the all-or-none vocalic anticipatory effect. A second factor involves the specific articulatory constraint for vowels which is reminiscent of the model proposed in [12].

4. CONCLUSION

The results of the first experiment indicate that the prosodic structure shows a hierarchical effect on the consonant articulation in the SC with cumulative increase in linguapalatal contact and seal duration for stop consonant in higher domain-initial positions. It also supports the hypothesis in [13] that the foot and syllable boundaries can be distinguished by the articulatory strength for consonant production instead of the acoustic silent duration. The vocalic gesture is shown to be affected by the preceding consonant production and the boundary condition immediately on its right. When the initial consonant is progressively strengthened the following vowel tends to be progressively reduced, but this is only a tentative conclusion because only one speaker’s data is used in the current study. The final-lengthening for vowel is accompanied by the strengthening of the vocalic gesture, which also supports the previous studies. The vocalic anticipatory effect is prominent up to the ip boundary but is constrained by the articulatory constraints for the vocalic gesture. However, the vocalic carryover effect is likely to be constrained in foot domain.

5. ACKNOWLEDGEMENTS

This research was supported by China Social Sciences Funds of the Ministry of Education (Grant 10JJD740007) and China National Natural Science Funds (Grant 61073085). We thank two anonymous reviewers for the comments on the earlier version of the paper.

6. REFERENCES

[1] Byrd, D., Kaun, A., Narayanan, S., Salzman, E. 2000. Phrasal signatures in articulation. In Broe, J.B., Pierrehumbert, J.B. (eds.), *Papers in Laboratory Phonology, Vol. V: Acquisition and the Lexicon*. Cambridge: Cambridge University Press, 70-87.

[2] Cho, T., Keating, P. 2001. Articulatory and acoustic studies of domain-initial strengthening in Korean. *J.Phon* 29, 155-190.

[3] Cho, T. 2004. Prosodically conditioned strengthening and vowel-to-vowel

coarticulation in English. *J.Phon* 32, 141-176.

[4] Cho, T., Keating, P. 2009. Effects of initial position versus prominence in English. *J.Phon* 37, 466-485.

[5] Fontdevila, J., Pallares, M.D., Recasens, D. 1994. The contact index method of electropalatographic data reduction. *J.Phon* 22, 141-154.

[6] Fougeron, C. 2001. Articulatory properties of initial segments in several prosodic constituents in French. *J.Phon* 29, 109-135.

[7] Hardcastle, W.J., Gibbon, F.E., Nicolaidis, K. 1991. EPG data reduction methods and their implications for studies of lingual coarticulation. *J.Phon* 19, 251-266.

[8] Li, A.J. 2002.Prosodic analysis on conversations in Standard Chinese. *Zhong Guo Yu Wen* 291, 525-535.

[9] Keating, P., Cho, Fougeron, Hsu. 2003. Domain-initial strengthening in four languages. In: Local, J., Ogden, R., Temple, R. (eds.), *Papers in Laboratory Phonology, Vol. 6: Phonetic Interpretation*. Cambridge: Cambridge University Press, 145-163.

[10] Kondo, Y. 2006. Within-word prosodic constraint on coarticulation in Japanese. *Language and Speech* 49, 393-416.

[11] Kuang, J.J., Wang, H.J. 2006. An acoustic study on the third tone sandhi across various prosodic boundaries. *Proc. 7th Phonetic Conference of China, Beijing*.

[12] Recasens, D, Pallares, M.D., Fontdevila, J. 1997. A model of lingual coarticulation based on articulatory constraints. *J. Acoust. Soc. Am.* 102, 544-561.

[13] Wang, H.J. 2008. *Non-linear Phonology of the Standard Chinese*. Beijing: Beijing University Press.

An Electropalatographic and Acoustic Study of the Tonal Effects on Vowel Production in Standard Chinese: A Pilot Study

Li Yinghao, Kong Jiangpin

Abstract

This paper carried out a pilot study for the tonal effects on the production of monophthongal vowels in Standard Chinese. The electropalatographic (EPG) and acoustic signals for six vowels /i, u, y, i1, i2, ɤ/ in monosyllables uttered in four tones were recorded, and the articulatory and acoustic measures were taken at three time points in zero-initial syllables and mid-portion frame/point in non-zero-initial syllables. The results show: (1) the articulatory variation in vowel production was mainly induced by the low tone (T3) compared with other tone. (2) Two strategies for articulatory adjustment were identified: the tongue lowering and retraction was found for vowels in T3 syllables when the post-dorsum of tongue is involved in vowel articulation, as shown in /u, ɤ, i1/; the increased tongue raising was found in T3 syllables when the pre-dorsum of tongue is involved in vowel production, as shown in /i, y, i2/. (3) The T3 influence on the articulatory measures for vowels tends to be present in the better part of the vocalic length, regardless of the pitch changes. (4) Consistent tonal effect on F1 for the vowels was found for which T4 tends to increase the F1. This might be attributed to the larynx raising that leads to the shortening of the vocal tract.

Index Terms: tone, vowel production, electropalato-graphy (EPG), Standard Chinese

1. Introduction

The tone-vowel interaction is manifested by the bi-directional influences between supra-laryngeal and laryngeal articulators in speech production [1]. On the one hand, the supra-laryngeal articulatory gestures influence the laryngeal gestures, exemp-lified by the intrinsic F0 of vowels, for which the F0 is positively correlated with the tongue body height of vowels [2, 3]. On the other hand, the production of tone does affect the supra-laryngeal gestures of vowels. In an acoustical study, Zee [4] found that vowels (except /a/) had higher F1 in high-falling tone in Taiwan Mandarin. Recent EMMA studies found that the tonal effect on supra-laryngeal articulators was not consistent compared with the robust effect of intrinsic F0 [5]; the tongue body for /a/ and /u/ tended to be retracted and lowered in low tone (T3) in Standard Chinese [6,7]. In Ningbo Mandarin [1] similar tonal effect was found for low vowel versus high front vowel in low short tone and low rising tone. The current paper aims at investigating the tonal effects on the vowel production in Standard Chinese.

The articulatory mechanism of tonal effect of tongue gestures for vowels was attributed to the larynx lowering [1, 6, 7, 8]. It is conjectured that the retraction and lowering of tongue body when producing low/back vowels in the T3 tone might be related with the downward vertical movement for the larynx, for which the contraction of related muscles drag the hyoid bone downward, and the muscles controlling the tongue configuration and connected to the hyoid bone are affected. This mechanism might explain the tongue retraction and lowering for low/back vowels, but for high front vowel /i/ or /y/ the situation might be different because the articulatory undershoot would occur if the tongue were retracted and/or lowered for the high front vowels. Thus it is still not clear the mechanism of tonal effects on the articulatory and acoustical properties of vowels in Standard Chinese.

Recent X-ray study shows that in Standard Chinese the F0 contour in four tones is closely correlated with the vertical movement of the larynx. Except the high-level tone (T1), the vertical larynx movement is positively correlated with the F0 [9]. In the Figure 3 in [9] the larynx is continually and monotonously rising in producing high-level tone, while in other three tones the trajectories of larynx vertical movement are rather similar with the corresponding F0 contours. The larynx vertical movement trajectory for the rising tone (T2) is slight different from that for T1 in that a slight downward movement of the larynx is found for some samples, which indicates that the low tonal component in T2 might be physiologically realized as larynx lowering. These results shows that fine adjustments of larynx vertical positions are involved in producing tones in Standard Chinese with the four tones being associated with contrasting laryngeal movement patterns that contribute the phonemic contrasts among four tones.

The voluntary and controlled larynx vertical movement does affect the articulatory and acoustic properties of vowels in Standard Chinese. The relevant articulatory studies have mentioned above. Acoustically, the vertical movement of the larynx would result in the variation of vocal tract length, which would affect the formant of vowels, as is shown in [4]. However, the combined articulatory and acoustic study on the tonal effects on vowel production in Standard Chinese is rare, and the existing results do not cover as many vowels in Standard Chinese. So the larynx lowering mechanism might not be able to explain the tonal effects on different vowel production.

In this paper, the electropalatography (EPG) was used to record the dynamic process of tongue-palate contact. On the one hand, such device has been rarely utilized in the tonal effects on vowel production, and EPG is more superior than EMA in reflecting the magnitude of tongue gesture, provided the linguopalatal contact can be captured during vowel production.

2. Method

The 62-electrode EPG system was used to record the linguopalatal contact for every 10ms, by which the tongue-palate contact is captured to reflect the tongue gestures for vowels in anterior-posterior dimension, reflected by the distribution of contact electrodes on the pseudo-palate, and high-low dimensions, reflected by the amount of contacted electrodes when the vowel is held constant.

2.1 Stimuli

The test vowels were /i, y, u, ɤ, i1, i2/ (/i1, i2/ were apical vowels, and their tongue configurations were respectively similar with the apical-alveolar and retroflex consonants). The above vowels were selected because meaningful linguopalatal contact can be registered for investigating the tonal effects on vowel production. Figure 1 shows the example of the tongue-palate contact for the six vowels in mid-portion location in the syllables. The vowel /a/ was not selected because no or little linguopalatal contact can be found for this vowel, regardless of the consonantal environment.

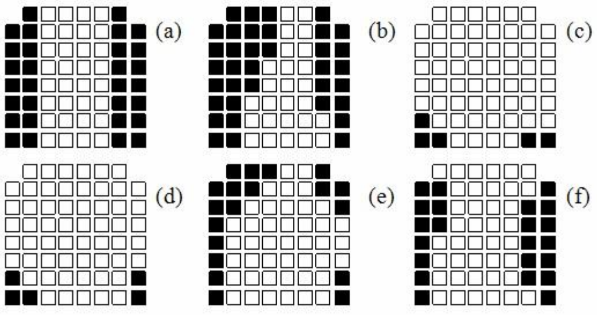


Figure 1: *The linguopalatal contact patterns for the six vowels /i, y, u, ɤ, i1, i2/ (From (a) to(f)).*

Table 1 shows the monosyllable stimuli in the current study. The Chinese monosyllables that have four tones, namely, high-level (T1), rising (T2), low (T3), and falling (T4), or at least have three tones including T3, were selected as the stimuli. For examples, the syllable /ti/ has four tones: /ti 55/ (high-level tone, meaning "to kick"), /ti35/ (rising tone, meaning "to lift"), /ti214/ (low tone, meaning "body"), /ti51/ (falling tone, meaning "to take the place of"). In total, 42 syllables were selected. Among those 36 syllables have four tones, which corresponded to 144 existing morphemes in Standard Chinese. Six syllables have three tones. Among those, two syllables (/mu/ and /ly/) do not have T1 tone, and four syllables (/ts□/, /s□/, /ku/, and /k^{h}u/) do not have T2 tone. In Table 1 the phonotactic combinations were possible in some cells, but the tone type is either less than three or lacking in T3. Therefore no syllables in those cells were selected. Some empty cells in Table 1 show that there is no possible syllable because of the phonotactic constraints in the Standard Chinese. For example, /i/ is not allowed to appear after the initials like apical, retroflex, and velar sounds. The inclusion of the monosyllables with zero initial was oriented to examine the tonal effects on the vowel production without the coarticulatory influence of the preceding initials.

2.2 Recording and speaker

The recording was carried out at the Phonetics Lab of Peking University. One female speaker, who was 27 years and was the announcer at the Peking University TV Station, was recruited in the experiment. A custom-made 62-electrode pseudo-palate was installed onto the speaker's palate, and the speaker was instructed to practice to pronounce properly for at least half an hour. Then the speaker was instructed to read aloud the Chinese characters in blocks on the computer screen. Each block contained 6 to 8 characters with a dummy token /ta51 tu51/ at the end of the block for the temporal alignment between the EPG and speech signals [10]. The stimuli were designed in random order and each character was read at least by two times. The speaker was instructed to read the stimuli list by using normal speech rate with the inter-character interval of at least 500ms. The unnatural production for stimuli items was marked by the experimenter for elimination in the following analysis, and re-grouped in blocks and recorded when all blocks were finished. The EPG, speech and electroglottographic (EGG) signals were simultaneously recorded. The sampling rate for EPG signal was 100 Hz, and for speech and EGG signals 22050 Hz.

Table 1. *Monosyllable Stimuli.(Six consonantal places of articulation for the initials were considered regarding the phonotactic constraints and the availability of Chinese monosyllables with at least three tones (and including T3). Monosyllables with zero initials were also included. LA stands for bilabial (labial-dental), AP for apical, AL for alveolar, RE for retroflex, PA for alveolo-palatal, and VE for velar. The syllables with an asterisk on the right have no T1 tone; the underlined syllables have no T2 tone.)*

Vowel	Monosyllables uttered by at least three tones						
	LA	AP	AL	RE	PA	VE	Zero
i	/pi/ /pʰi/ /mi/		/ti/ /tʰi/ /li/ /ni/		/tɕi/ /tɕʰi/ /ɕi/		/i/
y			/ly/*		/tɕy/ /tɕʰy/ /ɕy/		/y/
u	/pu/ /pʰu/ /fu/ /mu/ *		/tu/ /tʰu/ /lu/	/tʂu/ /tʂʰu/ /ʂu/		/ku/ /kʰu/ /xu/	/u/
ɤ				/tʂɤ/ /tʂʰɤ/ /ʂɤ/		/kɤ/ /kʰɤ/	/ɤ/
i1		/tsɿ/ /tsʰɿ/ /sɿ/					
i2				/tʂʅ/ /tʂʰʅ/ /ʂʅ/			

2.3 EPG and acoustic measures

The signal processing was conducted at the *EPGAnalyzer*, a Matlab-based EPG analysis platform developed by the Phonetics Lab of Peking University. The acoustic boundary of vowels was first demarcated by hand. The EPG frame with the largest linguopalatal contact (MaxFrame) was selected for further analysis by the computer program from the five consecutive EPG frames in the mid-portion of the acoustic interval of vowels. In the case of the zero initials, besides the MaxFrame, two other frames were selected for examination: one was at the one third of the vocalic interval, and the other at the two thirds.

The EPG parameters consisted of the contact area and dispersion indices for the key EPG frame(s). Total contact (TC) was defined as the ratio between the on-electrode number and 62. The contact dispersion indices included Center of Gravity (CoG), Contact Anteriority (CA), Contact Posteriority (CP) and Contact Centrality (CC). CoG was defined in [11] and other three indices in [12].

The acoustic measures included the fundamental frequency (F0) and formant frequency (F1 and F2). The F0 contour was calculated with the EGG signal by the computer program. However, manual adjustment was involved in extracting the T3 contour, for most samples had double-peak EGG wave interval when creaky voice appeared. The vocalic duration of vowels was normalized by using 15 points. The F1 and F2 trajectories were calculated by using LPC method, and manually adjusted for further analysis.

2.4 Data analysis

The F0 contours across vowels and tones were examined first to see the tonal variation. The consonantal environment was not considered because the effect of the preceding consonants on the realization of tones was weaker than the syllable vowels [13]; besides, the consonant-induced F0 perturbation is often found for the F0 contour onset frequency, and temporal alignment with the syllable onset in connected speech [14]. Two strategies were used for two groups of speech tokens. For vowels preceded by zero initials, the articulatory and acoustic measures were compared respectively at the one third, mid-portion, and nine-tenth temporal points in the vocalic interval through two-way ANOVA with between-group factors of Tone and Position. The selection of temporal points will be explained later. The apical and retroflexed vowels will be analyzed similarly with the zero-initial vowels because of the homorganic preceding consonants. For vowels preceded by non-zero initials (apical and retroflexed vowels excluded), the mid-portion parameters were compared through the two-way ANOVA with the between-group factors of Tone and Place of articulation.

The Bonferroni method was used in the ANOVA analysis

3. Results

Figure 2 shows the 15-point normalized F0 contours for syllables across vowels and tones. A glance at the F0 contours indicates that the F0 contour patterns are rather stable across vowels. The F0 contours for T1 syllables tend to be correlated with the tongue height: the F0 contour for the high vowels (/i, u, y/) tend to be slightly higher compared with that for the mid vowel and apical vowel /i1/, both of the latter two vowels involving low tongue body height. The F0 contours for syllable with other three tones are respectively interwoven across vowels, indicating a consistent prosodic pattern by the speaker and stable F0 patterns. Through the direct observation of the F0 contour pattern across vowels, we assume that the magnitude of the tone contour variation is more related to the vowel height than the initials that precede the vowels in question. Regarding the temporal point selection of zero-initial vowels, it can be found that the one-third point into the vocalic length has meaningful correspondence with the four pitch contours: the pitch contour for T2 starts to rise, and for T3 approaches the lowest F0 value, and for T4 starts to fall. The selection of nine-tenth point corresponds to the near-peak F0 value for T2, T3, and T4, which can also be compared with the result in [1].

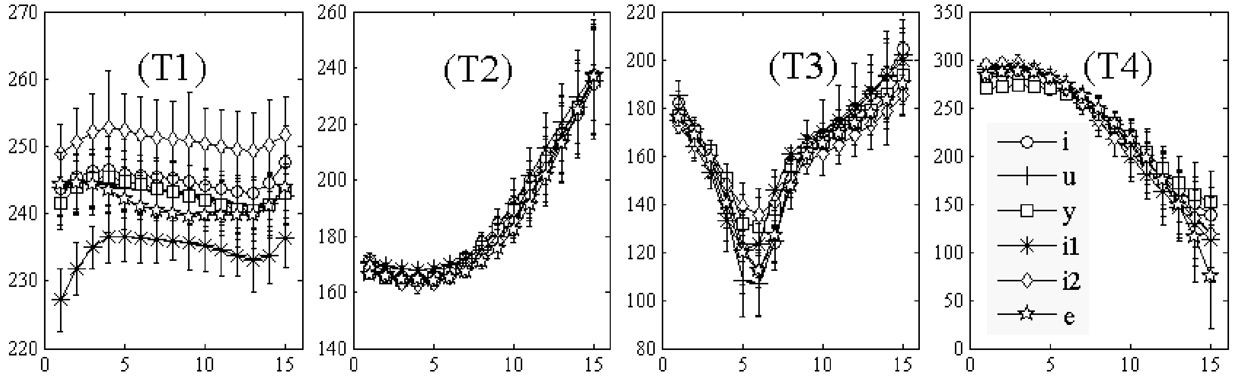


Figure 2: *The F0 contour patterns for four tones across vowels (The abscissa axis represents the frequency and the ordinate axis represents the normalized time points. The error bar stands for one standard error.)*

3.1 Vowel /i/

The two-way ANOVA results show that only the main effect of tone is significant for all measures except CP, indicating that tonal effect is the only source for the articulatory and acoustic variation in producing /i/ regardless of temporal points into the vowel. The post hoc comparisons show that the four out of five EPG indices in T4 syllables are significantly lower than those in syllables with other three tones ($p<0.05$). Meanwhile the CC in T3 is significantly lower than that in T1 ($p<0.05$). These results show that more linguopalatal contact in the anterior region of the palate and higher tongue position are found when the vowel is uttered in T1, T2, and T3 versus T4. Additionally, the tongue height is lower in T3 than in T1. Acoustically, the F1 in T4 is significantly higher than that in other tones ($p\leq0.05$). The F2 in T3 is significantly smaller than that in T1 and T4 ($p<0.05$). Taken together, the most prominent tonal effect on the production of /i/ preceded by zero initial is manifested by the slight different tongue gesture in T4, which is charac-terized by lower tongue height specifically in the anterior part of tongue, thus leading to the F1 increase [15]. The lower tongue height is also found in T3 as is evidenced by lower CC, which might lead to F2 decrease; however, this result is worthy of further examination because other EPG indices do not differ significantly with T1 and T2.

For /i/ preceded by non-zero initials, the Tone and Place (LA, AL, and PA) factors may both affect the vowel production. Because the focus the current paper is on the tonal effects on vowel production, it is expected the Tone factor affects the vowel production in the same pattern regardless of consonantal environment. Thus no Tone × Place interaction effect is expected to exist. However, the post hoc comparison results in Table 2 yielded rather inconsistent tonal effects on the vowel production. Similar tonal effect is found for TC and CC, which is shown by the significantly higher TC and CC in T2 and T3 versus T1 and T4. The significant interaction effect calls for the simple effect analysis when Place was held constant. Results show that the tonal effect on TC and CC are the same as in Table 2 for /i/ preceded by labial initials; only TC is found to be significantly different in /i/ preceded by alveolo-palatal initials (T3>T1,T2,T4, p<0.05); no tonal effect is found for /i/ preceded by alveolar initials. The lower CoG in T4 versus other tones resembles the results when zero initials precede /i/. The significant difference for CP is only found for alveolar initials. CA is found not to be affected by tones.

The F1 difference basically resembles the above result, indicating that lower tongue height in T4 (and possibly in T1) might result in higher F1. The tonal effect on F2 varies in different initial conditions. In case of alveolar initials, the tonal effect is shown to be T2<T4 (p<0.01); in case of alveolo-palatal initials, the tonal effect is shown to be T1<T3, T4 (p<0.01). No tonal effect is found in case of labial initials.

Taken together, the tonal effects on /i/ production is significant, though they are complicated by Place factor. In general, T3 induces more linguopalatal contact in case of LA and PA initials, and higher CP in case of AL initials, and more contact centrality in case of LA initials. Secondly, T4 induces less linguopalatal contact in case of LA and PA initials, which is mainly reflected by less contact in the anterior area of palate (smaller CoG but slightly larger CP), and less contact centrality. The tonal effect for T1 and T2 on /i/ is not consistent either articulatorily and acoustically.

Table 2. *Two-way ANOVA results for the measures of /i/ preceded by initials of three places (* p<0.05, **p<0.01, ***p<0.001; n.a. means no significant difference.)*

Measure	ANOVA results		
	Tone	Place	Interaction
TC	T1,T4<T2,T3*	PA>LA,AL**	*
CoG	T4<T1,T2*	PA>LA>AL***	n.a.
CA	n.a.	PA>LA >AL**	n.a.
CP	T1<T4,T3*	PA<LA<AL***	*
CC	T1,T4<T2,T3*	PA> LA,AL ***	**
F1	T4>T2,T3*** T1>T2***	n.a.	n.a.
F2	T1<T3,T4** T2<T4**	PA<LA,AL***	*

To summarize, T3 and T4 have significant effect on the articulation of /i/: T3 increases the linguopalatal contact and contact centrality while T4 decreases both. The tonal effect on F1 is salient for T4. The F1 is higher in T4 versus other tones. However, tonal effect on F2 of /i/ is not consistent. This might be attributed small capacity of speech samples.

3.2 Vowel /y/

Like /i/, only the main effect of tone on the EPG measures for /y/ is found. Post hoc comparison results show that the five EPG measures in T3 are significantly higher than other measures (TC: T3>T1,T2; CoG: T3,T4,T1>T2; CA : T3>T2; CP : T3>T1; CC: T3>T1, T2; p<0.05). In sum, the tongue position of /y/ in T3 is higher, more anterior, and more centralized compared with that in other tones. Regarding the acoustic parameters, the F1 in T4 is significantly higher than that in other tones (T4>T1,T2,T3, p<0.001). However, the Time×Tone interaction effect is marginally significant (p=0.07). In Figure 3 it can be found that the F1 contour for /y/ in T4 is not stable in the vocalic interval. The initial F1 is prominent high at the outset and gradually decreases toward the end of syllable. One-way ANOVA results yielded that significant difference of F1 only exist at the one-third and mid-portion points. Referring to Figure 2 and the larynx vertical position at the outset of T4 [9], the gradual decrease of F1 might be attributed to the vocal tract length variation [16] or the effect of source coupling [17].

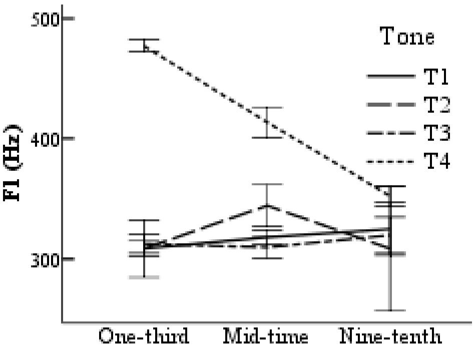


Figure 3: *The F1 for /y/ preceded by zero initial across four tones and three positions in the syllable (Error bar stands for one standard error).*

Because /y/ can only be preceded by alveolar initial /n, l/ and alveolo-palatal initials, the tonal effect was analyzed through one-way ANOVA respectively for each initial place. When the initial is lateral /l/, there is no T1 syllable. Results show that T3 has significant effect on the articulatory measures while T4 on the acoustic measure. The production of /y/ in T3 is shown to have more linguopalatal contact, more advanced and centralized tongue position. The tonal effect on F1 is the same as in /i/ and /y/ with zero initial.

Table 3. *One-way ANOVA results for the measures of /y/ preceded by /l/ and three PA initials.*

Measure	ANOVA results	
	Tone (/l/)	Tone(PA)
TC	T3>T4 (p=0.07)	n.a.
CoG	T3,T2>T4 **	n.a.
CA	n.a.	n.a.
CP	n.a.	n.a.
CC	T3>T4 *	n.a.
F1	T4>T2, T3 *	T1,T4>T2,T3*
F2	n.a.	n.a.

No significant difference was obtained for the articulatory measures for /y/ in case of non-zero initials. However, in Figure 4 we can find that the linguopalatal contact and centralization in T3 is prominently higher than those in other tones. Additionally, the tongue is more advanced in T3, as is shown by AC in Figure 4 and CoG (Not present). The combined findings indicate that tongue body is more anterior and raised in T3 versus other tones. The F1 of /y/ in T3 and T2 was significantly lower than in T1 and T4.

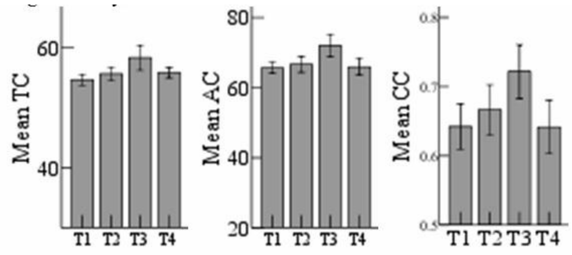


Figure 4: *The TC and CC for /y/ with non-zero initials* initials.

3.3 Vowel /i2/

In analyzing /i2/, we compare the measures at three time points over the vocalic interval because it is preceded by homorganic initials. In general, the post-alveolar constriction for /i2/ becomes gradually released toward the end of the syllable. Meanwhile, the factor of Position seldom interacts with Tone.

Table 4. *Two-way ANOVA results for the measures of /i2/. In the Position column,* ①, ②, ③ *respectively stands for one-third, mid-portion, and nine-tenth point.*

Measure	ANOVA results		
	Tone	Position	Interaction
TC	T4,T3<T2,T1**	①②>③**	n.a.
CoG	T4,T1<T3,T2**	①②>③**	n.a.
CA	T4<T1,T3,T2***	①②>③**	n.a.
CP	T4,T3,T2<T1*	n.a.	n.a.
CC	T4<T1,T3<T2*	①②>③**	*
F1	T4,T1>T2,T3***	②<③*	n.a.
F2	T1<T3*	n.a.	n.a.

The post hoc comparison show rather complicated tonal effects on the articulation of /i2/, which might be attributed to the articulatory instability of this vowel uttered by the speaker [18]. Nonetheless, the vowel in T3 tends to have less linguopalatal contact and advanced tongue gesture. The vowel in T4 has lowered tongue height, which is similar as in /i, y/. The vowel in T1 has the highest CP, indicating an active raise of tongue body toward hard palate. And the vowel in T2 has the highest level of centralization at the three time points. By comparing the results for vowels above, we argue that active tongue advancement is evidenced in T3, while the same strategy tends to be adopted in T2. The tonal effect of T4 is manifested by the tongue height lowering with all five EPG measure being the lowest compared with those in other tones. The tongue gesture adjustment in T1 is still unclear. It might be patterned with T4 by having more retracted tongue gesture (CoG, CA, CP) and a lower tongue height (CC), though it has highest TC.

The acoustic results resemble those for /i, y/ regarding F1. And the lower F2 in T1 versus T3 further indicates more retracted tongue body.

3.4 Vowel /u/

The post hoc comparison results for /u/ preceded by zero initial show a much clear picture. The EPG measures in T3 are significantly lower than other tones (TC: T3<T1,T4,T2; CoG: T3<T2,T4; CA: T3<T2,T4; CP: T3<T4,T1,T2; CC: T3<T4,T1,T2; p<0.05). These results show that the tongue gesture in T3 is protrudingly lowered and retracted, which supports the results in the previous studies [6,7]. The tonal effect on F1 is the same as in the vowels above with F1 in T4 significantly higher than that in other tones (p<0.05). The F2 in T4 is significantly higher than that in T1 (p=0.09).

Table 5 shows a rather clear picture of tonal effects on the articulation and acoustics of /u/. Like the case in zero initials, tongue retraction and lowering gesture is found for /u/ in T3. This tongue gesture adjustment strategy may also be partially utilized in the production of T2, for tongue retraction, but not lowering, is found for /u/ in T2. The acoustic comparison results show that the F1 and F2 are significantly lower in T2/T3 versus T1/T4. However, the higher F1 in T4 (and possibly T1) is argued to be associated with shortened vocal tract as a result of active larynx upward movement or coupling effect of F0.

Table 5. *Two-way ANOVA results for the measures of /u/ preceded by initials of four places*

Measure	ANOVA results		
	Tone	Place	Interaction
TC	T3,T2<T1* T3<T4***	LA<RE,AL**	n.a.
CoG	T3,T2<T4,T1*	LA<VE,RE,AL*	n.a.
CA	T3,T2<T1*** T3<T4*	LA<RE,AL**	n.a.
CP	T3<T2,T4,T1*	n.a.	n.a.
CC	T3<T1,T2*	n.a.	n.a.
F1	T2,T3<T4,T1***	LA>RE,VE,AL*	n.a.
F2	T2,T3<T4,T1***	n.a.	n.a.

3.5 Vowel /i1/

Like the retroflexed vowels /i2/, the apical vowel /i1/ is preceded by the homorganic apical initials, thus we compared the measures at three time points.

Table 6. *Two-way ANOVA results for the measures of /i1/.*

Measure	ANOVA results		
	Tone	Position	Interaction
TC	T3<T4, T1,T2**	①②＞③***	**
CoG	T3>T1,T4*	n.a.	n.a.
CA	T3<T1, T2*	①②＞③***	***
CP	T3<T4, T2, T1***	n.a.	n.a.
CC	T3<T1, T2*	①②＞③**	***
F1	T2<T1,T3<T4**	②＜③*	***
F2	T3,T2,T1<T4*	n.a.	n.a.

As shown in Table 6 four out of seven measures are unstable in the vocalic interval, which is mainly caused by articulatory perturbation toward the end of /i1/. However, the tonal effect is rather clear-cut for the articulation of /i1/, with T3 induces less linguopalatal contact in the posterior area on the palate, indicating a lowered tongue body gesture. Meanwhile, the lesser degree of centralization in T3 is caused by less anterior contact as indicated by lower CA (r=0.983, p<0.001). The highest CoG in T3 is negatively correlated with reduced posterior contact (CP, r=-0.702, p<0.001, also shown in Figure 5).

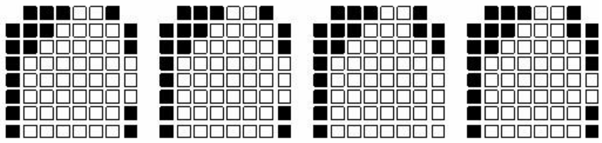


Figure 5: *The palatograms of the mid-portion frame /i1/ in the syllable /tsɿ/ uttered in four tones (From left to right are the palatograms in T1, T2, T3, and T4 tones).*

The tonal effect on F1 is similar with the results above, and F2 variation as a result of tonal conditions resembles that for /u/. Taken together, the tonal effects on /i1/ are represented by lowered tongue body and tongue tip gestures in T3. The retraction of tongue in T3 is not supported because CoG is not reduced as is in /u/. T4 has significant effect on F1/F2, which might be associated with either larynx upward movement in T4 or F0 coupling effect.

3.6 Vowel /ɤ/

Table 7 shows the post hoc comparison results for /ɤ/ preceded by the zero initial. In general, the articulatory measures decrease monotonously toward the end of the vowel, and the F1 increases and F2 decreases toward the end of the vowel, indicating the gradual releasing of the vowel gesture in the vocalic interval. However, the effect of T3 is consistent along the vowel length for four out of five EPG measures, showing a lowered and retracted tongue gesture in T3 (and possibly T2). Since the interaction effect is significant for CA, Figure 6 shows that the significant difference exists between T1/T4 and T2/T3 (p<0.001) at the onset, but the difference turns marginally significantly at mid-point (T2<T4, p=0.07), and is nullified toward the end of the vowel. Acoustically, the effect of T4 on F1 is significant, making a higher F1 than in T2 and T3. The F2 variation is a reflection of tonally- and temporally-conditioned vocalic gesture: lower F2 is found for T2 versus T1 at least toward the mid-point of the vowel, but no difference is found near the end of the vowel.

Table 7. *Two-way ANOVA results for the measures of /ɤ/.*

Measure	ANOVA results		
	Tone	Position	Interaction
TC	T3<T4 T1*	①＞②＞③***	n.a.
CoG	n.a.	①②＞③**	n.a.
CA	T2,T3<T4**	①＞②＞③**	**
CP	T3<T1*	①＞②＞③*	n.a.
CC	n.a.	①＞②＞③*	n.a.
F1	T2, T3<T4*	①②＜③*	n.a.
F2	T1>T2*	n.a.	n.a.

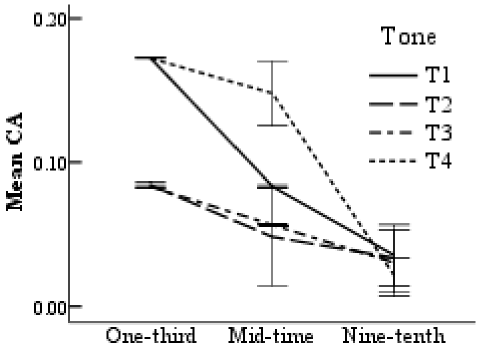


Figure 6: *The CA for /ɤ/ across four tones at three temporal points over the vowel interval.*

For /ɤ/ preceded by non-zero initials, only syllables with velar (VE) and retroflex (RE) initials were selected. Thus one-way ANOVA was carried out respectively for these two types of initial condition. No tonal effect is found for /ɤ/ preceded by VE initials. However, the tonal effect on F1 is consistent to the findings above, with the highest F1 in T4. Tonal effect is significant for /ɤ/ preceded by RE, which is shown by the lowered and retracted tongue gesture in T3 compared with other tones. The tonal effect on F1 is similar to the result above.

In sum, the tonal effect on the production of /ɤ/ is rather similar to that of /u/: on the one hand, the lowered and retracted tongue gesture is found in T3 versus other tones; and higher F1 in T4 versus other tones on the other.

Table 8. *One-way ANOVA results for the measures of /ɤ/ preceded by VE and RE initials.*

Measure	ANOVA results	
	Tone (VE)	Tone(RE)
TC	n.a.	T3<T2,T1,T4*
CoG	n.a.	n.a.
CA	n.a.	T3<T1**
CP	n.a.	T3<T2,T1,T4***
CC	n.a.	n.a.
F1	T2<T1,T3, T4* T1<T4 (p=0.05)	T2<T1,T3, T4*
F2	n.a.	n.a.

4. Discussion and Conclusions

To summarize, several patterns are found for the tonal effects on the vowel articulation and acoustics in the present paper: (1) the articulatory variation in vowel production was mainly induced by the low tone (T3) compared with other tone. (2) Two strategies for articulatory adjustment were identified: the tongue lowering and retraction was found for vowels in T3 syllables when the post-dorsum of tongue is involved in vowel articulation, as shown in /u, ɤ, i1/; the increased tongue raising was found in T3 syllables when the pre-dorsum of tongue is involved in vowel production, as shown in /i, y, i2/. (3) The T3 influence on the articulatory measures for vowels tends to be present in the better part of the vocalic length, regardless of the pitch changes. (4) Consistent tonal effect on F1 for the vowels was found for which T4 tends to increase the F1. This might be attributed to the larynx raising that leads to the shortening of the vocal tract.

The results of the current paper basically confirm the previous results of tonal effects on the vowel articulation [1,5,6,7] and vowel acoustics [4, 16]. Furthermore, the current paper augments the previous research in that two articulatory strategies might be involved in vowel production. Apart from the lowered and retracted tongue gesture in T3 for vowels by which the posterior tongue is active in vowel production, the high and front vowels tend to involved heightened, or advanced, tongue gesture in vowel production in T3. Meanwhile the tonal effect of rising tone (T2) on the vowel production tends to be patterned with T3, which, in many cases, differ significantly from T1 and T4. This indicates that tongue gesture is associated with the starting pitch value in that the tongue gesture might be different depending on whether the initial F0 is high or low. However, this hypothesis should be called into caution because only one speaker and one type of speech task is involved in this study. The final conclusion is still pending before more stimuli are to be analyzed and data from more speakers are collected.

The tonal effect on the vowel acoustics is mainly manifested by F1. T4 shows a consistent effect in raising F1. In Figure 2 it can be conjectured that this may be associated with high starting F0 that is often much higher than that in T1. This result may be attributed to the both constriction size and vocal length in vowel production [19], as lowered tongue gesture is found in high and front vowel on the one hand, and the active larynx vertical rise shortens the vocal tract length on the other.

5. Acknowledgements

This research was supported by the China Ministry of

Education Humanities and Social Sciences Grant (No. 11YJAZH055), China National Social Sciences Grant (No. 11BYY111), and China National Natural Science Funds (No. 61073085). We thank the constructive comments from three anonymous reviewers.

6. References

[1] Hu, F., "Tonal effect on vowel articulation in a tone language", Proc. Tonal Aspects of Languages, Beijing, 79-82, 2004.

[2] Ohala, J.J. and Eukel, B.W., "Explaining the intrinsic pitch of vowels", in Channon and Shockey [ED], In Honor of Ilse Lehiste, 207-215, Dordrecht, 1987.

[3] Whalen, D.H. and Levitt, A.G., "The universality of intrinsic F0 of vowels", J. Phonetics, 23, 349-366, 1995.

[4] Zee, E., "Tone and vowel quality", J. Acoust. Soc. Am, 62(S1), 1977.

[5] Torng, P.C., "Supralaryngeal articulator movements and laryngeal control in mandarin Chinese tonal production", Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 2000.

[6] Erickson, D.,Iwata, R., Endo, M. and Fujino, A., "Effect of tone height on jaw and tongue articulation in Mandarin Chinese", Proc. Tonal Aspects of Languages, Beijing, 53-56, 2004.

[7] Hoole, P. and Hu, F., "Tone-vowel interaction in Standard Chinese", Proc. Tonal Aspects of Languages, Beijing, 89-92, 2004.

[8] Honda, K., Hirai, H., Masaki, S. and Shimada, Y., "Role of vertical larynx movement and cervical lordosis in F0 control", Language and Speech, 42, 401-411, 1999.

[9] Wang, G. and Kong, J., "The relation between larynx height and F0 during the four tones of Mandarin in X-ray movie", Proc. ISCSLP, Taiwan, 335-338, 2010.

[10] Li, Y. and Pan, X., "Temporal alignment algorithm for electropalatographic and acoustic signals in long utterances", Intl. Conf. on Audio, Language and Image Processing, Shanghai, 2012. (Submitted for review).

[11] Hardcastle, W.J., Gibbon,F. and Nicolaidis, K., "EPG data reduction methods and their implications for studies of lingual coarticulation", J. Phonetics, 19, 251-266, 1991.

[12] Fontdevila, J.and Pallares, M.D., Recasens, D., "The contact index method of electropalatographic data reduction", J. Phonetics, 22(2), 141-150, 1994.

[13] At, H., "The acoustic variation of Mandarin Tones", Phonetica, 33(5), 353-367,1976.

[14] Xu, C.X. and Xu, Y., "Effects of consonant aspiration on Mandarin tones", Journal of International Phonetic Association, 33 (2), 165-181, 2003.

[15] Pickett, J.M., "The Acoustics of Speech Communication: Fundamental, Speech Perception Theory, and Technology", Allyn & Bacon, 1999.

[16] Zhang, J.L., "The intrinsic fundamental frequency of vowels and the effect of speech modes on formants", Acta Acoustica, 14(6), 401-406, 1989;

[17] Atkinson, J.E., "Aspects of intonation in speech: implications from an experimental study of fundamental frequency", Unpublished doctoral dissertation, University of Connecticut, 1973.

[18] Li, Y., "Electropalatographic study on segmental coarticulation in Standard Chinese", Unpublished doctoral dissertation, Peking University, 2011.

[19] Stevens. K., "Acoustic Phonetics", MIT Press, 1998

The Supraglottal Constriction in Tibetan Chants*

——Electroglottographic Evidences

YOSHINAGA Ikuyo, KONG Jiangping

Abstract

The phonation of Tibetan chants was examined using electroglottographic analysis. The supraglottal constriction, which was considered to generate their peculiar phonation qualities, was examined by a comparison with the vocal fold adduction. The results suggest that the frequency of the ventricular fold oscillation was the same as F0, and the closing peak of the ventricular fold adduction occurred approximately 171 degree after the vocal fold adduction. As the ventricular fold adduction immediately followed the glottal release, it generated the glottal pulse with double peaks in the corresponding glottal airflow. Low OQ_{egg} value was observed because the ventricular fold adduction, which occurred during the glottal release, lowered its value.

Keywords:

supraglottal constriction; ventricular folds; open quotiend (OQ_{egg}); speed quotient (SQ_{egg}); phase difference

1. Introduction

Tibetan lamas in their red robes chant sutras. The pure sounds of their low sonorous pitch heal listeners. *Sabda-vidya*, which deals with ancient Indian linguistic and grammatical studies including singing sutras, was one of the five fields of academics study in ancient India and was deeply treasured and successfully handed down by Tibetan Buddhists.

The supraglottal constriction is estimated to occur in Tibetan Buddhist chants using electroglottographic (EGG) and acoustic analyses [1]. These supraglottic phonations have been reported to occur in certain singing modes, for instance, Mongolian *Kargyraa* in 'throat singing' [2], ethnic and pop style singing [3], and Japanese traditional Noh singing [4]. The supraglottal constriction is widely considered to be caused by movements of the ventricular and aryepiglottic folds (Fig.1), the ventricular folds oscillate at the speed of F0, F0/2 or F0/3 in vocal-ventricular mode (VVM), so do the aryepiglottic folds at the frequency of F0/2 in growl voice [1, 2, 3, 4, 5]. The earlier EGG assessments of phonation types revealed that period-doubling EGG waveforms are the characteristics of VVM [6]. A certain mode of Tibetan chants is assessed as harsh voice [1], in which the ventricular folds generally become involved in the phonation of the true vocal folds [7]. These irregular supraglottic phonations have been found not only in singing but also in vocal fry, voice instabilities, and infant vocalizations. These irregular vocalizations are often interpreted as period-doubling bifurcations, and the corresponding acoustical signals often show sudden jumps to subharmonic regimes [8, 9].

In this paper, we deal with a certain singing mode of Tibetan Buddhist chants. The phase difference between the adduction of the vocal folds and that of the ventricular folds and EGG-based parameters of those adductions are examined using electroglottographic analysis to investigate the supraglottal movements and their contributions for the production of peculiar sounds.

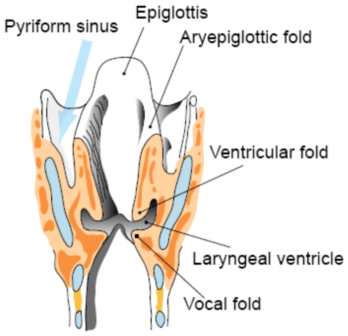


Figure1. Coronal view of the larynx, as seen from behind [3]

2. Material and Method

This section describes the voice materials, calculation method of Electroglottography (EGG)-based parameters, and data processing procedure.

2.1 Voice Material

The phonation of Tibetan Buddhist chants was studied in one Tibetan male monk from *Kumbum* Monastery of *Dge-Lugs-Pa*. The monk was 31 years old, with 18 years of priest experience, when the recording was performed. He was also a teacher at the Monastery with an excellent reputation for his chanting.

The sustained vowel /a/ phonated at 93.7 Hz (≒F2#), whose EGG signal was formed as clear period-doubling waveforms, was studied in this research.

The data acquisition took place at Kumbum Monastery in Qinghai province, China. The EGG signal was obtained by an EGG system (Electroglottograph Model 6103; Kay, USA). The audio signal was recorded by a Sony Electret Condenser Microphone. Those signals were simultaneously recorded and digitized at 16-bit resolution at a sampling frequency of 44.1 kHz.

2.2 Parameter Calculation Method

The EGG signals provide meaningful information only when the vocal folds repeat contact and de-contact during vibration. Therefore, contact-based analysis is the common algorism. A few parameters can be extracted from the EGG waveform that roughly correspond to the open quotient (OQ) and speed quotient (SQ). Because the EGG and airflow waveforms differ from each other qualitatively, OQ_{egg} and SQ_{egg} are employed in this study as the EGG-based parameters. Fig.2 shows that a period of EGG signal can be divided into contact and de-contact phases. Furthermore, the contact phase can be divided into contacting and de-contacting.

Basically, three kinds of EGG calculation methods are proposed, i.e., criterion-level [10], DEGG [11, 12, 13, 14, 15, 14] and the combination of the criterion-level and DEGG methods, called the hybrid method [17, 18]. The DEGG method is employed in this research, in which the glottal closing instance (GCI) and glottal opening instance (GOI) are determined as the maximum and minimum values of the DEGG waveform (Fig.2).

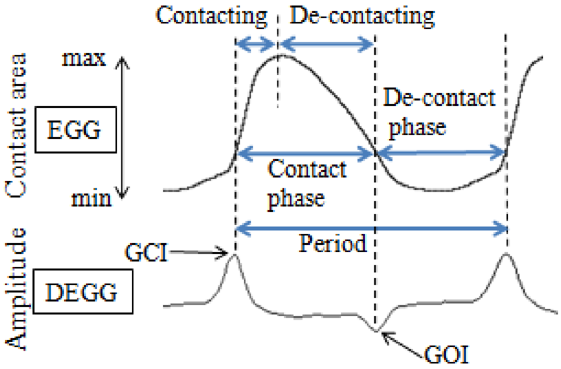


Figure 2. EGG waveform and phases of vocal fold contact.

Three EGG-based parameters are extracted: F0, OQ_{egg} and SQ_{egg}. The definitions of F0 and OQ_{egg} are described as: F0=1/period and OQ_{egg}%=de-contact phase/period*100. Although the SQ_{egg} can be varied in detail across researchers, the definition used in this study is SQ_{egg}%=de-contacting/contacting*100 [19].

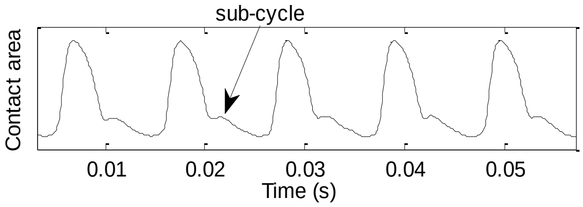


Figure 3. Five vibratory cycles of the EGG signal from vowel /a/ phonated at F2# (92.5 Hz).

The EGG waveform in the data from Tibetan chants demonstrates period-doubling phenomena (Fig.3). These phenomena are quite similar to what was observed in VVM [5, 6]. Furthermore, the ventricular or the aryepiglottic folds are considered to be involved in the supraglottal constriction, however, the aryepiglottic folds oscillate at the speed of $F_0/2$ because the aryepiglottic folds locate higher than the ventricular folds with less developed muscles. Indeed, it is quite reasonable to admit that sub-cycles observed in this study are caused by the oscillation of the ventricular folds.

The time-based parameters of the DEGG waveforms also yield information about periodicity and time patterns of the vibratory events. They are generally referred to as period time (T0), GCI, and GOI. And we call the maximum instance of the vocal fold contact area 'MI' in this study (see Fig.4).

Fig.5 shows the DEGG waveform which corresponds to the period-doubling EGG waveform. The parameters of the vocal fold adduction and the ventricular fold adduction are extracted. To avoid confusion between them, ‘$_2$’ is suffixed to the parameters of the latter. Two DEGG-based parameters are extracted: T0=GCI(n+1)-GCI(n) and T0$_2$=GCI$_2$(n)—GCI(n), ‘n’ is the number of the glottal cycles. T0 indicates the duration of each glottal cycle, T0$_2$ indicates the duration between the closing peak of the vocal fold and that of the ventricular fold closure. The phase difference between them is given as: (T0$_2$/T0)*360 °.

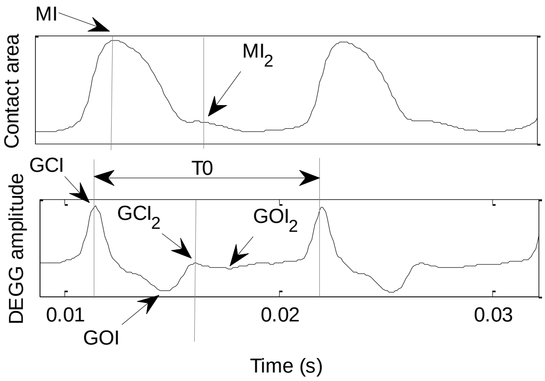


Figure 4. Time-based parameters of the EGG and DEGG waveforms.

According to the definitions of EGG-based parameters described above, three parameters of the vocal fold oscillation are given as: F0=1/T0, OQ$_{egg}$%=(T0-(GOI-GCI))/T0*100, and SQ$_{egg}$%=(GOI-MI)/(MI-GCI)*100. And OQ$_{egg2}$ and SQ$_{egg2}$ of the ventricular fold oscillation are given as: OQ$_{egg2}$%=(T0-(GOI$_2$-GCI$_2$))/T0*100 and SQ$_{egg2}$%=(GOI$_2$-MI$_2$)/(MI$_2$-GCI$_2$)*100.

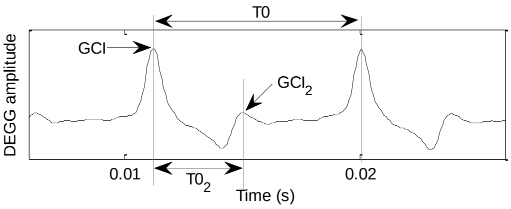


Figure 5. Definitions of T0 and T0$_2$

2.3 Data Processing

The recorded file was downsampled to 11,025 Hz. And the EGG rumble, which was caused by up and down laryngeal movements, was filtered out by a high-pass filter with the cutoff frequency set at 60 Hz because it could affect or mislead the parameter extraction. The parameter values for all of the cycles were extracted using the DEGG method and were saved in an Excel file, the parameter values of 247 data points were extracted from the file concretely. The data processing was performed by a Matlab-based program.

3. Phase Analysis

This section describes the phase difference between the oscillation of the vocal folds and that of the ventricular folds using DEGG analysis.

3. 1 Duration of T0 and T0$_2$

Fig.6 shows the durations of T0 and T0$_2$, they are presented by black and gray dots respectively. The abscissa indicates time, and the ordinate indicates the duration of T0 and T0$_2$ of each cycle.

It seems that T0 is more constant than T0$_2$, the standard deviation (SD) of T0$_2$ is 0.16ms which is 0.09 ms higher than that of T0 (Table II). The maximum value of T0$_2$ is 5.44 ms, and the minimum value is 4.72 ms. The range between them ups to 0.73 ms, on the other hand, the range of T0 is only 0.36 ms which is about a half of T0$_2$. These results suggest that the duration from the vocal fold closure to the ventricular fold closure is not as stable as the duration of each glottal cycle. The mean values of T0 and T0$_2$ is 10.67 ms and 5.07 ms. In other words, the ventricular fold closure occurs approximately 5.07 ms after the vocal fold closure, and the next vocal fold closure follows 5.6 ms after the ventricular fold closure.

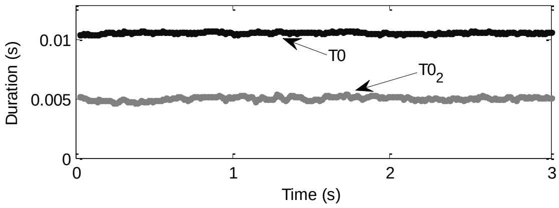


Figure 6. Durations of T0 and T0$_2$.

TABLE III MEAN AND SD VALUES OF T0 AND T0$_2$

	Mean (ms)	SD (ms)
T0	10.67	0.07
T0$_2$	5.07	0.16

3.2 Phase difference between the vocal fold and the supraglottal oscillations

Glottal cycles can be described with 0-360 degree scale. In this study, GCI is described as 0 ° as well as 360 °because it is the end of the glottal cycle as well as the beginning of the next glottal cycle.

Fig.7 shows the phase difference between the adduction of the vocal folds and that of the ventricular folds, it is presented by black dots. The abscissa indicates time, and the ordinate indicates the phase difference. Fig.7 roughly indicates that the supraglottal closure occurs approximately 180 ° after the vocal fold closure, and it shows that the values are not very stable. Table IV shows the mean and SD values of the phase difference between the adduction of the vocal folds and that of the ventricular folds. The mean value of the phase differences is 171.17 ° with the SD of 5.26 °. The earlier researches on VVM using the high-speed videoendoscopy revealed that the ventricular fold closure occurred during the 480 °-560 ° interval during the vocal folds were open [5]. In other words, the ventricular fold closure occurred during the 120 °-200 ° interval every other glottal cycle because the frequency of the ventricular fold oscillation was F0/2 in the above case. The mean value of the phase differences between the adduction of the vocal folds and that of the ventricular folds in this research is 171.17 °, the frequency of the ventricular fold oscillation is equal to that of the vocal fold oscillation. The phase difference of 171.17 ° is also in the 120 °-200 ° interval, and the ventricular fold adduction also occurs during the glottal release in this case.

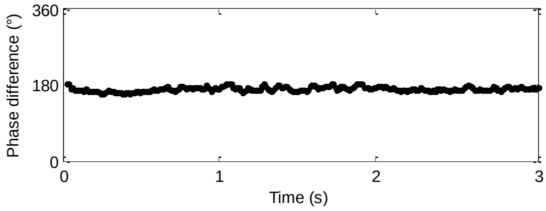


Figure 7. Phase difference between the vocal fold and ventricular fold adductions

TABLE V MEAN AND SD VALUES OF THE PHASE DIFFERENCE BETWEEN THE VOCAL FOLD AND VENTRICULAR FOLD ADDUCTIONS.

	Mean (°)	SD (°)
The phase difference between the vocal fold and ventricular fold adductions	171.17	5.26

[20] has reported on the relationship between the glottal air flow and vocal fold contact area. According to the result, opening of upper fold margins corresponds to the first airflow of the pulse, and the pulse terminates closely before lower fold margins close. Indeed, an upside-down image of the EGG waveforms roughly demonstrates the glottal airflow. Fig.8 shows the estimated glottal airflow based on the results from the phase analysis. An upside-down image of the EGG waveforms is shown as the estimated glottal airflow. 0 °and 171 ° indicate the decreasing peaks of the airflow, the former is caused by the vocal fold closure, and the latter is caused by the ventricular fold closure. A pulse can be described as follow: the airflow starts to be released, the volume velocity of the glottal pulse decreases closely before 171 °, and the peak volume velocity of the pulse follows it, then the pulse terminates. As a result, the glottal airflow is formed with two peaks.

As mentioned above, the SD values of the phase difference between the adduction of the vocal folds and that of the ventricular folds are not so stable that the volume velocity of the glottal airflow at the first peak, which occurs closely before 171 °, is affected by the unstable phase difference. In short, the glottal airflow is transformed by the ventricular fold constriction which seems to play an important role in generating their peculiar voice qualities.

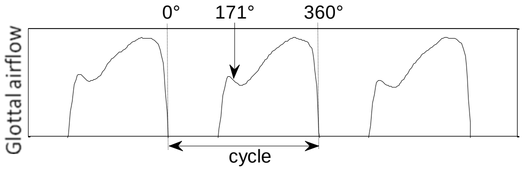


Figure 8. Estimated glottal airflow.

4. Parameter Analysis

This section describes the characteristics of EGG-based parameters of the vocal fold and ventricular fold oscillations to observe their adduction mechanisms.

4.1 OQ_{egg}

As mentioned above, OQ_{egg} indicates OQ_{egg} of the vocal fold oscillation, OQ_{egg2} indicates OQ_{egg} of the ventricular fold oscillation in this study. Fig.9 shows the OQ_{egg} and OQ_{egg2} values, black dots indicate OQ_{egg}, and gray dots indicate OQ_{egg2}. The abscissa indicates time, and the ordinate indicates the OQ_{egg} and OQ_{egg2} values of each cycle.

Table VI shows the mean and SD values of OQ_{egg} and

OQ_{egg2}. The mean OQ_{egg} is 67.64%, the SD value is 1.73%. The mean OQ_{egg2} is 74.59%, the SD value is 0.28%. The mean OQ_{egg2} is 6.95% higher than OQ_{egg}, it suggests that the contact phase of the ventricular fold oscillation is shorter than that of the vocal fold oscillation. Furthermore, the OQ_{egg2} value is more stable than OQ_{egg}, for the SD value of the former and the latter are 0.28% and 1.73%. The stable T0 and unstable $T0_2$ were the result from the earlier section, unstable OQ_{egg} seems to be caused by the unstable $T0_2$. It seems plausible that the ventricular fold constriction, which shortly follows the glottal release, influences the vocal fold movements. Correlation analysis between OQ_{egg} and $T0_2$ reveals that there is a negative correlation of -0.492 ($p<0.01$) between them.

The real OQ_{egg} value can be extracted by subtracting the contact quotient (CQ_{egg}) value of the ventricular fold oscillation from OQ_{egg} because the ventricular fold constriction occurs at the de-contact phase of the vocal fold oscillation. The mean real OQ_{egg} is 42.23% which is characterized as significantly low OQ_{egg}. It is quite natural that the phonation containing the supraglottal constriction has low OQ_{egg}.

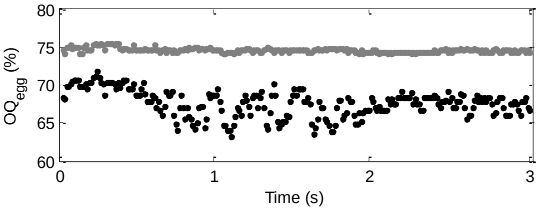


Figure 9. OQ_{egg} values of the glottal and the supraglottal adductions.

TABLE VII MEAN AND SD VALUES OF OQ_{egg}

	Mean OQ_{egg} (%)	SD (%)
The glottal oscillation	67.64	1.73
The supraglottal oscillation	74.59	0.28

4.2 SQ_{egg}

As mentioned above, SQ_{egg} indicates SQ_{egg} of the vocal fold oscillation, SQ_{egg2} indicates SQ_{egg} of the ventricular fold oscillation in this study. Fig.10 shows the SQ_{egg} and SQ_{egg2} values, black dots indicate SQ_{egg}, and gray dots indicate SQ_{egg2}. The abscissa indicates time, and the ordinate indicates the SQ_{egg} and SQ_{egg2} values of each cycle.

Table VIII shows the mean and the SD values of SQ_{egg} and SQ_{egg2}. The mean SQ_{egg} is 150.52% with the SD value of 11.00%, and the mean SQ_{egg2} is 100.73% with the SD value of 2.17%. The mean SQ_{egg} is 49.79% higher than SQ_{egg2}. These results suggest that the strong force of the glottal adduction raises the energy at the high frequency region in the vocal fold oscillation. On the other hand, low SQ_{egg2} resulted from less forced adduction because the ventricular folds are incapable of becoming tense, since they contain very few muscle fibres [21]. Fig.10 shows that the SQ_{egg} values are unstable, the SD value reaches 11% which is 8.83% higher than that for the ventricular fold oscillation. Correlation analysis reveals that there is a negative correlation of -0.629 ($p<0.01$) between the OQ_{egg} and the SQ_{egg}.

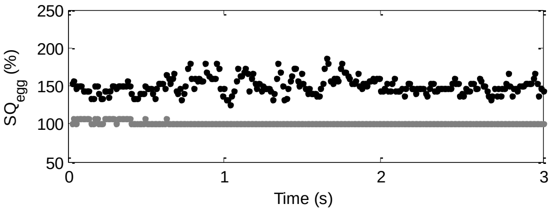


Figure 10. SQ_{egg} values of the glottal and the supraglottal adductions.

TABLE IX MEAN AND SD VALUES OF SQ_{egg}

	Mean SQ_{egg} (%)	SD (%)
The glottal oscillation	150.52	11.00
The supraglottal oscillation	100.73	2.17

EGG-based parameter values of the ventricular fold oscillation are rather stable compared to those of the vocal fold oscillation. It is quite natural that the values of the latter are complicated because the muscles of the vocal folds and those to control the vocal folds are much more developed than those of the ventricular folds. The mean value of real OQ_{egg} is 42.23%, its significantly low OQ_{egg} is one of the characteristics of pressed phonation. The OQ_{egg} of the vocal fold oscillation and the phase difference between the adduction of the vocal folds and that of the ventricular folds are negatively correlated, and the OQ_{egg} and SQ_{egg} of the vocal fold oscillation are negatively correlated. In short, the ventricular fold oscillation occurs unstably in phases, but the values of OQ_{egg2} and SQ_{egg2} are very stable. On the contrary, the duration of each glottal cycle is quite stable, but the values of OQ_{egg} and SQ_{egg} are not very stable because of the unstable occurrence of the ventricular fold oscillation in phases.

5. Conclusion

The ventricular fold constriction was estimated to occur approximately at 171 ° phase difference after the vocal fold closure. As the ventricular fold constriction immediately followed the glottal release, the estimated glottal airflow was formed with double peaks of the glottal pulse. Though F0 was very stable, the OQ_{egg} and the SQ_{egg} of the vocal fold oscillation were somehow unstable because of the unstable phase difference between the adduction of the vocal folds and that of the ventricular folds. The ventricular fold constriction contributed to lower the OQ_{egg} value. These phonation techniques involving the supraglottal structures made a great role in generating their peculiar phonation qualities.

Further physiological research using tools such as high-speed cameras is needed to clarify the ventricular adduction. Synthesize and perceptual evaluations are also expected as future work.

6. Acknowledgment

Thanks are due to Gazangduojie from Kumbum Monastery for the recording and valuable advices on this research.

This study was supported by China Social Sciences Funds of the Ministry of Education (Grant No: 10JJD740007).

7. References

[1] I. Yoshinaga and J. Kong, "Some phonatory characteristics of Tibetan Buddhist chants," Journal of the Phonetic Society of Japan, vol. 15(2), pp. 83-90, 2011.

[2] P. Lindestad, M. Södersten, B. Merker, and S. Granqvist, "Voice source characteristics in Mongolian 'throat singing' studied with high-speed imaging technique, acoustic spectra, and inverse filtering," Journal of Voice, vol. 15(1), pp. 78–85, 2001.

[3] K-I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama, "Growl voice in ethnic and pop styles," In Proceedings of the International Symposium on Musical Acoustics. Nara, 2004.

[4] I. Yoshinaga and J. Kong, "Laryngeal vibratory behavior in traditional Noh singing," Tsinghua Science and Technology, vol. 17(1), pp. 94-103, 2012.

[5] L. Fuks, B. Hammarberg, and J. Sundberg, "A selfsustained vocalventricular phonation mode: acoustical, aerodynamic and glottographic evidences," KTH TMHQPSR, vol. 3, pp. 49–59, 1998.

[6] J. H. Esling, "Laryngographic study of phonation type and laryngeal configuration," Journal of the International Phonetic Association, vol. 14(02), pp. 56–73, 1984.

[7] J. Laver, The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press, 1980.

[8] H. Hollien, J. Michel, and E. D. Thomas, "A method for analyzing, vocal jitter in sustained phonation," Journal of Phonetics, vol. 1, pp. 85–91, 1973.

[9] I. R. Titze, R. J. Baken, and H. Herzel, "Evidence of chaos in vocal folds vibration in Vocal fold physiology," in New Frontiers in Basic Science, I. R. Titze, Eds. San Diego: Singular Publishing Group, 1993, pp. 143–188.

[10] M. Rothenberg and J. J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," Journal of Speech and Hearing Research, vol. 31, pp. 338–351, 1988.

[11] N. Henrich, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," Journal of the Acoustic Society of America, vol. 115, pp. 1321–1332, 2004.

[12] D. G. Childers, D. M. Hicks, G. P. Moore, and L. Eskenazi, "Electroglottography and vocal fold physiology," Journal of Speech and Hearing Research, vol. 33, pp. 245–254, 1990.

[13] D. G. Childers and A. K. Krishnamurthy, "A critical review of electroglottography," Critical Reviews in Biomedical Engineering, vol. 12, pp. 131–161, 1985.

[14] D. G. Childers and J. N. Larar, "Electroglottography for laryngeal function assessment and speech analysis," IEEE Transactions on Biomedical Engineering, vol. 31, pp. 807– 817, 1984.

[15] D. G. Childers, G. P. Moore, J. M. Naik, J. N. Larar, and A. K. Krishnamurthy, "Assessment of laryngeal function by simultaneous, synchronized measurement of speech, electroglottography and ultra-high speed film," In Proceedings of the Eleventh Symposium on Care of the Professional Voice. New York, 1983.

[16] D. G. Childers, J. M. Naik, J. N. Larar, A. K. Krishnamurthy, and G. P. Moore, "Electroglottography, speech and ultra-high speed cinematography," In Vocal Fold Physiology and Biophysics of Voice, I. R. Titze and R.

Scherer, Eds. Denver: Denver Center for the Performing Arts, 1983, pp. 202–220.

[17] D. M. Howard, "Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers," Journal of Voice, vol. 9(2), pp. 163–172, 1995.

[18] D. M. Howard, G. A. Lindsey, and B. Allen, "Toward the quantification of vocal efficiency," Journal of Voice, vol. 4, pp. 205–212, 1990.

[19] J. Kong, On Language Phonation [Lunyuyanfasheng]. Beijing: Central Nationalities University Press, 2001.

[20] M. Rothenberg, "Some relations between glottal air flow and vocal fold contact area," In Proceedings of the Conference on the Assessment of Vocal Pathology. National Institutes of Health, 1979.

[21] K-I. Sakakibara, H. Imagawa, S. Niimi, and N. Tayama, "Physiological study of the supraglottal structure," In Proceedings of the International Conference on Voice Physiology and Biomechanics. Marseille, 2004.

Laryngeal Vibratory Behavior in Traditional Noh Singing*

YOSHINAGA Ikuyo, KONG Jiangping

Abstract

The phonation of Noh, a traditional Japanese style of singing, was investigated using electroglottographic and acoustical analyses. The dynamics of the laryngeal vibratory behaviors were analyzed for the singing voice of the Noh play compared with natural speech based on the electroglottography (EGG) parameters, EGG waveform, spectrum and spectrogram. The result shows that Noh singing is characterized by low OQ_{egg} and high SQ_{egg}. Three types of phonations are used in the singing with pressed, vocal-ventricular mode (VVM), and growl voices. It was hypothesized that the period doubling observed in the EGG signal was reflective of VVM, which was caused by the phase difference in the vocal and ventricular fold oscillations, while the damped peak amplitude in every other cycle in the EGG signal was the result of the oscillations of the aryepiglottic folds at a frequency of half of the fundamental frequency. Subharmonics generated by the supraglottal oscillations add unique timbre to the sounds. The results suggest that the combination of phonation types is the key factor in generating their peculiar voice qualities.

Key words:

pressed voice; vocal-ventricular mode (VVM); growl voice; electroglottography (EGG); open quotient (OQ_{egg}); speed quotient (SQ_{egg})

Introduction

Noh is a traditional performing art which has been handed down orally since the Nara Era. A typical Noh play involves a small chorus and orchestra, a shite (main role), and a waki (supporting role) wearing masks. The Noh artistic style has a solemn atmosphere. The Shite have 5 schools at present, among which the Hosho and Kanze school are most representative. The Hosho has an excellent reputation for its solid performance style and long history. From the oral arts viewpoint, the singing voice in Noh is influenced by Buddhist chants.

Research on Noh singing is quite rare. The singer's formant has been observed to be 3-4 kHz[1]. The duration of the consonants is lengthened in Noh singing compared with classical European singing. The Mahalanobis generalized distance on the LPC cepstrum was used to evaluate the distances between conversational and singing voices in Noh[2]. They showed that the distance between them is much closer than in classical European singing. Thus, expert singers use their characteristic phonation of singing even in their conversational speech.

Recent investigation using high-speed camera, X-ray, Kymography, and other methods has revealed ventricular and aryepiglottic fold oscillations in Asian vocal cultures and in some ethnic and pop music. The ventricular folds oscillate at speeds of the fundamental frequency (F_0), $F_0/2$ or $F_0/3$ in the vocal-ventricular mode (VVM), with the aryepiglottic folds oscillatery at $F_0/2$ in the growl voice (also called the voiced aryepiglottic trill[3])[4-6]. These kinds of phonations are found not only in the singing technique but also in the vocal fry, voice instabilities, and infant vocalizations. These irregular vocalizations are often interpreted as period-doubling bifurcations, with the corresponding acoustical signals often showing sudden jumps to subharmonic regimes[7,8].

Electroglottography (EGG) has been widely used to analyze vocal fold functions since Fabre reported on its use in 1957[9]. Five Chinese phonation types are described using EGG parameters[10]. Table 1 shows that the open quotient (OQ_{egg}) and speed quotient (SQ_{egg}) are key factors distinguishing the five phonation types. For example, the vocal fry is described by high OQ_{egg} and SQ_{egg}. This suggests that they are also very important parameters for voice quality assessments.

Figure 1 gives a simplified illustration which shows the relations between the voice qualities and physiological, EGG and spectral properties. The relationship between the EGG waveform and the frontal section of vocal folds is described according to Ref. [11]. The relationship between the superior view of vocal folds and the spectral tilt is determined with reference to Ref. [12]. Generally, low OQ_{egg} represents a pressed voice with more contact area in the vocal folds that can be seen at the frontal section of the vocal folds in Fig. 1. A high OQ_{egg} represents a breathy voice with less contact area in the vocal

* Published: Tsinghua Science and Technology, vol. 17(1), pp.94-103, 2012

folds, because a longer glottal release duration will result in more airflow. Low SQ_{egg} indicates the glottal closing is slower and the voice has less energy without the forceful glottal closure. This can be described as a steep spectrum tilt from an acoustical point of view. A high SQ_{egg} has a forceful, quick glottal closure and is described by a small spectrum tilt. Thus, these EGG parameters can be used to judge voice qualities.

Table 1 Distinctive features of the source parameters for the five Chinese phonation types compared with the modal. "–" indicates lower and "+" indicates higher than modal. A high pitched voice is indicated as "High" in this table.

	F_0	OQ_{egg}	SQ_{egg}
Fry	–	+	+
Breathy	–	+	–
Pressed	–	–	+
Modal	±	±	±
High	+	–	–

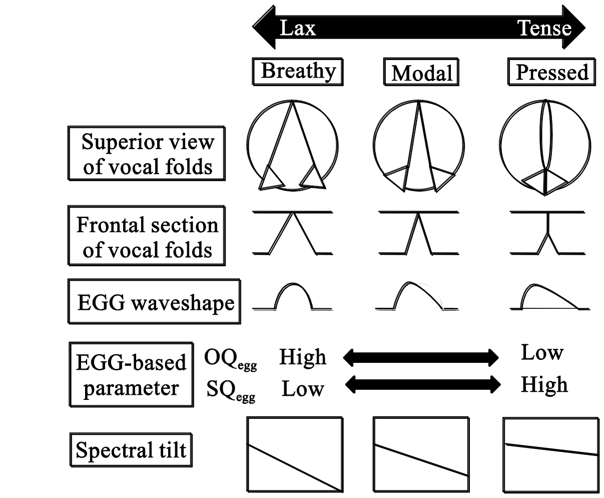


Fig. 1 Simplified illustration of the vocal folds, EGG waveforms, and parameters and the spectral tilt related with the phonation types

This study used electroglottographic and acoustic analyses to study the laryngeal vibratory behavior and to describe the phonatory characteristics of voice qualities in Noh singing.

1. Methods and Materials

This section describes the calculational method to get the EGG parameters, the details of the voice materials, and the processing procedure.

1.1 arameter calculation method

Voice quality is key to judging various singing styles. Perceptual assessment and a variety of instrumental (acoustical and physiological) methods are used in the definition of voice qualities[13]. EGG, which measures the electrical conductance changes between a pair of electrodes placed on the neck, is a noninvasive technique used to observe vocal fold vibratory patterns. The EGG signals provide meaningful information only when the vocal folds repeat contact and de-contact during vibration. Therefore, contact-based analysis is the common algorism. Because the EGG and airflow waveforms differ from each other qualitatively, OQ_{egg} and SQ_{egg} are employed in this study as the EGG-based parameters. A rise in the EGG signal corresponds to the closing of the glottis, while dropping corresponds to opening of the glottis as shown in Fig. 2.

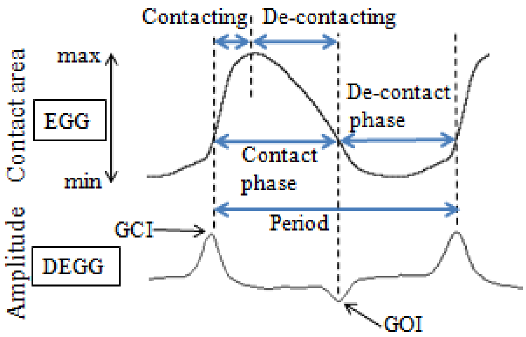


Fig. 2 EGG waveform and phases of vocal fold contact

Three EGG-based parameters are extracted: F_0, OQ_{egg}, and SQ_{egg}. The definitions of F_0 and OQ_{egg} are described as follows: F_0=1/period and OQ_{egg}%=de- contact phase/period*100%. Although the SQ_{egg} can be varied in detail across researchers, the definition used in this research is SQ_{egg}%=de-contacting/contacting* 100%[10].

There has been much discussion on the definition of the glottal closing instance (GCI) and glottal opening instance (GOI)[14-18]. Three kinds of proposed EGG calculational methods are the criterion-level algorithm[19], the differential of the EGG signal (DEGG)[15,20-24], and a combination of the criterion-level and the DEGG method, called the Hybrid algorithm[17,18]. The DEGG is considered the best method reflecting the GCI and GOI, but it is not reliable with imprecise or double GCIs and GOIs[15]. EGG waveforms of Noh singers often contain sub-cycles (Fig. 3).

In this case, the DEGG signals show double GCIs or GOIs, and the precise setting of the criterion level is necessary so that each instance can be detected. Therefore, the EGG cycle segmentations were set at 35% to define the GCIs or GOIs as shown in Fig. 4.

1.2 Voice material

The Noh singing voice was studied using 2 males and 2 females who were professional Noh singers of the Hosho school. The two males (subjects A and B) were successors of Yoshio Hosho, a living national treasure. They started singing Noh at the age of 4. The females (subjects C and D) also started singing at an early age. They were all in their twenties with about 20 years of singing experience when the recordings were made.

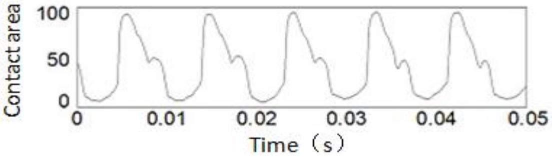


Fig. 3 5-cycle EGG waveform of vowel /e/ at G3 (196 Hz) from Tsurukame (Subject A)

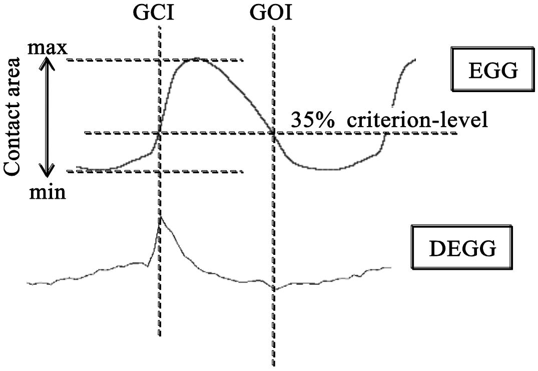


Fig. 4 Definitions of GCI and GOI on the EGG signal by the 35% criterion and a differential method

The voice materials consist of (1) the Noh play “Tsurukame (crane and turtle)” and (2) the sustained vowels /a, e, i, o, u/. Tsurukame is a story describing the New Year’s prayer at the palace of Emperor Xuanzong of the Tang Dynasty. The subjects read and sang the lyrics of Tsurukame in their comfortable pitch. In this case, gaps of F_0 and intensity changes occurred between the singing and speaking. Since F_0 and intensity influence OQ_{egg} and SQ_{egg}, this method can’t be used to compare OQ_{egg} and SQ_{egg} between singing and speaking with the first set of materials. The subjects spoke 5 sustained vowels /a, e, i, o, u/ using speaking and Noh singing styles. They sustained every semitone from their lowest pitch to the highest and they were requested to use the same loudness in singing and speaking. Thus, the F_0 gap between singing and speaking and the intensity differences were eliminated in case 2. In this case, the recordings covered a range of over an octave, starting at about F2# (92 Hz) for the males and E3 (164 Hz) for the female subjects. Since the recordings for case 2 covered their entire pitch range, it contained the pitch height which is not used in actual Noh singing. Moreover, the sustained phonation did not contain some of the Noh singing techniques which were used in the actual singing. However, comparison of the singing and speaking using the case 2 material will illustrate the inherent voice qualities of Noh singing phonation. The phonation types of actual Noh singing were studied in case 1.

The recordings took place at the Nohgaku stage in Tokyo, Japan. The EGG signal was obtained on an EGG by LARYNGOGRAPH Ltd BF; Kay, UK. The audio signal was recorded by a Sony Electret Condenser Microphone. These signals were simultaneously recorded and digitized on 16 bits at a sampling frequency of 44.1 kHz.

1.3 Data processing

The recorded files were prepared for the acoustical analysis by downsampling to 11.025 Hz. Then, the EGG rumble caused by the up and down laryngeal movements was filtered out by a high-pass cutoff frequency of 60 Hz, because it could confuse the calculated results. The files were divided into smaller pieces to prepare for the batch processing to obtain the EGG parameters. Since a large amount of data processing was needed and the recorded files were different lengths, 30 parameter values were extracted from each file. Wavelet transforms were applied to each file to adjust the noise and warp of the EGG signals, which might also cause miscalculations, before extracting the EGG parameters. The data processing was performed using the Matlab based SpeechLab developed by the Linguistic Lab of Peking University.

Since the singing data files were 11 min long which was twice the length of the speaking data, the parameter were extracted using 1140 data points from each singing and speaking file. For case 2, 30 parameter values were extracted for each sustained vowel.

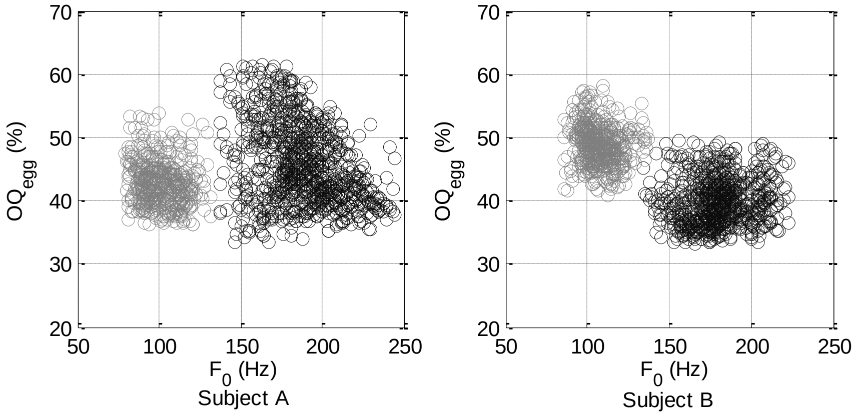
2 Parameter Analysis

The EGG parameters were then analyzed by comparing the singing and speaking parameters to study the Noh singing voice quality.

2.1 Pitch range analysis for Tsurukame

The F_0 distributions for the singing and speaking of Tsurukame are shown in Figs. 5 and 6. The discussion is limited to F_0 to focus on the pitch range. The F_0 distributions for male subjects A and B are shown in Fig. 5, while the F_0 distributions for female subjects C and D are shown in Fig. 6. The singing parameters are shown by 1140 black circles, while the speaking parameters are shown by 1140 gray circles in Figs. 5 and 6. Table 2 lists the average, minimum, and maximum of F_0 for each subject. Semitone as well as hertz was used to measure F_0. Figures 5 and 6 show a large difference in F_0 between the singing and speaking. F_0 for the singing is significantly higher and its range is wider than for speaking with all the subjects. The F_0 range in the males is 11-12 semitones for singing, and 6 semitones for speaking. Thus, the singing range is almost double the speaking range. The F_0 range for the females is 11-12 semitones for singing and 4-5 semitones for speaking. Thus, the singing range is more than double the speaking range. The F_0 range for singing is then wider than that for speaking regardless of gender. The average F_0 of the males is about 182-188 Hz for singing and 103-108 Hz for speaking. Thus, the F_0 for the singing is 9-10 semitones higher than that of speaking. The average F_0 for the females is 288 Hz for singing and 193-198 Hz for speaking. Thus, the F_0 of singing is 7 semitones higher than that of speaking for the females.

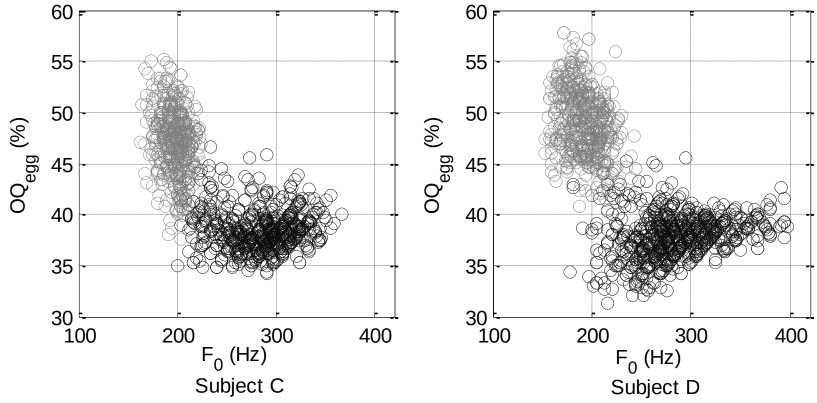
Therefore, the F_0 range for singing is 11-12 semitones, almost double that of speaking. The average F_0 for singing is 9-10 semitones higher in males and 7 semitones higher in females than that of speaking. Though singing voice in Noh sounds quite low, the F_0 of singing actually rather high than that of speaking.



Subject A

Subject B

Fig. 5 Distribution of F_0 for singing (black) and speaking (gray) of Tsurukame by male.



Subject C

Subject D

Fig. 6 Distribution of F_0 for singing (black) and speaking (gray) of Tsurukame by female.

Table 2 Average, minimum, and maximum of F_0 in Tsurukame

Subject	Average (Hz)		Min (Hz)		Max (Hz)	
	Singing	Speaking	Singing	Speaking	Singing	Speaking
A	187.7	102.7	89.2	73.4	275.6	149.0
B	181.5	107.7	95.9	74.6	268.9	144.8
C	288.1	197.7	190.1	160.3	376.2	232.6
D	288.2	193.1	161.0	141.7	419.7	263.4

2.2 EGG parameter analysis: Sustained vowel

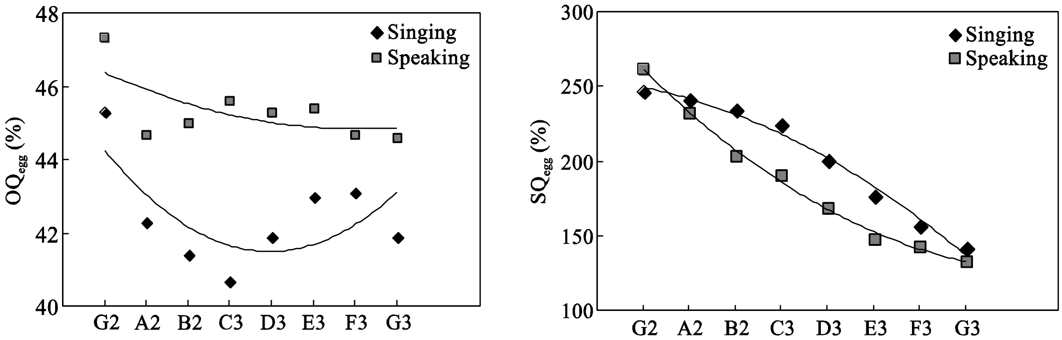


Fig. 7 Average of OQ_{egg} (left) and SQ_{egg} (right) for an octave (males)

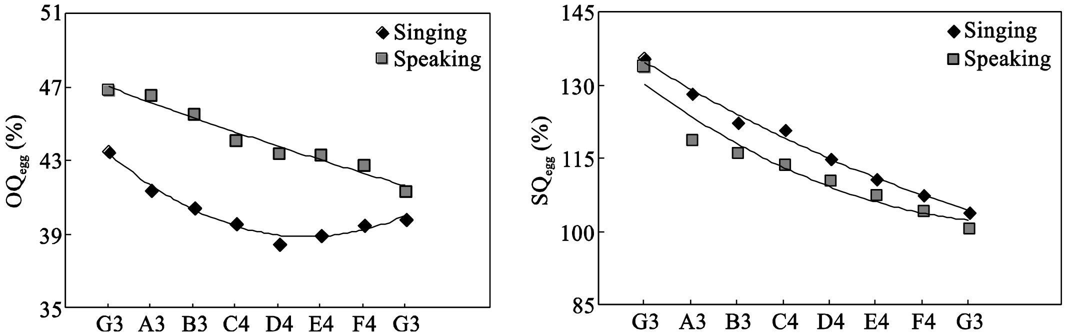


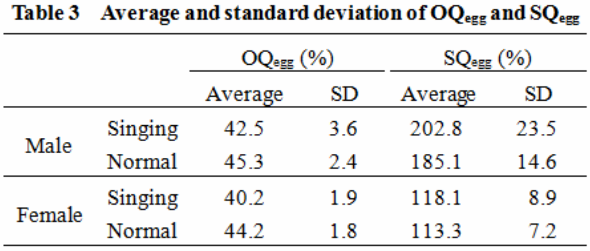
Fig. 8 Average of OQ_{egg} (left) and SQ_{egg} (right) for an octave (females)

The pitch range shown in Figs. 7 and 8 covers 1 octave from G2 (98 Hz) to G3 (196 Hz) for the males in Fig. 7, and from G3 (196 Hz) to G4 (392 Hz) for the females in Fig. 8. The average of OQ_{egg} and SQ_{egg} is given in black for singing and in gray for speaking. Polynomial fitting curves are also given in Figs. 7 and 8.

The data in Figs. 7 and 8 shows that OQ_{egg} of singing is lower than that of speaking both in males and females, while SQ_{egg} of singing is higher than that of speaking. Thus, singing is characterized by low OQ_{egg} and high SQ_{egg}. The OQ_{egg} of speaking tends to drop steadily as F_0 increases. For the singing, the fitted polynomial curves for OQ_{egg} are rather arched with the lowest OQ_{egg} at C3 (131 Hz) for the males and D4 (294 Hz) for the females. The lowest OQ_{egg} for the males reaches 40.7% while that for the females reaches 39.6% with a maximum 5% gap between singing and speaking. Thus, very strong adduction of the vocal folds occurs during singing, especially around C3 for the males and D4 for the females. SQ_{egg} for both singing and speaking tends to drop steadily as F_0 increases. Thus, Noh singing can be characterized as low OQ_{egg} and high SQ_{egg}.

The overall averages and standard deviations, SD, are given in Table 3. The average OQ_{egg} for singing is 42.5% for the males and 40.2% for the females, which is 3%-4% lower than that for speaking. Strong adduction of the vocal folds with longer glottis closures is achieved in singing. The average SQ_{egg} for singing is 202.8% for the males and 118.1% for the females, which is 5%-18% higher than those for speaking. In this case, singing involves quick and forceful glottis closure. The standard deviation of OQ_{egg} is higher in singing, reaching 3.6% in males. The standard deviation of SQ_{egg} is also quite different between singing and speaking especially in males, with 23.5% for singing and 14.6% for speaking. These results illustrate the dynamic activities occurring in the glottis during Noh singing.

Table 3 Average and standard deviation of OQ_{egg} and SQ_{egg}



		OQ_{egg} (%)		SQ_{egg} (%)	
		Average	SD	Average	SD
Male	Singing	42.5	3.6	202.8	23.5
	Normal	45.3	2.4	185.1	14.6
Female	Singing	40.2	1.9	118.1	8.9
	Normal	44.2	1.8	113.3	7.2

It is difficult to represent each parameter linearly, because natural human voices always have source variation. However these attempts can help recognize the inherent characteristics of the voice source parameters. The low OQegg and high SQegg are the key features of Noh singing phonation.

The source characteristics can be described in terms of the glottal pulse shapes in the time domain, such as in the LF model[25], and can also be described in terms of their spectral effects in the frequency domain. The shape of the glottal source is an important determinant of voice quality[26,27]. Figure 9, which is based on the LF model, shows a schematic of the glottal flow and the differentiated glottal flow for speaking and Noh singing. UP and EE are equivalent to the peak volume velocity of the glottal pulse and the excitation strength. A low OQ_{egg} and a high SQ_{egg} in the EGG, which are characteristics of Noh singing, are related to a small UP and high EE. Thus, Noh singing has a small peak volume velocity of the glottal pulse with a short glottal release while a strong excitation strength with forceful glottis closure. Thus, Noh singing characterized by the low OQ_{egg}, and high SQ_{egg} can be assessed as a pressed voice.

3. Phonation Analysis

In this section, the Noh singing phonation types are investigated using the EGG waveforms, EGG parameters, and acoustic spectra and spectrograms.

3.1 EGG parameter analysis for each phonation type

Actual singing phonation involves more voice qualities than just sustained phonation, because it has rhythm and emotional factors. The Noh singing voice was classified as pressed phonation in Section 2 by analyzing the singing and speaking parameters. The voice quality of the actual singing voice from Tsurukame is assessed here.

Some Asian oral characteristics involve oscillations of the ventricular folds and aryepglottic folds. The EGG waveform for VVM includes periods doubling[4,28,29], while growl is described as a damped amplitude every other cycle[29]. The period doubling is caused by the phase difference between the vocal and ventricular fold oscillations. When the ventricular folds oscillate at $F_0/2$ or $F_0/3$, subharmonics appear at $F_0/2$ or $F_0/3$ in the spectrum or spectrogram[4,29]. When the growl voice is analyzed by X-ray videofluoroscpy[6,29], the larynx position is usually high, the aryepglottic region is compressed antero-posteriorly, and the tubercle of the epiglottis and arytenoid cartilages come into contact. The corresponding EGG waveform indicates less contact area in every other cycle[6]. These supraglottal vibrations are also thought to be used in Noh singing. The EGG waveforms for VVM and growl observed in Noh singing are shown in Fig. 10. The period doubling is found in VVM while damped amplitudes in every other cycle in growl.

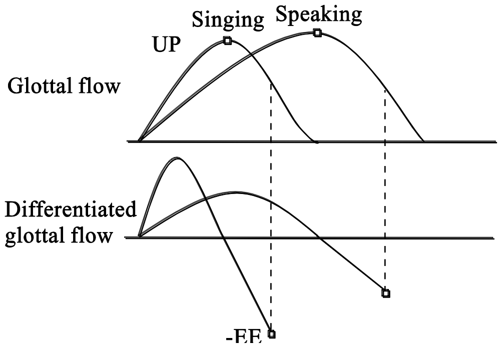


Fig. 9 Schematic of glottal flow and its derivative for singing and speaking

The three kinds of EGG waveforms observed in Noh singing are growl, VVM, and pressed voice. Growl appears in the initial part of almost every phrase, while VVM and pressed voice appear during the rest of the sound. The F_0 distributions for each phonation type are shown in Fig. 11. F_0 of growl is the lowest and the range is the narrowest with an average F_0 of

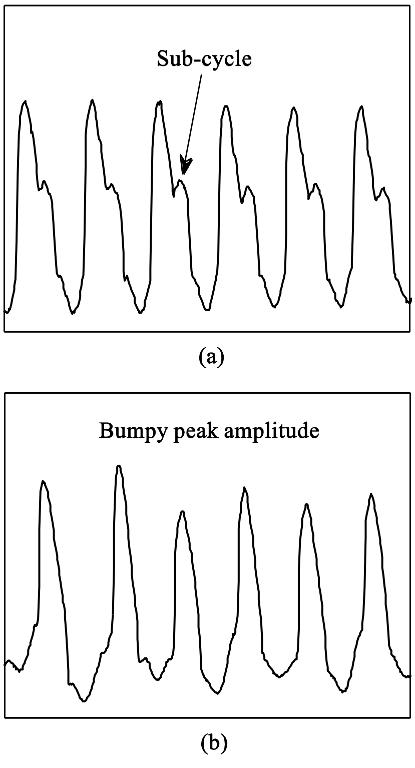


Fig. 10 6 cycle EGG waveform for (a) VVM and (b) growl from Tsurukame (subject A)

141.6 Hz for the males and 220.0 Hz for the females. The VVM frequency is in the middle with an average F_0 of 165 Hz for the males and 249 Hz for the females. The F_0 of the pressed voice is the highest and the range is the widest. Overlap occurs between each phonation type since each type covers a rather wide range, while it implies that switching between phonation types involves not only the pitch height but also techniques to add certain vocal effects to the singing voice. Thus, the F_0 are characterized as: growl < VVM < pressed.

The OQ_{egg} variations in Fig. 12 show that growl has the highest OQ_{egg}, then pressed, and VVM with the lowest OQ_{egg}. The average OQ_{egg} for growl is 53.0% for the males and 45.5% for the females. The average OQ_{egg} for VVM is 39.4% for the males and 36.7% for the females. There is a significant difference between OQ_{egg} for growl and that for the other phonation types. For example, OQ_{egg} for growl is 9%-14 % higher than that for VVM. Moreover, though F_0 for growl and the other phonation types overlap as shown in Fig. 11, less overlap occurs for OQ_{egg} in both males and females as shown in Fig. 12. Thus, OQ_{egg} is characterized as: VVM < pressed < growl.

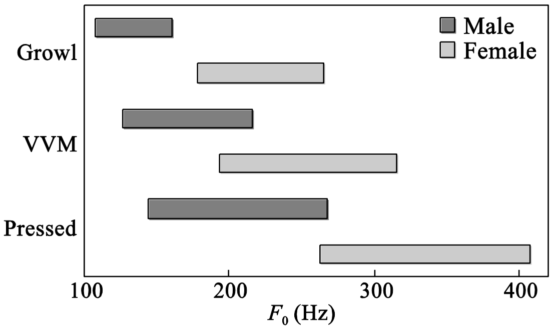


Fig. 11 F_0 Distributions for each phonation type

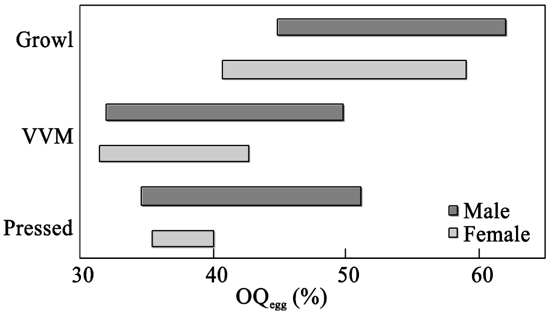


Fig. 12 OQ_{egg} distribution for each phonation type

The SQ_{egg} for VVM is significantly higher than that of growl or pressed voices in both males and females as shown in Fig. 13 and covers a range of 170.2% in the male subjects. Though the F_0 range of the pressed voice is the widest among the three phonation types (Fig. 11), the SQ_{egg} variation is the most stable as shown in Fig. 13. Therefore, the vocal fold vibrations for the pressed voice can be characterized as stable, while the VVM involves dynamic laryngeal behavior result in very unstable SQ_{egg}. SQ_{egg} can be characterized as: pressed < growl < VVM.

The distinctive features of each phonation type are shown in Table 4. As noted in Section 3, Noh singing is classified as pressed phonation with low OQ_{egg} and high SQ_{egg}, whereas the growl voice has a peculiar phonation quality with high OQ_{egg} and low SQ_{egg} caused by less contact area of the vocal folds resulting from the aryepiglottic fold oscillations. Although the EGG parameters indicate a significant difference between growl and the other two phonation types, the supraglottal oscillation is a common feature in both growl and VVM. VVM is characterized by the lowest OQ_{egg} and the highest SQ_{egg}. As a result, VVM and pressed voice are classified as the pressed type phonation. VVM is characterized as the more pressed, energetic and dynamic phonation type. On the other hand, growl is characterized as a rather lax phonation type.

3.2 Switching of phonation types and acoustic analysis

Subharmonics are often found in voice instabilities, infant vocalizations, some paralinguistic features of speaking, and singing techniques[4,7,8,29]. Subharmonics are also observed in Noh singing. Figure 14 shows the switching of

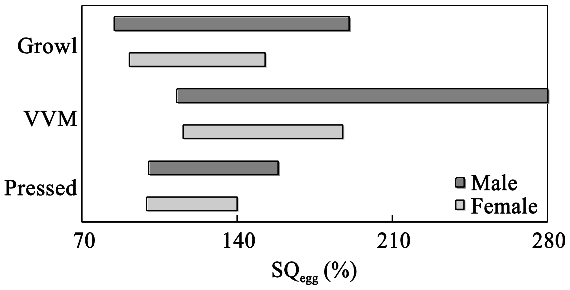


Fig. 13 SQ_{egg} distribution for each phonation type

Table 4 Distinctive features of the source parameters for the three phonation types. “-” indicates low, “+” means high, “±” is middle.

	F_0	OQ_{egg}	SQ_{egg}
Growl	−	+	±
VVM	±	−	+
Pressed	+	±	−

phonation types in actual Noh singing, (a) F_0 (indicated in black), (b) OQ_{egg} (indicated in gray) and SQ_{egg} (indicated in black), and (c) corresponding narrow-band spectrogram. The parameters are all normalized. Figures 14a and 14b provide a typical example of each phonation type's source features with growl having high OQ_{egg} and low SQ_{egg} and VVM having low OQ_{egg} and high SQ_{egg}. Subharmonics are observed both in spectrograms in Fig. 14c and the spectrum in Fig. 15 in the growl region. As seen in Fig. 15, the 63 Hz undertone is relatively weaker than the original harmonic peak at 126 Hz with the spectrum including clear subharmonics, indicated by the black arrows in Fig. 15, up to the high frequency region.

The frequency of the ventricular fold oscillations is the same as F_0 in both males and females, with $F_0/2$ also observed in females. Clear subharmonics appear due to the ventricular fold oscillations at $F_0/2$ in Fig. 16.

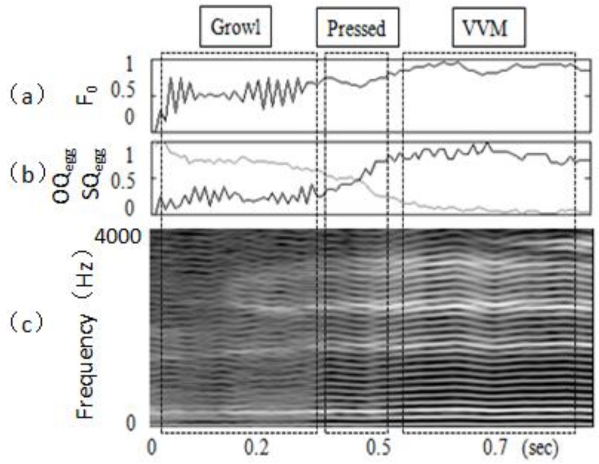


Fig. 14 Switching of phonation types in Tsurukame (subject A) (a) F_0 (black), (b) OQ_{egg} (gray) and SQ_{egg} (black), and (c) spectrogram

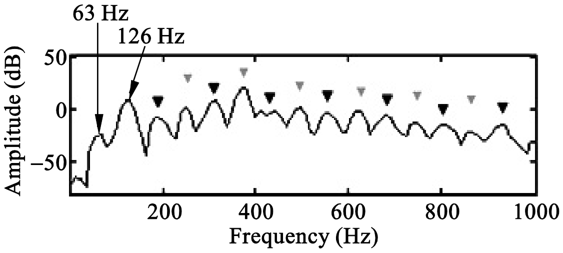


Fig. 15 Spectrum (range 0-1000 Hz) of growl from subject A. Black arrows indicate subharmonic peaks, gray arrows indicate original harmonic peaks.

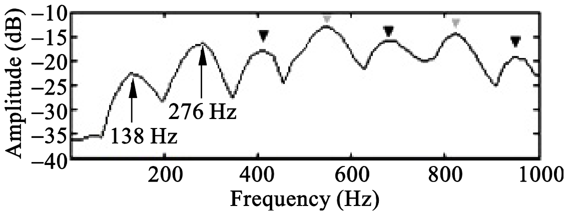


Fig. 16 Spectrum (range 0-1000 Hz) of VVM from subject C with an undertone observed at 138 Hz

F_0 is 276 Hz, with the subharmonics at 138 Hz which is equivalent to $F_0/2$ and other harmonics at higher frequencies.

A vibrato example from Tsurukame is shown in Fig. 17. Since the main perceptual effect of the vibrato depends on the frequency modulation[30], in Noh F_0 also shows sinusoidal modulation in Fig. 17a. Period doubling is observed in the EGG signal in Fig. 17d in the low pitch region in Fig. 17a with low OQ_{egg} and high SQ_{egg} in Fig. 17b. Thus, the vibrato involves not only the F_0 modulation but also a combination of phonation types.

Thus, vibrato is achieved by the combination of F_0 modulation and the switching of phonation types. Some VVM and growl subharmonics add low pitched impressions to the sounds.

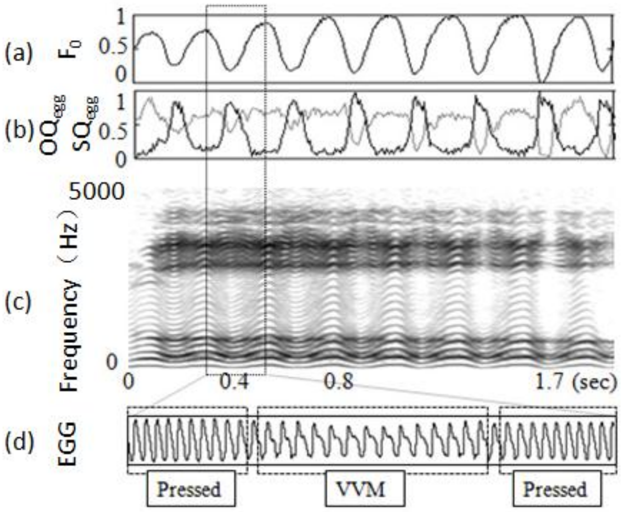


Fig. 17 Vibrato from subject A. (a) F_0 (black), (b) OQ_{egg} (gray) and SQ_{egg} (black), (c) spectrogram, and (d) selected EGG signals

4. Conclusions

This study analyzed the phonation mechanisms of Japanese traditional Noh using physiological and acoustical methods.

Three types of phonation occur in Noh singing: pressed, VVM and growl. Among the so-called extended vocal techniques, VVM is the representative Asian traditional vocal technique which sounds solemn and ancient, while growl is the animalized sound which delivers passionate and dramatic vocal effects to the singing. Noh singing is characterized by low pitched sounds, however, the actual F_0 is rather high and the pitch range is rather wide. Subharmonics, generated by the ventricular and aryepiglottic fold oscillations at a frequency of $F_0/2$, add the low pitched sound effect to the singing. The unique mixed vocal timbre resulting from the combination of phonation types creates a compelling effect to Noh singing.

Further physiological measurements using other techniques such as high-speed cameras are needed to clarify the laryngeal behaviors of these peculiar phonations. Synthesize and perceptual evaluations are also expected in future work.

5. Acknowledgements

The authors thank Yoshio Hosho of the living national treasure from the Hosho school of Noh, his pupils and Mr. and Mrs. Ueda from Toshiba for the recordings and valuable advice on this study.

6. References

[1] Nakayama I. Comparison of vocal expressions between Japanese traditional and western classical-style singing, using a common verse. *The Journal of the Acoustical Society of Japan*, 2000, **56(**5): 343-348.

[2] Yoshioka N, Nagahata D, Yanagida M, et al. Differences among vowels used in Noh, Kyogen and European classical singing. Technical report of IEICE, 2001, **122**: 1-8.

[3] Esling J H. Pharyngeal consonants and the aryepiglottic sphincter. *Journal of the International Phonetics Association*, 1996, **26**(2): 65-88.

[4] Fuks L, Hammarberg B, Sundberg J. A self-sustained vocal-ventricular phonation mode: Acoustical, aerodynamic and glottographic evidences. *KTH TMH-QPSR*, 1998, **3**: 49-59.

[5] Lindestad P-Å, Södersten M, Merker B, et al. Voice source characteristics in mongolian “throat singing” studied with high-speed imaging technique, acoustic spectra, and inverse filtering. *Journal of Voice*, 2001, **15**(1): 78-85.

[6] Sakakibara K, Fuks L, Imagawa H, et al. Growl voice in ethnic and pop styles. In: Proceedings of the International Symposium on Musical Acoustics. Nara, Japan, 2004.

[7] Hollien H, Michel J, Thomas E D. A method for analyzing, vocal jitter in sustained phonation. *Journal of Phonetics*, 1973, **1**: 85-91.

[8] Titze I R, Baken R J, Herzel H. Evidence of chaos in vocal folds vibration in vocal fold physiology. In: New Frontiers in Basic Science. San Diego: Singular Publishing Group, 1993: 143-188.

[9] Fabre P. An electrical inscription of the glottis during phonation: glottography from high frequency. *Journal of the National Academy of Medicine*, 1957, **141**: 66-69. (in French)

[10] Kong Jiangping. On Language Phonation. Beijing: Central Nationalities University Press, 2001: 173-189. (in Chinese)

[11] Titze I R. Interpretation of the electroglottographic signal. *Journal of Voice*, 1990, **4**: 1-9.

[12] Stevens K N. Physics of larynx behavior and larynx modes. Phonetica, 1977, **34**: 264-279.

[13] Kent R D, Ball M J. Voice Quality Measurement. California: Singular Publishing Group, 2000: 117-118.

[14] Baken R J, Orlikoff R F. Clinical Measurement of Speech and Voice. New York: Delmar learning, 2000: 413-427.

[15] Henrich N. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of Acoustic Society of America*, 2004, **115**: 1321-1332.

[16] Herbst C. Evaluation of various methods to calculate the EGG contact quotient [Dissertation]. Stockholm: KTH Speech, Music and Hearing, 2004.

[17] Howard D M, Lindsey G A, Allen B. Toward the quantification of vocal efficiency. *Journal of Voice*, 1990, **4**: 205-212.

[18] Howard D M. Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. *Journal of Voice*, 1995, **9**(2): 163-172.

[19] Rothenberg M, Mahshie J J. Monitoring vocal fold abduction through vocal fold contact area. *Journal of Speech and Hearing Research*, 1988, **31**: 338-351.

[20] Childers D G, Hicks D M, Moore G P, et al. Electroglottography and vocal fold physiology. *Journal of*

Speech and Hearing Research, 1990, **33**: 245-254.

[21] Childers D G, Krishnamurthy A K. A critical review of electroglottography, *Critical Reviews in Biomedical Engineering*, 1985, **12**: 131-161.

[22] Childers D G, Larar J N. Electroglottography for laryngeal function assessment and speech analysis. *IEEE Transactions on Biomedical Engineering*, 1984, **31**: 807-817.

[23] Childers D G, Moore G P, Naik J M, et al. Assessment of laryngeal function by simultaneous, synchronized measurement of speech, electroglottography and ultra-high speed film. In: The Eleventh Symposium: Care of the Professional Voice. New York: Voice Foundation, 1983: 234-244.

[24] Childers D G, Naik J M, Larar J N, et al. Electroglottography, speech and ultra-high speed cinematography. In: Vocal Fold Physiology and Biophysics of Voice. Denver: Denver Center for the Performing Arts, 1983: 202-220.

[25] Fant G, Liljencrants J, Lin Q. A four parameter model of glottal flow. *STL-QPSR*, 1985, **4**: 1-13.

[26] N í Chasaide A, Gobl C. Voice Source Variation. The Handbook of Phonetic Sciences. Oxford: Blackwell Publishers Ltd, 1997: 427-461.

[27] Fant G. Speech Acoustics and Phonetics. Boston: Kluwer Academic Publishers, 2004: 249-300.

[28] Esling J H. A laryngographic investigation of phonation type and laryngeal configurations. Working papers of the Linguistic circle. Canada: University of Victoria, 1983: 14-36.

[29] Sakakibara Ken-Ichi, Imagawa Hiroshi, Niimi Seiji, et al. Physiological study of the supraglottal structure. In: Proc. International Conference on Voice Physiology and Biomechanics. Marseille, France, 2004.

[30] Johan Sundberg. The Science of the Singing Voice. Illinois, USA: Northern Illinois University Press, 1987: 163-176.

Some Phonatory Characteristics of Tibetan Buddhist Chants*

YOSHINAGA Ikuyo, KONG Jiangping

要旨

歌声を形成する言語情報（歌詞）と非言語情報（音高・音長情報を含む旋律及び声質）のうち非言語情報である声質に焦点を当て、電気声門図及び音声信号を基にチベット声明の発声における特徴の解明を試みた。その結果、声明の音源パラメータは低声門開放率及び高声門開閉速度率の特徴をもち、それらに対応する音響特徴として H1-H2 及び H1-A3 が共に低い数値を示した。これらは声明におけるりきみ発声の特徴を表している。次に、電気声門図波形及びスペクトル解析により、単なるりきみだけでなく、声門上構造物の振動によると推測されるザラザラ感（harsh）のある発声も確認された。ここではその周波数は声帯振動と同じ F0 であった。声明における音域は通常発話時の半分の音域に値する 2 半音であり、音高も発話時より 2 半音低めであった。このような特徴からチベット声明は喉詰型発声の伝統を汲むことがわかった。

Keywords

pressed voice, harsh voice, supraglottal constriction, electroglottography (EGG), open quotient (OQ_{egg}), speed quotient (SQ_{egg})

1. INTRODUCTION

Tibetan lamas chant sutras with devout devotion and passion. ‘Outsiders’ listening to these chants often cannot help but be impressed by the unique sounds of their low sonorous pitch. *Shomyo* (sabda-vidya, in Sanskrit), as it has been called, was one of the five fields of academic study in ancient India and was deeply treasured and successfully handed down by Tibetan Buddhists.

In ‘throat singing’, Mongolian *Kargyraa* is the common label for low, bass-pitched singing, and a similar style is found in Tibetan Buddhist chants (Lindestad et al. 2001, Sakakibara 2003). A study conducted in the 1960s used sonograms to hypothesize that the ‘odd harmonics’ found in the chants of Tibetan lamas were produced by double oscillators or asymmetrically vibrating vocal folds (Smith et al. 1967). High-speed video endoscopy and electroglottography (EGG) of non-Tibetan ‘throat singing’ revealed that the ventricular folds oscillated at half of the frequency of vocal folds in a typical phonation mode, which was judged to be perceptually identical to that used in Tibetan Buddhist chants (Fuks et al. 1998). In Mongolian ‘throat singing’, the ventricular fold vibrations were observed via high-speed imaging techniques and kymography (Lindestad et al. 2001). The ventricular folds oscillate at a frequency of F0, F0/2, or F0/3 in vocal-ventricular mode (VVM) (Fuks et al. 1998, Lindestad, et al. 2001, Sakakibara, et al. 2004). These supraglottic phonations have been found not only in singing techniques but also in vocal fry, voice instabilities, and infant vocalizations. These irregular vocalizations are often interpreted as period-doubling bifurcations, and the corresponding acoustical signals often show sudden jumps to subharmonic regimes (Hollien et al. 1973, Titze et al. 1993).

The phonatory characteristics of voice qualities are very important in defining singing techniques. However, perceptual assessments of voice qualities remain ambiguous. Objective assessments, such as acoustical analysis, synthesis, and physiological observation, are needed (Sakakibara 2003).

This paper describes the electroglottographic and acoustic analyses conducted to reveal the voice production mechanisms of certain singing modes of Tibetan Buddhist chants and describes the phonatory characteristics of these voice qualities.

2. METHODS AND MATERIALS

This section describes EGG, the primary experimental method of this research, the calculation method of EGG-based parameters, voice materials, and the data processing procedure.

2.1 Electroglottography

Voice quality is a key issue in describing various singing styles. Perceptual assessment and a variety of instrumental (acoustical and physiological) methods are applied in the definition of voice qualities (Raymond and Martin 2000). EGG, which measures electrical conductance changes between a pair of electrodes placed on the neck, is a noninvasive technique for the observation of vocal fold vibratory patterns. One of the

*Published: Journal of the Phonetic Society of Japan, vol. 15(2), pp. 83-90, 2011.

authors has described five Chinese phonation types using EGG parameters. Table 10 shows that the open quotient (OQ_{egg}) and the speed quotient (SQ_{egg}) are key factors in distinguishing these five phonation types: vocal fry, breathy voice, pressed voice, modal voice, and high-pitched voice. For example, vocal fry is characterized as high OQ_{egg} and SQ_{egg}. These parameters are very important to voice quality assessments.

Table 10 Distinctive features of the source parameter in five phonation types compared with the modal voice. "–" indicates lower and "+" indicates higher than the modal voice (Kong 2001).

	Fry	Breathy	Pressed	Modal	High
F0	–	–	–	±	+
OQ_{egg}	+	+	–	±	–
SQ_{egg}	+	–	+	±	–

Fig. 1 Simplified illustration of the vocal folds, EGG waveform and parameter, and spectral tilt related to the phonation type.

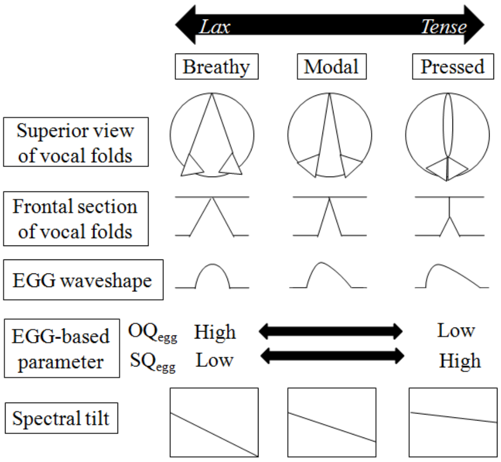


Fig. 1 is a simplified illustration that shows the relationships of voice qualities to physiological, electroglottographic, and spectral properties. The relationship between the EGG waveform and its corresponding frontal section of vocal folds is described according to Titze (1990). The relationship between the superior view of vocal folds and its spectral tilt is determined with reference to Stevens (1977). Low OQegg represents pressed voice with the larger contact area of the vocal folds that can be seen at the frontal section of the vocal folds in Fig. 1. In contrast, high OQegg represents breathy voice because more airflow is released with the longer de-contacting duration. A voice with lower SQegg has weaker energy because of the reduced speed when the vocal folds come into contact. Acoustically, this is reflected by the steeper spectral tilt. In contrast, higher SQegg has more forceful and quicker glottal closure and is accompanied by a more gradual spectral tilt. Thus, the characteristics of these EGG parameter values are reflected in the acoustic features.

2.2 Parameter Calculation Method

The EGG signals provide meaningful information only when the vocal folds repeat contact and de-contact during vibration. Therefore, contact-based analysis is the common algorism. A few parameters can be extracted from the EGG waveform that roughly correspond to the open quotient (OQ) and speed quotient (SQ). Because the EGG and airflow waveforms differ from each other qualitatively, OQ_{egg} and SQ_{egg} are employed in this study as the EGG-based parameters. Fig. 2 shows that a period of EGG signal can be divided into contact and de-contact phases. Furthermore, the contact phase can be divided into contacting and de-contacting.

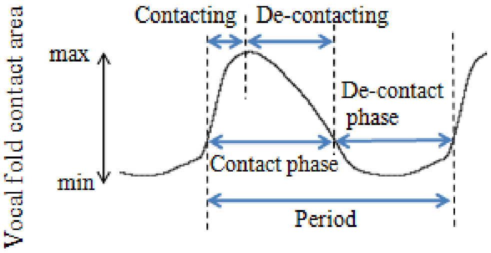


Fig. 2 EGG waveform and phases of vocal fold contact.

Three EGG-based parameters are extracted: F0, OQ_{egg}, and SQ_{egg}. The definitions of F0 and OQ_{egg} are described as follows: F0=1/period and OQ_{egg}%=de-contact phase/period*100. Although the SQ_{egg} can be varied in detail across researchers, the definition used in this research is SQ_{egg}%=de-contacting/contacting*100 (Kong 2001).

There have been discussions on the definition of the glottal closing instance (GCI) and glottal opening instance (GOI) (Baken and Orlikoff 2000, Henrich 2004, Herbst 2004, Howard et al. 1990, Howard 1995). Three kinds of EGG calculation methods are proposed, i.e., criterion-level (Rothenberg 1988), derivative of the EGG signal (DEGG) (Henrich 2004, Childers, Hicks, Moore and Eskenazil 1990, Childers, Moore, Naik, Larar and Krishnamurthy 1983, Childers, Naik, Larar, Krishnamurthy and Moor 1983, Childers and Krishnamurthy 1985, Childers and Larar 1984) and the combination of the criterion-level and DEGG methods, called the hybrid method (Howard et al. 1990, Howard 1995). The DEGG is considered the ideal method to reflect the GCI and GOI, but it is not reliable in the case of imprecise or multiple GCIs and GOIs (Henrich 2004). The EGG waveform in the data from Tibetan chants demonstrates period-doubling phenomena (Fig. 3). In this case, the DEGG signals show double GCIs or GOIs, and the precise setting of the criterion level is necessary so that each instance can be detected. Therefore, the criterion-level method of the 35% threshold is employed in this study, in which the threshold level is determined between the maximum and minimum values of the EGG waveform (Fig. 4).

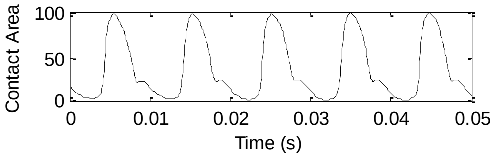


Fig. 3 Five vibratory cycles of the EGG signal from vowel /a/ phonated at G2 (98 Hz).

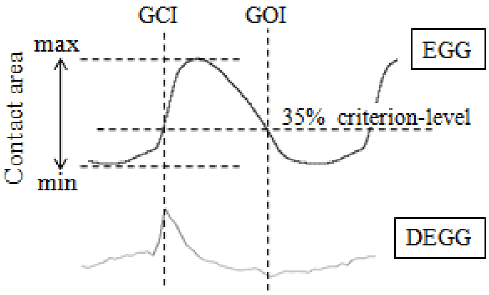


Fig. 4 Definition of GCI and GOI with a 35% criterion-level and derivative EGG (DEGG) waveform.

2.3 Voice Material

The phonation of Tibetan Buddhist chants was studied in one male monk from *Kumbum* Monastery of *Dge-Lugs-Pa*, which is one of the best monasteries in China. The monk was 31 years old, with 18 years of priestly experience, when the recording was performed. He was also a teacher at the monastery with an excellent reputation for his chanting.

The voice materials consist of two types: 1) a sutra, *Gadanlajima*, and 2) sustained vowels /a, e, i, o, u/. Gadanlajima is a representative sutra in the Kumbum Monastery. The subject chanted and read the sutra at his comfortable pitch.

The subject sustained five vowels /a, e, i, o, u/ using two styles, chanting and speaking, in material 2. Each semitone was produced from the lowest to highest for the subject's range while attempting to maintain the same volume in chanting and speaking. Thus, the factors that might influence the values of source parameters were eliminated.

The data acquisition took place at Kumbum Monastery in Qinghai province, China. The EGG signal was obtained by an EGG system (Electroglottograph Model 6103; Kay, USA). The audio signal was recorded by a Sony Electret Condenser Microphone. Those signals were simultaneously recorded and digitized at 16-bit resolution at a sampling frequency of 44.1 kHz.

2.4 Data Processing

To prepare the recorded files for acoustical analysis, the files were down-sampled to 11.025 kHz. Next, the EGG rumble, which was caused by up and down laryngeal movements, was filtered out by a high-pass filter with the cutoff frequency set at 60 Hz because it could affect or mislead the parameter extraction. The files were divided into smaller pieces in preparation for the batch processing to obtain the value of the EGG parameters. The parameter values for all of the cycles were extracted using the criterion-level method of 35% and were saved in an Excel file. Because a large amount of data processing was needed and the lengths of the recorded files were inconsistent, parameter values at 30 data points were also extracted from each piece of the recorded file and saved in an Excel file. Before extracting the values of the EGG-based parameters, the wavelet transform was applied to each file to reduce the high-frequency noise of EGG signals, which might cause miscalculations in detecting the highest peaks and contacting and de-contacting peaks (Kong and Liew 1998). The data processing was performed by Matlab-based VoiceLab, which was developed by the Linguistic Lab of Peking University.

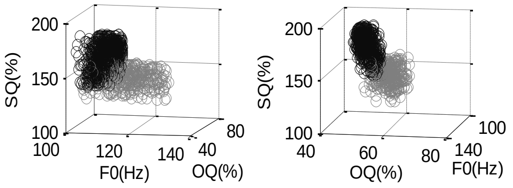
The lengths of the recorded data for chanting and speaking were approximately four minutes for each in material 1. The parameter values of 700 data points were extracted from each data file. Data that indicated abnormally low or high values for parameters were deleted because they may not be from vowels but from voiced consonants. In the case of material 2, parameter values at 30 data points were extracted from each sustained vowel.

3. PARAMETER ANALYSIS

In this section, the EGG parameters are compared with comparisons between chanting and speaking to recognize the inherent features of glottal source in chanting.

3.1 Parameter Distribution of Gadanlajima

Fig. 11 shows the parameter distribution of Gadanlajima for chanting and speaking. The x-, y- and z-axes in Fig. 11a represent F0, OQ_{egg} and SQ_{egg}, and those of Fig. 11b represent OQ_{egg}, F0 and SQ_{egg}, respectively. The parameters for chanting are shown by 700 black circles, and those for speaking are shown by 700 gray circles. Table 11 shows the mean and range of F0, OQ_{egg} and SQ_{egg}. In Fig. 11a, the data for chanting are located at a lower F0 region than that for speaking. The mean F0 of chanting is 102.3 Hz, which is 2 semitones lower than speaking. The F0 range of chanting is a little over 2 semitones (14.9 Hz), which is only half of the range of speaking. Fig. 11b shows the distribution of OQ_{egg}, demonstrating that the values are lower for chanting than for speaking. The mean OQ_{egg} of chanting is 52.8%, which is lower than that of speaking by 4.4%. The range of OQ_{egg} is 7% for chanting and 10.1% for speaking. Because the F0 range of speaking is wider than chanting, it is quite natural that the OQ_{egg} range of speaking is also wider. The SQ_{egg} for chanting is significantly higher than that of speaking. The mean SQ_{egg} value of chanting is 232.1%, which is 88% higher than that of speaking. Its range is 49.4%, which is 14.5% wider than that of speaking.



(a) (b)

Fig. 5 The distribution of parameters for chanting (black) and speaking (gray).

Table 11 The mean and range of F0, OQ_{egg} and SQ_{egg}.

		F0 (Hz)	OQ_{egg} (%)	SQ_{egg} (%)
Chanting	Mean	102.3	52.8	232.1
	Range	14.9	7.0	49.4
Speaking	Mean	116.1	57.2	144.1
	Range	28.0	10.1	34.9

To summarize, the chanting of Gadanlajima is characterized by low F0, low OQ_{egg} and high SQ_{egg} compared to speaking (see Table 12).

Table 12 Parameter characteristics of Gadanlajima.

	F0	OQ_{egg}	SQ_{egg}
Chanting	–	–	+
Speaking	+	+	–

3.2 Parameter Distribution in Sustained Vowels

It is common for OQ_{egg} and SQ_{egg} to co-vary with F0. Because there is a pitch range difference between chanting and speaking for Gadanlajima, the sustained vowels with the same pitch height are examined for chanting and speaking in this section. The results from sustained vowels in his entire pitch range show that the distribution of OQ_{egg} and SQ_{egg} in his high-pitch region do not show a significant difference between chanting and speaking. This is because 2~4 semitones near the lowest pitch region are used for actual chanting and speaking (cf. Table 11). Therefore, a significant difference is observed in the parameter values obtained from the low-pitch region. Thus, the pitch range compared here is limited from F2# (92.5 Hz) to B2 (123.5 Hz). Parameter values of 520 data points are extracted from both chanting and speaking. Fig. 6 shows the distribution of OQ_{egg} and SQ_{egg} of chanting (black) and speaking (gray). The x-axis and y-axis of Fig. 6 represent OQ_{egg} and SQ_{egg}. Table 13 shows the mean value and range of parameters. The OQ_{egg} in chanting is 3.5% lower than that in speaking. The OQ_{egg} range of chanting is 10.1% and that of speaking is 12.2%. The latter is slightly wider. The distribution of SQ_{egg} is separated between chanting and speaking. The mean SQ_{egg} of chanting is 19.9% higher, and the SQ_{egg} range is 15.4% narrower than speaking.

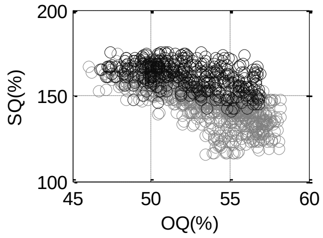


Fig. 6 The distribution of OQ_{egg} and SQ_{egg} in chanting (black) and speaking (gray) phonated at the range of F2#~ B2.

Table 13 The mean and range of OQ_{egg} and SQ_{egg} phonated at the range of F2#~ B2.

	OQ_{egg} (%)		SQ_{egg} (%)	
	Mean	Range	Mean	Range
Chanting	50.3	10.1	164.5	35.1
Speaking	53.8	12.2	144.6	50.5

Thus, the low OQ_{egg} and high SQ_{egg} are the common features in the chanting (see Table 14), which agrees with the result from Gadanlajima. The lower OQ_{egg} in chanting indicates the longer duration of vocal fold contact, suggesting that more pressed phonation is employed. The higher SQ_{egg} means that vocal fold contact is more rapid, which results in the higher energy in the higher frequency region.

Table 14 Parameter characteristics of sustained vowels.

	OQ_{egg}	SQ_{egg}
Chanting	Low	High
Speaking	High	Low

4. SPECTRAL MEASURE ANALYSIS

This section describes the acoustic manifestations that correspond to the results from the time domain analysis. The measurements include the H1-H2 and H1-A3.

4.1 H1-H2 Measurement

Spectral measure analysis is often used as an acoustic method to assess voice qualities. The difference in amplitude between the first and second harmonics (H1-H2) is a common measure to judge the tightness of vocal fold closure. The lower the H1-H2 is, the smaller the open quotient of vocal fold vibration becomes to produce pressed voice. For instance, breathy voice has high H1-H2, and creaky voice has low H1-H2. The lower OQ_{egg} is obtained for chanting from the EGG analysis, which suggests that more pressed phonation is used. The H1-H2 is expected to be low in chanting because it reflects the open quotient of glottal vibration (Bickley 1982).

Fig. 7 is the result of the H1-H2 values of sustained vowel /a/ at F2#~B2. The reason for adopting only vowel /a/ in this section is that the first formant of vowel /a/ has a higher frequency that hardly influences the values of H1 or H2. Fig. 7 shows that the H1-H2 of chanting is lower than speaking, as expected. The H1-H2 value of chanting is 7.6 dB and that of speaking is 11.4 dB, which is 3.8 dB higher than the former (Table 15). This result suggests that chanting has more pressed voice quality than speaking, which agrees with the result of low OQ_{egg} in chanting from EGG analysis.

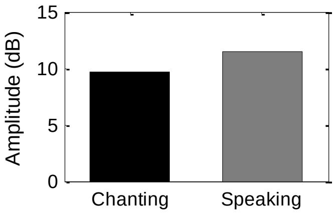


Fig. 7 H1-H2 value of vowel /a/ phonated at the range of F2#~B2.

Table 15 H1-H2 value of vowel /a/ phonated at the range of F2#~B2.

	Chanting	Speaking
Amplitude (dB)	7.6	11.4

4.2 H1-A3 Measurement

The difference in amplitude between the first harmonic and third formant frequency (H1-A3) is one of the common measures to judge the spectral tilt. The lower the H1-A3 is, the smaller the spectral tilt becomes. For instance, breathy voice has large H1-A3, which is indicated as a steep spectral tilt, unlike creaky voice, which has low H1-A3, as indicated by a small spectral tilt. The low H1-A3 is expected in chanting because the high SQ_{egg} is reflected by the low H1-A3 (Stevens and Hanson 1995).

Fig. 8 shows the H1-A3 values of sustained vowel /a/ at F2#~B2. The H1-A3 value of chanting is lower than that of speaking, as expected. The H1-A3 value of chanting is 21.8 dB and that of speaking is 28.4 dB, which is 6.6 dB higher than the former (Table 16). This suggests that chanting has a smaller spectral tilt than speaking. This finding agrees with the result of high SQ$_{egg}$ in chanting from the EGG analysis; namely, chanting has stronger energy in the high-frequency region.

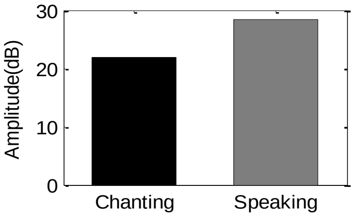


Fig. 8 H1-A3 value of vowel /a/ phonated at the range of F2#~B2.

Table 16 H1-A3 value of vowel /a/ phonated at the range of F2#~B2.

	Chanting	Speaking
Amplitude (dB)	21.8	28.4

Low H1-H2 and low H1-A3 in chanting from the spectral analysis (Table 17) agree with the results of low OQ$_{egg}$ and high SQ$_{egg}$ from the EGG analysis. Thus, chanting is described as a more pressed phonation with stronger energy in the high-frequency region. The results from the EGG parameter analysis in the time domain are reflected in the results from the acoustic parameter analysis in the frequency domain.

Table 17 Spectral characteristics of chanting and speaking.

	H1-H2	H1-A3
Chanting	Low	Low
Speaking	High	High

5. PHONATION ANALYSIS

In this section, the phonation mechanism of chanting is investigated by observing the EGG and DEGG waveforms and spectrograms.

5.1 Phonation Mode

EGG and acoustic parameter analyses were conducted in the earlier sections. The results indicate that pressed phonation with strong energy in the high-frequency region is one of the characteristics of chanting. To go a step further, the shape of EGG and DEGG waveforms are observed. The period-doubling pattern is a typical EGG waveform of chanting, as shown in Fig. 3. The subcycles in the period-doubling waveform seem to be derived from the phase difference between vocal fold and supraglottal oscillations. This feature of the EGG waveform is quite similar to what was observed in VVM (Fuks. et al. 1998).

The subcycles are more clearly seen in the DEGG waveform. Fig. 9 shows the EGG waveform of sustained vowel /a/ phonated at G2 (98 Hz) and the corresponding DEGG waveform. The EGG waveform is characterized as period-doubling, and the DEGG waveform is characterized as double GCIs. A question arises as to whether the GCI with lower amplitude yields because of supraglottal adduction. Although further physiological experiments are needed, it is not unreasonable to assume that the GCI with lower amplitude is caused by the supraglottal adduction. The glottal opening is immediately followed by the supraglottal adduction, and the supraglottal adduction is followed by the actual glottal release (de-contact phase).

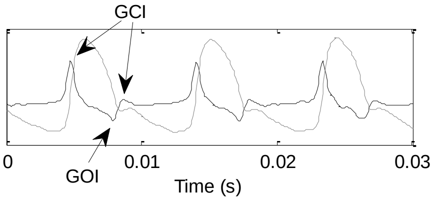


Fig. 9 EGG and DEGG waveform during sustained vowel /a/ phonated at G2.

5.2 Harshness

The Tibetan Buddhist chants are perceived as sounds with low and tense voice containing audible airflow noise. Regarding the harshness setting, which is one of the subcategories in the phonatory setting, Laver (1980) suggests,

‘This is the setting where the ventricular folds become involved in the phonation of the true vocal folds by squeezing closed the ventricle of Morgagni and pressing down on the true vocal folds, …… In order to bring the ventricular folds to this position, a high degree of muscular tension is needed, and the effect is normally to make phonation auditorily very harsh.’

Irregularity in pitch and spectral noise are the characteristics of harsh voice (Laver 1980). Fig. 10 compares a spectrogram of sustained vowel /a/ phonated at G2# (103.8 Hz) in chanting to that of speaking. The spectrogram of chanting shows noise lying over a wide frequency region. In contrast, significant noise is not found in the spectrogram of speaking.

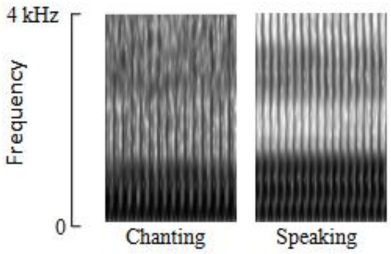


Fig. 10 Wide-band spectrograms of sustained vowel /a/ phonated at G2# in chanting and speaking.

Harsh and growl phonations are typical extended vocal techniques involving supraglottal oscillations. As mentioned above, the frequency of the ventricular fold oscillation is found as F0, F0/2, or F0/3 in previous literature. In the case of growl voice, the frequency of the aryepiglottic fold oscillation is found as F0/2 (Sakakibara et al. 2004). If we assume supraglottal oscillation in the chanting voice, its frequency is thought to be equal to that of vocal fold vibration, and supraglottal and glottal closures occur alternately. Then, the noise is due to excessive constriction and friction at the level of the supraglottal structure. The characteristics of these irregular vocalizations are often interpreted as period-doubling bifurcations and subharmonics. The subharmonics are usually recognized in the narrow-band spectrograms; however, they are not found in this study because the frequency of the supraglottal oscillation is equal to F0.

Period doubling in the EGG waveform, double GCIs in the DEGG waveform, and the widespread noise in the spectrograms are considered to result from the supraglottal activities that immediately follow the vocal fold opening.

6. CONCLUSION

Supraglottal constriction and adduction were estimated to occur in Tibetan Buddhist chants. Pressed phonation with a small spectral tilt resulting from the EGG and acoustic analyses, period doubling in the EGG waveform, double GCIs in the DEGG waveform, and the noise in the spectrograms were supportive evidence for these occurrences. Tibetan Buddhist chants of Kumbum Monastery maintain throat-singing traditions.

Further physiological research using tools, such as high-speed cameras, is needed to clarify the laryngeal vibratory mechanism of GCIs. Other styles of Tibetan Buddhist chants should also be investigated in future work.

7. ACKNOWLEDGEMENTS

This study was supported by China Social Sciences Funds (Grant No: 10&ZD125).

8. REFERENCES

[1] Baken, R. J. and R. F. Orlikoff (2000) *Clinical Measurement of speaking and voice*. New York: Delmar learning.

[2] Bickley, C. (1982) “Acoustic analysis and perception of breathy vowels.” *Speech Communication Group Working Papers*, 73-93. Cambridge, Mass: MIT, Research Laboratory of Electronics.

[3] Childers, D. G., D. M. Hicks, G. P. Moore and L. Eskenazi (1990) “Electroglottography and vocal fold physiology.” *Journal of Speech and Hearing Research* 33, 245–254.

[4] Childers, D. G. and A. K. Krishnamurthy (1985) “A critical review of electroglottography.” *Critical Reviews in Biomedical Engineering* 12, 131–161.

[5] Childers, D. G. and J. N. Larar (1984) “Electroglottography for laryngeal function assessment and speech analysis.” *IEEE Transactions on Biomedical Engineering* 31, 807–817.

[6] Childers, D. G., G. P. Moore, J. M. Naik, J. N. Larar and A. K. Krishnamurthy (1983) “Assessment of laryngeal function by simultaneous, synchronized measurement of speech, electroglottography and ultra-high speed film.” In *Proceedings of the Eleventh Symposium on Care of the Professional Voice (New York)*.

[7] Childers, D. G., J. M. Naik, J. N. Larar, A. K. Krishnamurthy and G. P. Moore (1983) “Electroglottography, speech and ultra-high speed cinematography.” In I. R. Titze and R. Scherer (eds.) *Vocal Fold Physiology and Biophysics of Voice*, 202–220. Denver: Denver Center for the Performing Arts.

[8] Fuks, L., B. Hammarberg and J. Sundberg (1998) “A self-sustained vocalventricular phonation mode: acoustical, aerodynamic and glottographic evidences.” *KTH TMH-QPSR* 3, 49–59.

[9] Henrich, N. (2004) “On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation.” *Journal of the Acoustic Society of America* 115, 1321-1332.

[10] Herbst, C. (2004) “Evaluation of various methods to calculate the EGG contact quotient.” Diploma Thesis, Department of Speech, Music and Hearing, KTH, Stockholm.

[11] Hollien, H., J. Michel and E. Doherty (1973) “A method for analyzing, vocal jitter in sustained phonation.” *Journal of Phonetics* 1, 85–91.

[12] Howard, D. M. (1995) “Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers.” *Journal of Voice* 9 (2), 163-172.

[13] Howard, D. M., G. A. Lindsey and B. Allen (1990) “Toward the quantification of vocal efficiency.” *Journal of Voice* 4, 205–212.

[14] Kong, J. (2001) *Lunyuyanfasheng* [On Language Phonation]. Beijing: Central Nationalities University Press.

[15] Kong, J and W-C. A. Liew (1998) “Wavelet analysis of speech phonations of Chinese speakers.” In *Proceedings of the 1st International Conference on Chinese Spoken Language Processing (Singapore)*.

[16] Laver, J. (1980) *The Phonetic description of voice quality*. Cambridge: Cambridge University Press.

[17] Lindestad, P., M. Södersten, B. Merker and S. Granqvist (2001) “Voice source characteristics in Mongolian ‘throat singing’ studied with high-speed imaging technique, acoustic spectra, and inverse filtering.” *Journal of Voice* 15(1), 78–85.

[18] Raymond, D. K. and J. B. Martin (2000) *Voice quality measurement*. California: Singular Publishing Group.

[19] Rothenberg, M. and J. J. Mahshie (1988) “Monitoring vocal fold abduction through vocal fold contact area.” *Journal of Speech and Hearing Research* 31, 338-351.

[20] Sakakibara, K-I. (2003) “Production mechanism of voice quality in singing.” *Journal of the Phonetic Society of Japan* 7 (3), 27-39.

[21] Sakakibara, K-I., L. Fuks, H. Imagawa and N. Tayama (2004) “Growl voice in ethnic and pop styles.” In *Proceedings of the International Symposium on Musical Acoustics (Nara)*.

[22] Smith, H., Stevens, K. N. and R. Tomlinson (1967) “On an unusual mode of chanting by certain Tibetan lamas.” *Journal of the Acoustic Society of America* 41, 1262-64.

[23] Stevens, K. N. (1977) “Physics of larynx behavior and larynx modes.” *Phonetica* 34, 264-279.

[24] Stevens, K. N. and H. Hanson (1995) “Classification of glottal vibration from acoustic measurements.” In O. Fujimura and H. Hirano (eds.) *Vocal fold physiology: Voice quality control*, 147–170. San Diego: Singular Publishing Group.

[25] Titze, I. R. (1990) “Interpretation of the electroglottographic signal.” *Journal of Voice* 4, 1-9.

[26] Titze, I. R., R. J. Baken and H. Herzel (1993) *Evidence of chaos in vocal folds vibration in Vocal fold physiology*. San Diego: Singular Publishing Group.

A Phonetic Study on Chanting of Chinese Five Syllable Modern-Style Poems

Yang Feng

ABSTRACT

This paper analyses the chanting of 21 five-syllable modern-style poems, which is Chinese traditional style of poem reciting which has long history and special melody. The phonetic analysis of chanting and its relationship with poetic metrics has not yet been fully studied. This paper is to find out the prosodic hierarchy according to pause duration, and probe the phonetic features and methods of chanting. Results reveal that pause exists after “level-level” tonal combinations, which is a kind of metrical pattern of Chinese poetry. The duration of syllables doubles in sentence final position. An exclamation is added when the sentence ends with a checked syllable. The pitch of syllables with level tone is lower than that of syllables with oblique tones, alternation of level and oblique tones forms the chanting melody. Sentences and poems with same metrical pattern have the same chanting melody.

Keywords: Chanting, Prosodic hierarchy, Five-syllable modern-style poems

1. INTRODUCTION

Chanting has a long history in China. It accompanies poetry as the traditional style of reciting poems and proses with cadence and pleasant melody. After the New Culture Movement chanting has gradually declined, today only a few very old scholars can chant.

Professor Zhao Yuanren made great contributions to chanting. He is the first scholar who recorded and studied chanting, claiming to save this kind of art and compose songs according to chanting[7]. Mr Tang Wenzhi set a special “Tang Melody” and trained many chanters. Yang Yinliu, Chen Bingzheng, Sun Xuanling and Du Yaxiong studied the relationship between chanting and music[2, 6, 8]. Wang Enbao, Chen Shaosong, and Qin Dexiang collected materials about chanting and studied the history, methods and melodies of chanting[1, 4].

None of researches mentioned above has employed the method of phonetic analysis. This paper carries out phonetic study of 21 five-syllable modern-style poems chanted by Professor Tu An, trying to find out the prosodic hierarchy according to silent gaps, and probe the phonetic features and methods of chanting[3].

2. METHOD

2.1 Material and chanter

Poems recorded are five-syllable modern-style poems. The chanter is Professor Tu An, a 87-year old male scholar. He was born in Changzhou. He had been well-educated in old-style private school with proficiency in classic Chinese literature.

2.2 Recording

These poems are recorded in the Phonetic Lab. Each poem is recorded three times. The first time is reciting in standard Chinese, the second time is reciting in the chanter’s dialect, the third time is chanting in the chanter’s dialect. The software used in analysis is Praat.

3. RESULTS

3.1 Prosodic hierarchy and metrics

First the duration of each pause in chanting is detected, then the pauses are classified according to their durations as the marker of boundaries of prosodic levels. Figure 1 shows that the duration of pauses is distributed in three ranges: within 100ms, 300ms, and 1700ms. According to the three classes of pause duration in chanting, three levels of prosodic units can be identified: foot, prosodic phrase, and prosodic sentence. Prosodic boundaries agree with the metrics of poems.

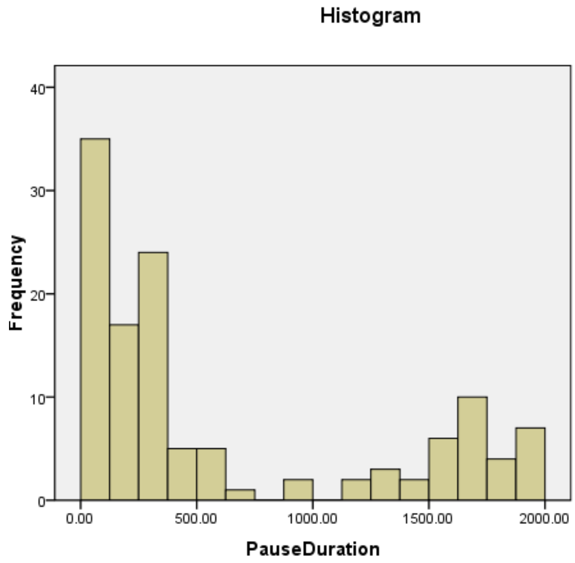


Figure 1: The distribution of duration of each pause in chanting

Boundaries between feet appear after "oblique-oblique" tonal combinations. As shown in Figure 2, the metrics of the sentence "ming cheng ba zhen tu" is "level-level-oblique-oblique-level", the boundary between feet is after "ba zhen", the combination of "oblique-oblique" syllables, with a gap of only 5ms. There is no obvious pause at the boundaries between feet, and no lengthening of syllables before foot boundaries.

Boundaries between prosodic phrases appear after "level-level" tonal combinations. That is, after the second or the fourth syllable of a five-syllable sentence. As shown in Figure 2, the longest pause appears after "ming cheng", the "level-level" tonal combination, and before the boundary the syllable "cheng" lengthens. In average, the gap of the boundaries of prosodic phrases is 0.350s, the syllables before the boundary lengthen for 12%.

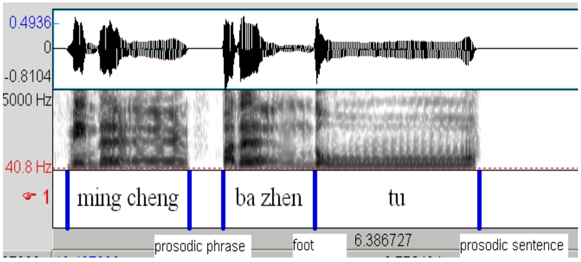


Figure 2: The wave and spectrograph of sentence "ming cheng ba zhen tu" in chanting, "level-level-oblique-oblique-level"

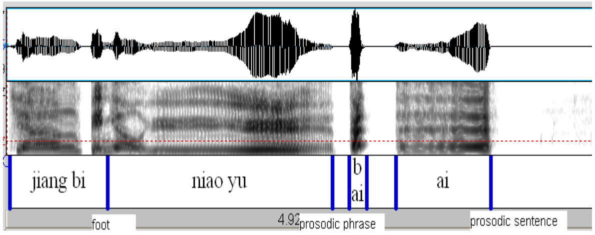
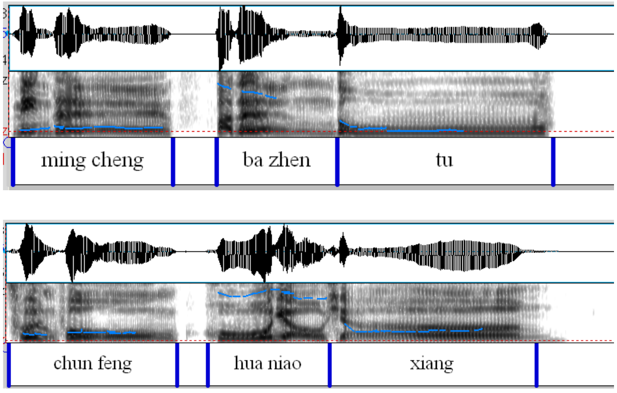


Figure 3: Sentence final checked syllable lengthens by adding an exclamation

Boundries between prosodic sentences appear at the final position of each line of the poem, being marked by doubling of the duration of syllables, which is the most obvious feature of chanting. As shown in Figure 2, the duration of the syllable "tu" doubles. The average duration of sentence final syllables is up to 1.7s. If the final syllable is a checked syllable, an exclamation is added to the syllable to reach the effect of lengthening. As in Figure 3, the sentence final syllable "bai" is a checked one in the chanter's dialect, therefore, an exclamation "ai" is added and lengthened.

3.2 Chanting melody and metrics

In chanting syllables of level tone have a lower pitch while syllables of oblique tones have a higher pitch, forming the rise and fall in melody through alternation of level and oblique syllables. Sentences of same metrics have the same pattern of melody, and poems of same metrics are similar in melody.



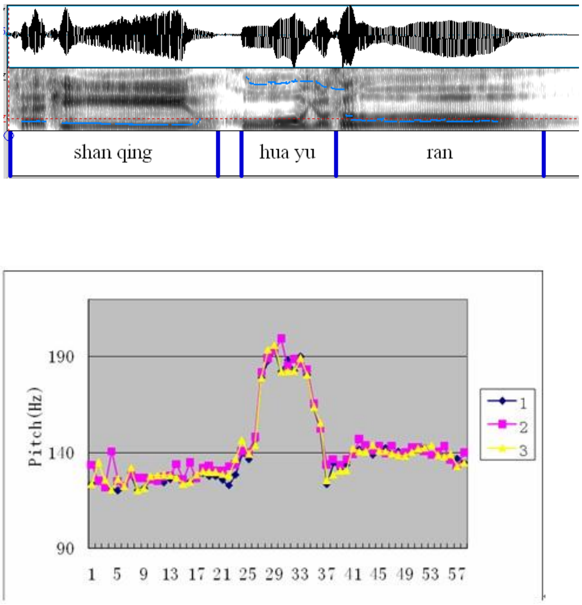


Figure 4: Three examples of "level-level-oblique-oblique-level" pattern and their melody curve

As shown in Figure 4, all the three sentences are of "level-level-oblique- oblique-level" pattern. The pitch of "level- level" tonal combinations "ming cheng", "chun feng", and "shan qing" is lower than the "oblique-oblique" combinations "ba zhen", "hua niao", and "hua yu". The three sentences have the same metrical pattern, therefore their melodies are similar, forming a "low-low-high-high-low" curve. By combining sentences of the same metrical pattern, poems of the same metrics have the same melody, therefore, chanters may chant poems of the same metrics with similar melody.

4. CONCLUSIONS

From analysis above, the chanting of five-syllable modern-style poems has the following features:

Pause exists after "level-level" tonal combinations. Before the break the duration of the syllable lengthens

The duration of syllables doubles in sentence final position. An exclamation is added when the sentence ends with a checked syllable.

The pitch of syllables with level tone is lower than that of syllables with oblique tones, alternation of level and oblique tones forms the chanting melody. Sentences and poems with same metrical pattern have the same chanting melody.

These are main features of the chanting of five-syllable modern-style poems, and also the main techniques of chanting. Chanting with this method and personal features would be a splendid works of voices.

This research is an analysis of one chanter, more chanters and various types of chanting of poems in different dialects will be included in future study.

5. ACKNOWLEDGMENT

This research was supported by China Social Sciences Funds of the Ministry of Education (Grant 10JJD740007) and National Social Sciences Funds (Grant 10&ZD125). We thank two anonymous reviewers for the comments on the earlier version of the paper.

6. REFERENCES

[1] Chen Shaosong. 2002. *Chanting of classical poetry.* Beijing: Social Sciences Academic Press.

[2] Du Yaxiong, 1990. *The relationship between language and music*, Chinese Music, 1.

[3] Kruckenberg, A. and Fant, G. 1993. *Iambic versus trochaic patterns in poetry reading,* Nordic Prosody VI, Stockholm, 1993, 123 – 135.

[4] Qin Dexiang. 2002. *Chanting and music.* Benjing: The China literary federation of press.

[5] Wang Li. 2000. *Metrical pattern of poetry.* Beijing: Commercial Press.

[6] Yang Yinliu. 1981. *History of Ancient Chinese Music.* Beijing: People's Music Publishing House.

[7] Zhao Yuanren. 1994. *Symposium of Zhao Yuanren's Music Study.* Beijing: The China literary federation of press.

[8] Zhang Ming. 1998. *Outline of Language musicology.* Beijing: Culture and Art Publishing House.

Voice Quality Features of Korean Students at Advanced Level of Chinese

Oh Hanna

Abstract

This paper investigates whether the Korean students at advanced level of Chinese produce Chinese with specific voice quality, and which voice quality parameters have significant difference. The study includes comparative analysis between the Korean and Chinese vowel /a/ in a frame sentence for Korean students. It can be made certain by measuring voice quality parameters by means of EGG. The results showed that the Korean learners use specific voice quality features for improving Chinese nativeness, and especially the phonation quality changes were obvious. They produced Chinese phonation quality characteristics depending on higher F0, lower OQ, SQ and H1-H2 compared with Korean. In addition, this study found that there are similarities between phonation quality features of Korean/Chinese contrasts and Korean lax/tense contrast features. It suggests that in the target language processing, phonetic features of the L1 contribute to “phonation quality” perception and production of a L2.

Index Terms: second language acquisition, voice quality, phonation, Chinese, Korean

1. Introduction

When the L2 learners produce the target language, should one consider the phonation quality features of the target language?

As is known to all, the main purpose of learning L2 is common to "communication", while the goal of the L2 learners is often reaching to a native speaker-like level. So, when we practice L2 speaking, we often try to imitate its specific phonetic characteristics to close to native pronunciation. Stockmal et al.[1] examined that bilinguals can produce two languages in significant different voice qualities respectively. Esling and Wong[2] and Stockmal et al.[3] also emphasized that the voice quality is one of important cues for evaluating L2 learners fluency, and occupied very important position in the speak processing as well. However, much of the L2 research to date has focused on articulation properties.

This study, therefore, according to Kong’s[4] classification of "Voice Quality", is divided into Phonation Quality and Articulation Quality, taking articulation quality data for reference data, mainly discusses whether the Chinese phonation quality of Korean speakers differ from Korean, and continue to make improvements on what kind of voice qualities are used to improve nativeness of Chinese.

2. Method

2.1 Subject

The participants in this study included 20 native speakers of Korean (10 males, 10 females from ages 25-35 years old) with, whose language background was limited to Seoul dialect. In addition, they all had HSK (Chinese proficiency test) certificates of an advanced level.

2.2 Speech Material

Recording materials consisted of two kinds of sentences with same meaning and different languages. To minimize the effect of F0 in Chinese tone, this study only selected the Chinese and Korean vowel "/a/", with no specific meaning. Each word was recorded in a frame sentence, which was as follows:

Chinese：这是汉语“啊”的发音。

[tʂɤʂʅ hɑny "a" tə fain]

Korean：이것은 한국어 ‘아’ 발음입니다.

[igʌsɯn hɑŋ̚gugʌ 'a' b̥ɑrmimnidɑ]

The sentences above mean “This is the ‘a’ sound of Chinese/Korean.”

2.3 Recording and Data Analysis

The recording was taken in a sound-proof booth. It used Kay company production PENTAX Model 6103 EGG and microphone. Both of them can put the EGG signal and the speech signal to audio processing software in computer at the same time with sampling rate of 22 kHz. Each material was read twice at normal speed, and 20 points were selected for each sample. The original signal file, after using cutting and noise reduction functions of Audition1.5, extracted the needed parameters by means of the Voicelab program (written on Matlab by Linguistics Lab. of Peking University). Next it was saved in Excel form. Then both Chinese and Korean parameters were compared, including male and female parameters. Finally Excel's *t*-tests for statistical analysis and graph functions were used.

3. Results

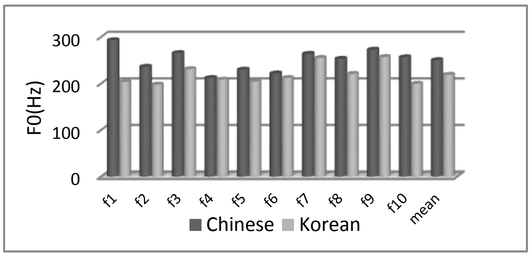
3.1 Phonation quality analysis

There are lots of acoustical parameters of phonation quality. Common parameters are: 1) F0; 2) OQ; 3) SQ and 4) H1-H2 [4].

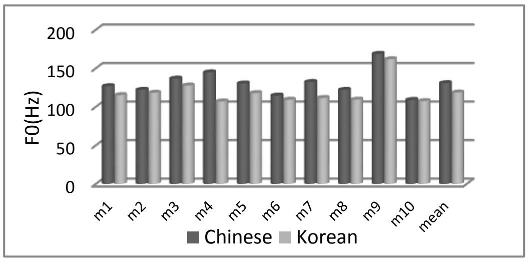
3.1.1 F0

Different types of phonation have different vocal cords vibration frequency, therefore, F0 is one of important aspects of reflecting voice quality features.

Figure 1 shows the F0 change between two languages for male and female speakers respectively.



(a)



(b)

Figure 1: F0 data comparison of the Korean/Chinese vowel /a/, (a) for female, (b) for male.

Generally the F0 values of modal voice ranges from 150 to 300 Hz for females and 100-200 Hz for males respectively. Clearly, all data is in range of modal voice, and results show that Chinese F0 values are higher than Korean F0 values without exception. The mean value of F0 for female is 251Hz in Chinese and 219Hz in Korean. As for males, the mean value is 131Hz in Chinese and 119Hz in Korean. These results were confirmed by statistical analysis (*t*-test). The change rate of F0 between Korean and Chinese, no matter what gender, are highly significant (p<0.01). Also the change rates of females are higher than males.

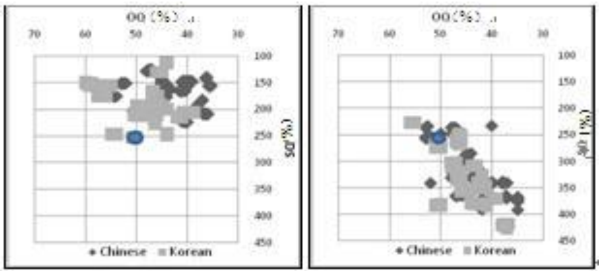
Additionally, to confirm whether these results are free from interference by the Chinese tone 1(high), the frame sentences were analyzed as well. The results showed that the mean values and whole graphs of F0 in Chinese sentences are also higher than in Korean, and change range of females are higher than males in the same way as above. The mean values are as follows: 211Hz in Korean and 268Hz in Chinese for females, 124Hz in Korean and 128Hz in Chinese for males.

In brief, Korean learners of Chinese raise F0 to achieve specific voice quality of Chinese. In addition, based on Laver [5]'s study, it was reported that lax voices tend to have lower pitch, and from the result of F0 value, we can infer that Korean vowel is laxer than Chinese.

3.1.2 OQ and SQ

OQ is defined as the ratio of open phase and pitch period， and SQ is defined as the ratio of opening phase and closing phase [4]. With OQ and SQ parameters, Figure 2 visualizes directly the difference of the phonation types between females and males as well as Korean and Chinese. Because of space limitations, bar charts like Figure 1 were omitted.

Based on Laver [5] and Kong [4], the OQ value of modal voice is around 50%, and around 250% for the SQ value. As shown in Figure 2, phonation types of most female data for Korean vowel seems to belong to the modal voice and the breathy voice boundary, while most parameters of the Chinese vowel is close to between the modal voice and the creaky voice boundary. Compared to females, most male data belong to between the modal voice and the creaky voice boundary, and there is a large part that is overlapped.



(a) (b)

Figure 2: The phonation type (OQ and SQ) chart of the Korean/Chinese vowel /a/, (a) for female, (b) for male.

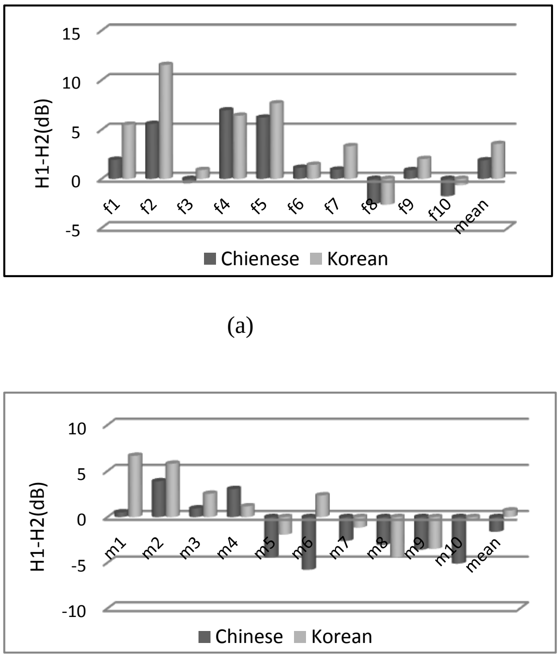
The specific results of this analysis showed that, OQ values of Korean are higher than Chinese especially for females, there is just one exception in males. The difference of OQ values between Korean and Chinese, no matter what gender, is shown to be highly significant (0.0002 for female, 0.0098 for male) statistically. Also the change was more obvious in female data, same as F0 result. The difference of the OQ value is one of a huge factor to distinguish different languages. According to above results, from the view of native speakers, it is possible that the change of voice quality in female speakers would be more easily identified.

As for the SQ value, the result is overall the same as the OQ data. Most of the SQ values in Chinese are lower than Korean. The only difference is that the SQ data of the two languages are shown to have significant differences only in females (p<0.01). Because some inter-speaker variability was found, the statistical analyses demonstrated no significance for males.

It is notable that, OQ values of Chinese in both females and males are more close to a pressed voice feature, compared to the Korean. However, the results of Figure 2 shows that the SQ values of Chinese much lower than Korean, instead. There is a similar result concerning this. Wang [6] found that the SQ value of the tense voice is not always higher than the lax voice. It is not so clear whether the reason is specific phonation quality of Korean speakers or other factors, it is to be solved through further study. But one thing is for sure, the way of processing Chinese phonation quality different from Korean, and Korean learners of advanced level could try to find out a proper way in own voice conditions to realize the target language features.

3.1.3 H1-H2

H1-H2 amplitude analysis has been widely used by linguists to infer the state of the glottis in distinguishing different phonation types [7], [4]. Physically, a higher value of H1-H2 would indicate pressed voice and a lower or negative value would indicate breathy voice in general.



(a)

(b)

Figure 3: H1-H2 amplitude comparison of the Korean/ Chinese vowel /a/,(a) for female, (b) for male.

Results showed H1-H2 values of Korean are higher than Chinese no matter what gender. The mean value of female is 1.891dB in Chinese, 3.554dB in Korean. As for males, the mean value is -1.62dB in Chinese, 0.719dB in Korean. These results were confirmed by statistical analysis (*t*-test). The change rate of H1-H2 between Korean and Chinese is highly significant (p<0.05).

From the above results, it can be said that the Korean learners employed a relatively pressed laryngeal setting in Chinese compared to the one in Korean.

In a word, the results of phonation quality parameters revealed that there are substantial differences between both languages for Korean speakers.

3.2 Articulation quality analysis

Fant [1] pointed out that, because speech organs work in a systemic way, phonation is interacting with articulation unavoidably. Articulation quality is defined as the height of tongue and roundness of lips. in acoustic analysis, articulation is quantified by formants structure [4].

3.2.1 Formants structure

Formants structure can reflect resonant characteristics of both of two languages. As the position of your tongue relatively large change in different vowels, it can affect the height of larynx, and causes change of phonation quality. And raised larynx can raise all formants value as well as F0 value. On the contrary, lowered larynx can bring all formants and F0 values down [8].

As can be seen below (Table1), the formants data between two languages are quite similar. The results show that there are no significant differences in each object statistically due to inter-speaker variability.

Table 1. Comparison of mean values of formants for Korean(K)/Chinese(C) vowel /a/ for Korean speakers, (a) for female, (b) for male data.

		Mean	Sig.	Mean	Sig.
F1 (Hz)	C.	977	0.3 (p<.05)	822	0.7 (p<.05)
	K.	952		813	
F2 (Hz)	C.	1497	0.7 (p<.05)	1437	0.4 (p<.05)
	K.	1486		1482	

(a) (b)

Based upon the above results of phonation and articulation quality analysis, it was found that generally Korean speakers didn’t rely on articulation quality for producing Chinese specific voice quality in vowel, while mainly use the strategy of changing phonation quality.

4. Discussion

Generally, phonation quality has not been emphasized in teaching and learning Chinese. However, the present study suggested that Korean students could notice differences of phonation quality in Chinese phonetics.

How could they observe differences between Korean and Chinese phonation quality? According to studies of Flege [9] and Best [10], the influence of L1 is essential to explaining this question.

The notable factor is, the Korean consonant system has a lax/tense contrast, and Cho et al. [11] and Kim et al. [12] studies indicated that the F0 and H1-H2 values of Korean tense consonants are higher than lax sounds.

The table given below made comparisons between phonation quality parameters of Korean lax/tense sounds in CV syllables [11] and Korean/Chinese phonation quality parameters of the study above.

Table 2. Comparison of mean values of F0 and H1-H2 in Korean lax/tense contrast and Korean(K)/Chinese(C) vowels for Korean speakers, (a) for lax/tense contrast, (b) for Korean/Chinese vowels. (Diff. stands for difference)

	Lax	Tense	Diff.	K.	C.	Diff.
F0(Hz)	124	138	11%	119	131	10%
H1-H2 (dB)	4,2	-4.8	214%	0.7	-1.6	346%

Obviously, there were overall trends of changes in F0 and H1-H2 values between Korean lax/tense contrasts and Korean/Chinese parameters of the above study. In a word, Korean tense sounds correspond to Chinese vowel sounds for Korean speakers.

Combined with phonetic learning and perception models, the results suggest that due to Korean language having lax/tense contrast with different phonation qualities, it is possible that Korean speakers could be more sensitive to those features. Therefore, they process Chinese phonation quality by taking Korean lax/tense features as a reference.

The learners of different language backgrounds produce different L2 performances. It is worthwhile for further studies on how phonation qualities of L1 effects on L2 phonetic acquisition.

5. Conclusions

The results showed that advanced Korean learners produce Chinese with different voice quality features, compared to Korean. They who depend on higher F0, lower OQ, SQ and H1-H2 produce Chinese phonation quality characteristics, while articulation quality change is not quite remarkable. The general trends of the voice quality processing strategy of Chinese is the same in both males and females, while overall, female voice quality changes are more obvious than in males.

In addition, this study found that, the performance of

Korean/Chinese phonation quality features for Korean students is quite similar with Korean lax/tense contrast features. It suggests that L1 transfer also plays a role in phonation quality perception and production of L2.

Finally, this study indicate that, due to phonetic features of L2 itself or in order to improve the nativeness of L2, the L2 learners at advanced levels should focus on the voice quality features of L2 naturally, and try to reflect the particular manners in L2 phonation.

6. Acknowledgements

This research was funded by China National Natural Science Funds (Grant 61073085). I am very grateful to the anonymous reviewers for the comments on the earlier version of the paper.

7. References

[1] Stockmal, V., Bond, Z.S., "Same talker, different language", Applied Psycholinguistics, 21:383–393, 2000.

[2] Esling, J., Wong, R., "Voice Quality Settings and the Teaching of Pronunciation", TESOL QUARTERLY, 17:89-95, 1983.

[3] Stockmal, V., Bond, Z.S., Moates, D., "Judging voice similarity in unknown languages", In Proceedings of the 17th Congress of Linguists, Prague, 2004.

[4] Kong, J.P., "On Language Phonation", Minzu University of China Press, 2001.

[5] Laver, J., "The Phonetic Description of Voice Quality", Cambridge University Press, 1980.

[6] Wang, F., Kong, J.P., "A study of lax/tense voice in the Wuding Yi", Status Report of Phonetic and Music Research, Linguistics Lab of Peking University, 2008.

[7] Kirk, P.L., Ladefoged, P., Ladefoged, J., "Using a spectrograph for measures of phonation types in a natural language", UCLA WPP 59, 102-113, 1984.

[8] Fant, G., Gauffin, J., "Speech Science and Speech Technology", Shangwu publishing house, 1994.

[9] Flege, J.E., "Second Language speech learning: Theory, findings and problems", in Strange, W. [ED], Speech Perception and Linguistic Experience: Issues in Cross-Language Research., 233-277, Baltimore: York Press, 1995.

[10] Best, C.T., "The emergence of native language phonological influences in infants: A perceptual assimilation model", in Goodman, J., Nusbaum, H.C. [ED], The Development of Speech Perception, 167-224, Cambridge: MIT Press, 1994.

[11] Cho, T.H., Jun, S.A., Ladefoged, P. 2002. Acoustic and Aerodynamic Correlates of Korean Stops and Fricatives. Journal of Phonetics. 30, 193-228.

[12] Kim S.H., Emily C., "Phonetic Duration of Korean /s/ and its Borrowing into Korean", Japanese/Korean Linguistics, 10:406-419, 2002.

A Forensic Aspect of Articulation Rate Variation in Chinese

Cao Honglin[1,2], *Wang Yingli*[3]

1 Key Laboratory of Evidence Science (China University of Political Science and Law),
Ministry of Education, China;
2Dept. of Chinese Language and Literature, Peking University, Beijing, China;
3Center of Criminal Technology, Public Security Bureau of Guangdong Province, China

ABSTRACT

This study presents the statistical data for the articulation rate (AR) of 101 male Chinese speakers. 100 spontaneous telephone speech samples produced by 100 speakers and 10 samples produced by another speaker are investigated to test the inter- and intra-speaker variation of AR respectively. Two separate histograms for the global AR and the mean AR are shown to be near normal distribution. It is found that the range of AR for the one speaker is small and relatively stable when the topic and style are similar. The global AR and mean AR can be used as discriminatory features for forensic speaker identification.

Keywords: Speaker identification, articulation rate, Chinese language, spontaneous speech

1. INTRODUCTION

Speech tempo is one of the prosodic features, which can be exhibited by two methods, one is speaking rate/speech rate/syllable rate (all terms can be abbreviated to SR), and another is articulation rate (AR). Both SR and AR can be defined as “the number of output units per unit of time” [1] (e.g., syllables per second). The biggest difference between SR and AR is that the former includes pause intervals but the latter does not [1-9].

The previous studies of speech tempo showed that AR had more speaker-discriminating power than SR in English [2] and in German [6-7]. When calculating AR, one important issue is how to deal with pause. It is known that pause basically can be silent/unfilled and filled (such as um and uh in English). However, the specific methods of different investigators are not the same. All studies in [1-9] exclude silent pauses, and all except Laver [4] and Cao [8] exclude filled pauses as well. In forensic studies, Künzel [6] proposed a formula to calculate AR, which was “number of syllables/ [duration – combined duration of all pauses]”. More recently, Jessen [7] refined the steps and criteria of the measurement of AR in German and made some rather persuasive conclusions.

Although the AR parameter is found to be powerful in forensic speaker identification in English and German, similar studies on Chinese are rare. The present study focuses on the AR variation of Chinese speakers and aims to provide some useful statistical data by using the method proposed in [7].

2. METHOD

2.1 Speech material

Considering that most of the forensic-phonetic casework relates to telephone recordings (TRs), a database named *FTRD 2010* was compiled at Peking University, which included a number of spontaneous TRs in the daily work. 100 different TRs of 100 male speakers (M1-100) and 10 different TRs of one male speaker (M101) were selected from the *FTRD 2010* database for evaluating the inter- and intra-speaker variation of AR respectively. The 10 TRs (all being talks with judges about legal cases) from M101 were similar in style.

The topics of the TRs were about the discussion of forensic cases in conversational style. All TRs were spoken in Mandarin Chinese with no evident regional features. The age of the 100 different speakers ranged from 22 to 55, according to a preliminary survey. And the speaker M101 was 29 years old. The speakers consisted of forensic scientists, judges, police officers, lawyers, interested parties and lab workers.

Each individual speaker’s speech was selected and saved as a single wav file through the Adobe Audition 3.0 software, i.e. the speech of irrelevant speaker (e.g. a female lab worker) was excluded. The durations of the final speech samples of speakers M1-100 and the speaker M101 were on average 51s (with standard deviation (SD) of 14s and range from 20s to 82s) and 39s (with SD of 12s and range from 26s to 57s) respectively.

2.2 Measurement

To get the AR data, three important issues have to be clarified. First, which linguistic unit should be counted? This is an easier question for the present study, because each Chinese character is concurrently one syllable [10]. So the AR will be measured in terms of monosyllabic Chinese characters per second in the present study.

The second issue is about the method and criteria for measuring of AR. We have followed Jessen [7]: The realized syllables, not canonical ones were counted. The size of speech intervals [1] were selected by the investigator's short-term memory to choose the number of syllables easily (i.e., the investigator goes through the speech signal and selects portions of fluent speech containing a certain number of syllables that can easily be retained in short-term memory.). Each memory selected stretch consisted of only fluent speech, excluding silent pauses, filled pauses, laughter sounds and any immoderate syllable lengthening. In order to minimize increasing the phrase-final lengthening effect of very short utterance on AR, the lowest number of syllables per stretch was set to be no less than four.

Third, both AR for the entire recording, which was called "global AR (GAR)" [7] and AR for each selected speech stretch, which was called "local AR (LAR)" [7] was to be calculated. To get GAR of one speech sample, the total duration of its all selected stretches was divided by the total number of syllables of all selected stretches. And LAR was calculated by the number of syllables dividing by the duration for each stretch.

Both the number of syllables and the duration of each stretch were extracted from the "TextGrid" file generated by the Praat (version 5143) [11]. The procedure is illustrated in Figure1.

The numbers of memory stretches counted were on average 30 for M1-100 (with SD of 8.5 and range from 12 to 54), and 26 for M101 (with SD of 7.7 and range from of 18 to 41). The number of syllables per stretch for M1-100 was range from of 4 to 22 and on average 7.8.

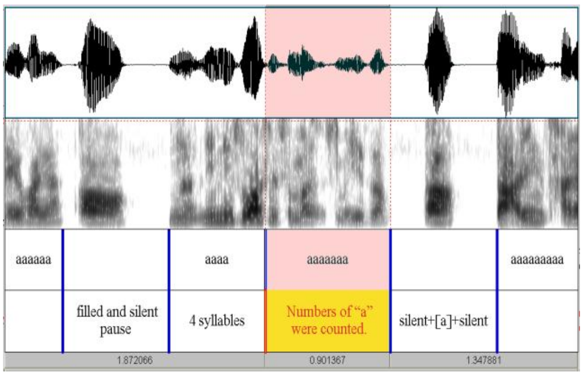


Figure 1: The annotation procedure for each memory stretch. One letter "a" stands for one syllable.

3. RESULTS AND DISCUSSION

The global articulation rate (GAR) value for the 101 speakers (M1-100 and M101) and the mean articulation rate value across memory stretches for each speech sample (LARmean) were calculated. In Figure 2-3, results are shown in form of histograms that stand for how many speakers lie within a particular interval of GAR and LARmean values.

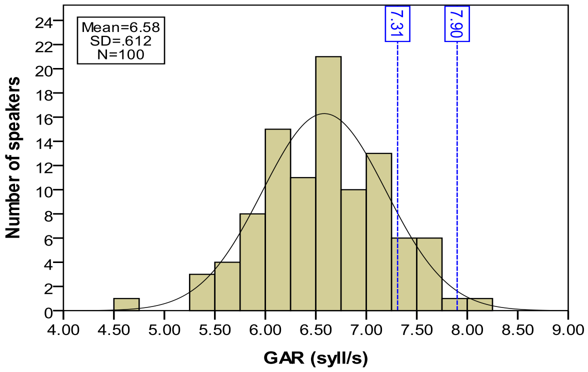


Figure 2: Histogram for GAR parameter. The blue dashed lines stand for the range of GAR values of M101 (see below).

Illustrated in Figure 2, the statistical results of the parameter GAR values of speakers M1-100 show an approximate normal distribution. For example, the values from 6.50 to 6.75 syll/s are found in many more speakers (21, 21% of

[1] As recommended by one reviewer, two possible types of speech intervals can be chosen for calculating AR, which are the "inter-pause intervals" and the "intonation phrase". However, they are "not without empirical or methodological problems"; the present method is simpler and more pragmatic (for more details see [7]).

100 speakers) than values at the lowest and highest margins of the distribution (both found in only one speaker, 1% of 100 speakers). This result provides valuable reference data for Chinese population statistics in forensic casework. Based on this statistical result, the GAR parameter does not successfully discriminate some of the speakers with GAR values in the central area. However, for those speakers who strongly deviate from the central trend, the GAR becomes a salient discriminatory parameter.

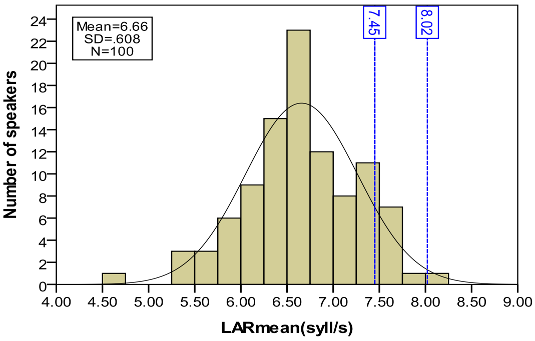


Figure 3: Histogram for LARmean parameter. The blue dashed lines stand for the range of LARmean values of M101 (see below).

Figure 3 shows that the result of the LARmean values of M1-100 also form a good approximation of a normal distribution. Not surprisingly, the distributions of GAR values and LARmean values are similar. Comparatively, 23 speakers are found in the center of the LARmean distribution (values from 6.50 to 6.75 syll/s). One speaker appears at the margin of each distribution. However, the two distributions are not exactly the same. Across the 100 speakers, mean values of GAR and LARmean are 6.58 syll/s and 6.66 syll/s respectively. The former is a little lower. After examining the two groups of data, an interesting result is found.

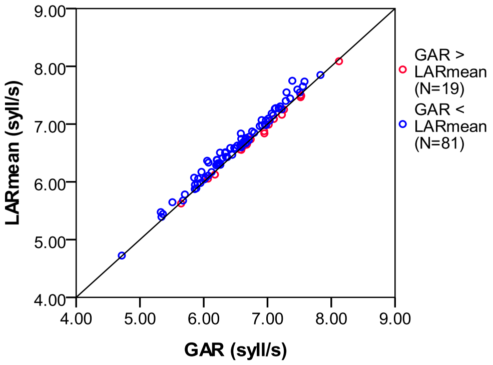


Figure 4: Scatterplot for GAR and LARmean values.

As shown in Figure 4, 19 speakers' GAR values are higher than their LARmean values (all difference between them are less than 0.07 syll/s), whereas the other 81 speakers' GAR values are lower than their LARmean values (the difference range from 0.00 to 0.36 syll/s with an average of 0.10 syll/s). This explains why the two distributions are different, e.g. in the range from 6.00 to 6.25 syll/s, the numbers of speakers are 15 and 9 in figure 2 and 3 respectively. Pearson correlations are run in order to determine whether and to what extent the GAR values and the LARmean values correlate with each other. A significant positive correlation was found (r=0.990, p=0.01) between the two AR values. However, given the difference between the two different calculating methods, it is unwise to mix them in one case at the same time, and the GAR values and LARmean values should not be compared.

Table 1: Literature on AR (syll/s) of male speakers in three languages (L). G – German, E – English, C – Chinese, Spont – spontaneous.

Study	Men	AR mean	L	Speech
[5]	27	5.74	G	Reading
[6]	5	5.89	G	Spont talk
[7]	100	5.19	G	Spont phone
[9]	47	5.2	E	Informal talk
[8]	2	5.65 [2]	C	Reading
present	100	6.58/6.66	C	Spont phone

The data in the previous studies of AR for German, English and Chinese are presented in Table 1. The number of male speakers in the present study is more than [5,6,8,9], excepted for [7]. The major difference among these studies in the literature lies in the average value of AR (GAR or LARmean). The present results are the highest in all studies. The difference may be caused by factors such as number of speakers, age of speakers, calculating method, language and speech style. Interestingly, as we follow the method in [7], the differences between [7] and the present results are still significant (both in the average values and the whole distribution). Compared with German and English, Chinese syllable structure is simple. The maximal Chinese syllable construction is #CVVC/V# and there are no consonant clusters [10]. Comparatively, a syllable in English/German can contain up to three consonants at the beginning, as in *stray/strick*, and

[2] The value 5.65 syll/s is the average value of the two speakers' AR values (5.3 syll/s for male 1 and 6.0 syll/s for male 2) in Cao [8].

up to four consonants at the end, as in *glimpsed/herbst* [10,12]. As pointed out in [7], "speakers of a language with simple syllable structure are expected to produce more syllables … than speakers of a language with more complex syllable structures, hence show higher AR", we can infer that except for Cao [8], language may be the most critical factor. As it is known that the syllable structures of Japanese and many African languages, like Yoruba, are less complex than Chinese, to find whether or not AR data of these languages speakers will be lower than the present result, more researches are needed. The low AR value of Cao [8] is easily explained since only silent pauses were excluded when the AR was measured and the subjects were confined to two male speakers.

Table 2 lists GAR and LARmean values of M101's 10 different speech samples. The minimal and maximal values of GAR and LARmean are illustrated in figure 2 and 3 (see the blue dashed lines) respectively. The ranges of the GAR and LARmean values are both relatively centralized and both occupy the position near the top margin of the two distributions. Because the topics and styles of the 10 speech samples are similar, the intra-speaker variations of GAR and LARmean are relatively stable and (in this case) smaller than the inter-speaker variations shown in figure 2 and 3. The intra-speaker variation may be larger if the styles of the speech samples differ. Forensically, when AR parameters are used, it is critical to get the most stylistically similar speech samples (including other possible factors, such as emotional factors), compared with the unknown samples.

Table 2: A list of GAR and LARmean values of M101's 10 different speech samples.

Number	GAR	LARmean
1	7.52	7.52
2	**7.31**	**7.45**
3	7.87	7.96
4	7.55	7.73
5	7.43	7.76
6	**7.90**	**8.02**
7	7.55	7.66
8	7.86	7.80
9	7.85	7.69
10	7.100	7.80
Mean	7.74	7.64
SD	0.17	0.21

CONCLUSIONS

This study provides valuable population statistics on Chinese speakers' articulation rates. Two histograms are shown for the articulation rates (GAR and LARmean) of 100 male Chinese speakers, which show approximate normal distributions. Our findings are not very similar with previous data in German or English, presumably because the syllable structure in Chinese is simpler than that in German and English. Both GAR and LARmean parameters, which are significantly correlated, can discriminate individual speakers. However in the present study, it is hard to estimate which one is more powerful in discriminating individuals. 10 different spontaneous speech samples of one speaker, of which the topics and styles are similar, were investigated. The results show that the intra-speaker variation of AR is relatively stable and lower than the inter-speaker variation. In forensic casework the investigators should pay attention to the possible mismatch in stylistic factors, which may cause high intra-speaker variation. Since more variables have to be included as shown in [9], this study is also a platform for further investigation.

4. ACKNOWLEDGEMENTS

This research was funded by the China National Natural Science Funds (Grant 61073085). We thank two anonymous reviewers for the comments on the earlier version of the paper. And we also thank Michael Jessen for providing us the article [6].

5. REFERENCES

[1] Tsao, Y.C., Weismer, G., Iqbal, K. 2006. Interspeaker variation in habitual speaking rate: Additional evidence. *JSLHR*, 49, 1156–1164.

[2] Goldman-Eisler, F. 1968. *Psycholinguistics. Experiments in Spontaneous Speech*, Academic Press, London

[3] Miller, J.L., Grosjean, F., Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41, 215–225.

[4] Laver, J. 1994. *Principles of Phonetics*, Cambridge University Press: Cambridge.

[5] Künzel, H.J., Masthof, H.R., Köster, J.P. 1995. The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition, *Science and Justice*, 35, 291–295.

[6] Künzel, H.J. 1997. Some general phonetic and forensic aspects of speaking tempo, *Forensic Linguistics*. 4, 48–83.

[7] Jessen, M. 2007. Forensic reference data on articulation rate in German. *Science and Justice*, 47, 50-67.

[8] Cao, J.F. 2003. Articulation rate and its variations (in Chinese), *Proceedings of the 6th National Conference on Modern Phonetics*, Tianjin, 143-148.

[9] Jacewicz, E., Fox, R.A., O'Neill. C., Salmons, J. 2009. Articulation rate across dialect, age, and gender, *Language Variation and Change*, 21, 233–256.

[10] Wang, H.J. 2008. *Chinese Non-Linear Phonology* (in Chinese), Peking University Press, Beijing.

[11] http://www.fon.hum.uva.nl/praat/ download_win.html

[12] Fox, A. 2005. *The structure of German*. 2nd edition. Oxford University Press, Oxford.

English Learners' Phonation in Chinese Narrow Focus Syllable "-/a/ / C_"

YANG Jin [1], *KONG Jiangping*

1 Dept. of Foreign Languages
USST, SHISU, PKU
Shanghai, China

Abstract

This paper conducts acoustic experiments from a signal processing perspective, to examine the "-/a/ /C_" syllable narrow focus processing strategy of English L2 learners. The research discusses F0, OQ and SQ as key phonatory acoustic parameters. L2 strategies are examined by measuring deviations from L1 group to L2 group. The matrix distance reflects L1's traces in the interlanguage system.

Keywords: F0, OQ, SQ, narrow focus, L2

1. Introduction

As in the past, phonetic researches on narrow focus are mainly focused on the acoustic parameters such as F0, duration and amplitude. Ladeforged (1982) and Fant (2004) claimed that power is an important parameter for judging voice quality and is closely related with F0 change. Chen (1974) found that the average pitch range of the four Chinese speakers was at least 1.5 times wider than that of the four English-speaking subjects when they spoke their native languages. Zhang et al. (2008) reported Mandarin speakers' production of lexical stress contrasts in English is influenced partly by native-language experience with Mandarin lexical tones, and partly by similarities and differences between Mandarin and English vowel inventories. There are also a few researches on the acoustic parameters of Chinese focus types. However, there seems to be little research on the narrow focus perception studies, not to say phonation studies. To the author's knowledge, only Yin (2011) used a nine-syllable Chinese sentence to test sentence final narrow focus processing strategy. Within the scope of Chinese sentences of fixed length, his result showed that male speakers mainly process narrow focus by raising high frequency energy while female speakers mainly realize narrow focus by raising low frequency energy. Yin's research is a pioneering work in Chinese L1 speakers' phonation pattern on narrow focus. Based on this, our study intends to make one step further. We will add two variables into the current study, i. e., L2 speakers and sentence length. We intend to follow Kong's (2001) definition of OQ and SQ in his research on Chinese tones and regards F0, OQ and SQ as three key parameters in locating English and Chinese "-/a/ / C_" Structure narrow focus as an important prosodic characteristics and therefore conducts experiments on English L2 learners of medium-high level. And three-dimensional Matlab plots of cross language F0, OQ and SQ are presented for the first time on narrow focus processing. We hope this crisscross study will shed light on the phonation nature of speech production and L2 phonetic acquisition.

2. Method

2.1 EGG

EGG equipment is composed of several components. High Frequency Oscillator is a current-controlled oscillator to yield micro high-frequency current. The current yielded passes through the Electrode Circuit that in our case is glottis and vocal folds. Then there is Automatic Gain Control to ensure a stable and appropriate sized electric signal feedback. At the output end is AM Detector to detect output signal. Output signal at this place is Gx signal which usually contains low-frequency jitter. Due to the fact that most EGG has set High-Pass Filters, Gx signal passed through this filter and changes into Lx signal. Researches show that EGG signal is positively related with vocal folds attack area. When vocal folds attack area (hereafter referred to as VAA) grow, EGG signal becomes stronger. It weakens while VAA decreases. This has been proved by high-speed glottis photography (Fourcin 1974, Baer, T. et al. 1983 & Gilbert, H. R. et al., 1984).

This paper is sponsored by USST Social Science Project 11XSY22, Shanghai Municipality Key Course "Translation" & "Interpretation for Science & Technology" 1K10305001 Project, and National Social Science Junior Project 09CYY017.

2.2 Parameters

F0 or pitch, is the inverse of a vocal folds vibration period. OQ refers to Open Quotient, is the opening phase of the glottis divided by the period as indicated in this formula – OQ = ac / ad. SQ refers to the Speed Quotient, is the opening phase divided by the closed phase as indicated in this formula – SQ = ab / bc. In the following prototypical glottis airflow signal illustrated in Fig. II.B.1 (courtesy of Kong, 2001), which could be analogical to glottis area signal, the abscissa axis indicates a timeline while the vertical axis indicates vocal folds attacking area. The whole figure reflects a change of glottal airflow or glottal area along the timeline. Point 'a' and 'd' stand for the initial point of glottis opening phase where glottis airflow value augment from zero. 'b' is the point when the glottal width reaches a peak value, thus corresponding to a peak area of glottis. 'c' represents the closure of vocal folds when glottal airflow resets to zero. Period 'ac' is the opening phase of glottis while 'cd' is the complete closure phase and the glottal airflow or glottis area is zero. Contacting event and de-contacting event are highlighted as the beginning of a complete closure of the glottis and the opening of the glottis. The area in-between is the closing phase of the glottis.

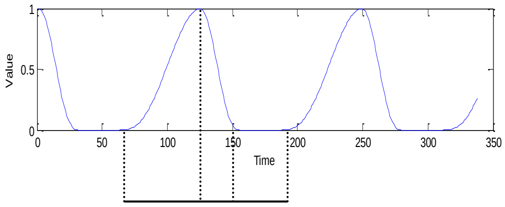


Fig. II.B.1 EGG signal

Another term is CQ – Contact Quotient in EGG signal processing. CQ refers to the ratio of glottis closure phase to the period. Fig. II.B.2 (courtesy of Kong, 2001) is a typical EGG signal. Its abscissa is time and the vertical axis is vocal fold contact area.

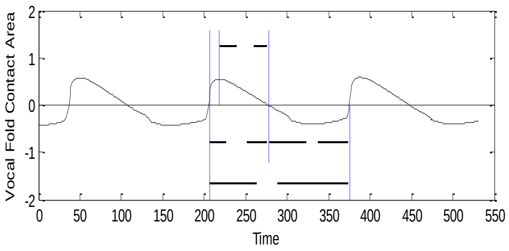


Fig. II.B.2 Defining EGG Signal

In this figure, A stands for a vocal fold vibration cycle, starting from the closing point of the glottis to the next closing point. It covers a period between two contacting events. B signifies a closed phase. It covers the period between the initial of glottis closure phase and the opening point. It's between contacting event and de-contacting event. C represents an open phase. It's between the glottis opening point to the next closure point, i. e., the period between de-contacting event to the next contacting event. D is a closing phase starting from the closure of glottis to vocal folds contact area augments to a peak value. E is the opening phase starting from the maximum contact area to the open point of the glottis. Therefore in this figure, OQ (Open Quotient) = C/A*100%. CQ (Contact Quotient) = B/A*100%. SQ (Speed Quotient) = E/D*100%. Since B + C = A, OQ + CQ = 1. In this sense, the formulas also state the complementary property of OQ and CQ (Henrich 2004).

From the above two illustrations, we can easily achieve this result – sound wave is in a reverse phase to EGG signal. This is because EGG signal is obtained via the resistance between the pair of vocal folds, i. e., it's determined by the opening degree of the glottis. Therefore, this will lag a semi-period to the sound wave which is yielded from vocal folds attack.

2.3 Participants

Four English and Chinese bilinguals participated in this experiment. They are two male and two female native American English speakers and two male and two female native Beijing dialect speakers. The four American English speakers were born and grow up in California in local American families. The four Chinese speakers were born and grow up in local Beijing families. They speak Beijing dialect whose phonetic system is the basis of Standard Mandarin Chinese, i. e., Chinese Putonghua, the official language of China. All of them are students recruited from Peking University. Their age ranges from 20 to 24. All American students have been studying Chinese for two to four years. Their Chinese proficiency is skilled. Since oral proficiency is not surely positively correlated with written language proficiency, we invited Oral English teachers to evaluate their spoken proficiency. The American students are of medium-high Chinese level.

None of subjects reported having any speech disorders. All subjects received payment.

2.4 Materials

Forty-four sets of question/answer pairs were constructed for this experiment. Questions were designed to set a disambiguating context to point out clearly the sentence final narrow focus marking location. The subject will click on the mouse to continue PPT questions and answer orally. Answers are all statements. All narrow focused syllables are composed of a truly existing consonant followed by /a/, thus 22 syllables in Chinese. Sentences are arranged randomly to appear before subjects, each appear totally for twice to ensure a reliable outcome.

Tab. II. D. 1 contains all the token forms. Ch_l means Chinese default last syllable at the long sentence. Ch_lf means Chinese narrow focus at the last syllable of the long sentence. Ch_s and Ch_sf refer to their counterparts in Chinese short sentences.

Ch_l	*Zhang Xiaosan zai wei qiang bian de kongdi shang xie le* “-/a/ /C_”. – (English equivalence: *Zhang Xiaosan* wrote “-/a/ /C_” on the ground by the wall.)
Ch_lf	*bu, Zhang Xiaosan zai wei qiang bian de kongdi shang xie le* “-/a/ /C_”. – (English equivalence: No, *Zhang Xiaosan* wrote “-/a/ /C_” on the ground by the wall.)
Ch_s	*wo xie de shi* “-/a/ /C_”. – (English equivalence: I wrote “-/a/ /C_”.)
Ch_sf	*bu, wo xie de shi* “-/a/ /C_”. – (English equivalence: No, I wrote “-/a/ /C_”.)

Tab. II. D. 1 Read Speech Material

2.5 Procedure

Trained proctors recorded all speech data. Questions appeared randomly in a laptop computer. Subjects wore Sony clip microphone, EGG neck belt, breast belt, abdomen belt, finger voltage collector and heart pulse collector. The Sony clip microphone is all-directional and located about 15 cm away from their mouths, and they were instructed to speak naturally at a normal rate and volume. Recording are carried out in a sound-attenuated room. The main recording equipment is a 16 bit Myoelectrigraph & Electroencephalograph Information Gatherer produced by Australia Powerlab Company. The recording software is Chart 5 which is carried by the equipment itself. Recording collects two channels of signals: 1) sound file collected by Sony microphone; 2) graph collected by EGG produced by Kay Corporation.

2.6 Results

Altogether 22*8*2*2 = 704 tokens are obtained for analysis. All tokens are normalized to 30 points before averaged according to F0, OQ and SQ separately. Results are shown in the following figures. In each graph, a red line denotes the “-/a/ /C_” syllable cumulative average value at the sentence final narrow focus place. The blue line is for their counterpart in the non-narrow focus sentence final position. Legends are indicated in this model such that “mA_ch_lf_F0ave” indicates male American English speakers’ average F0 across all “-/a/ /C_” syllables at the narrow focus position. “mC” group are male Chinese speakers. However, considering limit space of this paper, we’ll only extract English and Chinese male speakers’ reaction to the stimuli of Chinese long sentence. Other data are shown in Tab. II. F. 1 by function coefficient.

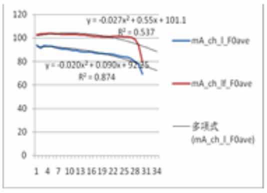


Fig. II. F. 1 mA_ch_l_F0

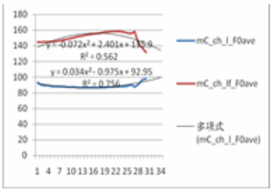


Fig. II. F. 2 mC_ch_l_F0

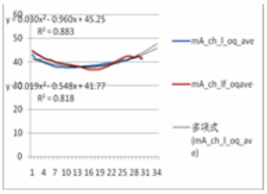


Fig. II. F. 3 mA_ch_l_OQ

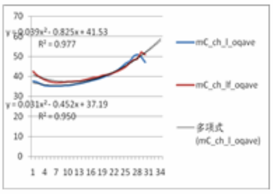


Fig. II. F. 4 mC_ch_l_OQ

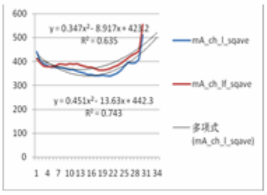


Fig. II. F. 5 mA_ch_l_SQ

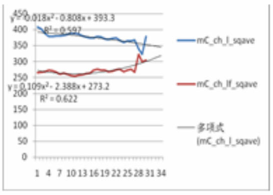


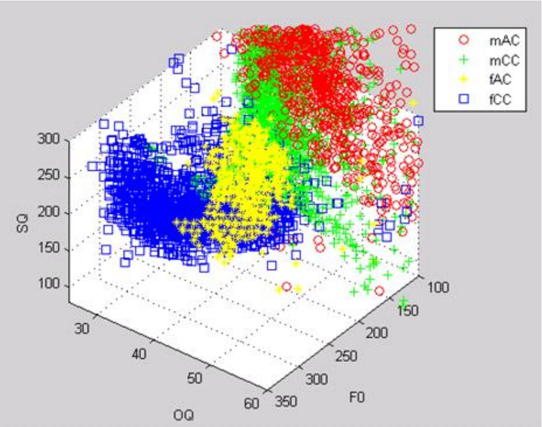
Fig. II. F. 6 mC_ch_l_SQ

For other data, different levels of variables are fitted into a curve function y = cx2 + tx + v (c stands for curvature, t for tilt rate and v for intercept). In Tab. II. F. 1, Ch_l stands for Chinese long sentences which contain fourteen syllables. Ch_s stands for Chinese short sentences which contain five syllables. mAf is male American speakers' narrow focus token. mA is male American speakers' default token (or, token at non-narrow focus position). Collected data are as follows.

	F0			OQ			SQ		
	c	t	v	c	t	v	c	t	v
Ch_l									
mAf	-0.027	0.55	101.1	0.03	-0.96	45.25	0.347	-8.917	423.2
mA	-0.02	0.09	92.25	0.019	-0.548	41.77	0.451	-13.63	442.3
mCf	-0.072	2.401	135.9	0.039	-0.825	41.53	0.109	-2.388	273.2
mC	0.034	-0.975	92.95	0.031	-0.452	37.19	-0.018	-0.808	393.3
fAf	-0.057	1.8	241.1	0.006	-0.115	50.62	-0.087	2.582	161.7
fA	0.017	-1.716	198.8	-0.019	0.656	52.98	-0.099	2.263	198.3
fCf	0.042	-0.874	228	0.001	-0.067	35.07	0.127	-2.817	191.2
fC	0.149	-5.147	193.6	-0.022	0.917	36.21	-0.104	3.6	197.1
Ch_s									
mAf	-0.019	0.37	111.9	0.031	-0.882	44.85	0.014	1.761	337.1
mA	-0.019	0.017	90.66	0.027	-0.689	41.58	0.389	-10.83	372.7
mCf	-0.008	0.752	128.5	0.04	-0.757	40.89	0	0.061	249.3
mC	0.052	-1.54	108.3	0.062	-1.36	40.9	-0.113	4.071	270.9
fAf	-0.103	3.365	257.3	0.004	0.029	49.89	-0.04	0.729	154.5
fA	0.012	-0.713	236.1	0.003	-0.132	51.3	-0.101	3.013	171.5
fCf	0.017	-0.068	222.5	0.01	-0.34	35.68	0.018	-0.184	180.5
fC	0.057	-2.65	198.3	-0.012	0.496	36.28	-0.021	2.069	164.8

Tab. II. F. 1 F0, OQ and SQ parameters

F0, OQ and SQ are also put into a three dimensional plot edited by Matlab software for convenience. In Plot II. F. 1, the red circles represent male American speakers Chinese speech signal, briefed in the legend as "mAC". The green plus represent male Chinese speakers' data, the yellow diamonds for female American speakers' and the blue squares for female Chinese speakers' signal.



Plot II. F. 1 Matlab 3-D distribution of F0, OQ and SQ

3. Analysis and discussion

First, F0 has a robust effect in narrow focus representation, despite L1 or L2, sentence length or sex difference. For details, the male American speakers' F0 variation seems smaller than the other groups. In other words, the difference between the highest and the lowest value is comparatively smaller than the other groups. This is true for either at narrow focus position or not.

The common feature of male and female Chinese speakers F0 variation proves that Chinese speakers distinguish clearly between narrow focus and non-narrow focus syllables. The common feature of female Chinese and English speakers F0 variation proves that female speakers share a good distinction on narrow focus. The male American speakers are neither L1 speakers nor female, therefore they are weak in F0 variation for a good narrow focus.

Sentence length is found contributive to narrow focus identification. We found in short Chinese sentences containing five syllables, F0 variation is not obvious as in long sentences. This may due to the fact that each syllable weighs more in the short sentence unit than the long sentence unit (the weight 1/5 > 1/14). This weight is correlated with contribution. So the rest four syllables in the short sentences are not so identifiably different from the final syllable "-/a/ / C_", thus causing less F0 variation.

Second, OQ distinguishes sex. Male speakers, American or Chinese, perform similar OQ curves at either narrow focus or not. Male OQ has an obvious rising tilt rate. Female speakers, on the other hand, perform typical OQ difference on narrow focus difference. Their OQ is lower at narrow focus position and higher at non-narrow focus position.

Chinese speakers, no matter sex, have a little bigger OQ variation range than Americans.

All male speakers do not vary much at OQ values. All females vary, though. Females' OQ are lower at narrow focus position. But the variation is not much, only slightly bigger than male speakers. Meanwhile, although females' OQ differ at narrow focus position, they do not take a rapid growth as males do. We observed only one exception – the female American speakers, when reading Chinese short sentences, suddenly rise from the one fourth point along the time axis. This might be the result of lab effect. We suppose the subject may psychologically hint herself of reaching a supposed expectation of experiment purposes.

The male Chinese and female American speakers rise more rapid in Chinese short sentences than long sentences. Since the short sentence only contains five syllables. Subglottal air pressure is still very strong at the sentence final position. There are still plenty of energy causing a rapid growth of OQ.

Third, SQ is stable on most occasions, differing only slightly for gender difference. Male American speakers' SQ are mostly stable but rise rapidly at the rear end. The rest three groups are fairly stable. Male Chinese speakers and female American speakers have a slight decline at the rear end of SQ. They fluctuate a little and don’t have rapid changes. However, this could not prove that male Chinese speakers are more close to female American speakers. In Chinese long sentences, all females decline at the rear end. But the male Chinese speakers SQ fluctuate and have a rising tilt rate at the rear end. In this sense, we believe all male speakers’ SQ rise fairly obviously at the rear end but Chinese’s rise is moderately lower in degree.

Above we have mentioned changes. Now let’s move to have a closer observation of the exact values. Males’ SQ is between 350-400 while females’ is about 250. This difference reveals that males have more low frequency energy than females.

Two male groups have higher SQ value in long Chinese sentences than short ones. This proves that at long sentence final position when subglottal air pressure declines, energy decline rapidly, therefore male compensate with a higher vocal folds vibration frequency. This is a male speaker strategy different from female speakers. However, we will not exclude another possible cause in this process. We’d like to report a byproduct in our experiment. We have observed a double peak phenomenon at the rear end of “-/a/ / C_” syllable. Since vocal folds vibrate quasi-periodically, frequency should also be fairly stable. Therefore double peaks within one period might be due to the ventricular folds vibration. Ventricular folds are above the vocal folds. When speech proceeds to the ending of speech, the subglottal air pressure declines to a minimum and the glottis begins to fall. The vocal tract withdraws downward. This will cause vocal folds to close first and pulls the ventricular to vibrate for a couple of times. Sometimes on other occasions ventricular vibration is intentionally used as a specific artistic phonation pattern as in Noh, a traditional Japanese opera.

Another surprising finding is a compensatory effect of OQ, SQ to F0. In other words, if F0 at narrow focus is higher than at non-narrow focus position, then a lower OQ, or a lower SQ, or a combination of both OQ and SQ, is witnessed across almost all conditions in our experiment. The only one exception is male American speakers. Even for them, the proposition is only moderately violated. They perform similarly at OQ and SQ. Hence according to this proposition we infer a similar F0 performance. And we do witness a smaller F0 variation range than all other groups.

For the rest three groups, compensatory effect are obvious. Male Chinese speakers have a stable OQ, so their F0 and OQ is negatively correlated, i. e., compensatory distributed. All female groups have lower OQ and higher F0 at narrow focus position. Since OQ varies not in a symmetrical degree as F0, their SQ adds weights on the scale for the smaller OQ variation. SQ and OQ together compensate F0 variation.

Female Chinese and American speakers’ F0 difference (F0 average peak minus low) is very close in the two positions:

f1_ch_lf_F0ave – f1_ch_l_F0ave = 73.2526
f2_ch_lf_F0ave – f2_ch_l_F0ave = 66.9312
f1_ch_sf_F0ave – f1_ch_s_F0ave = 47.9098
f2_ch_sf_F0ave – f2_ch_s_F0ave = 51.7211

Male and female Chinese speakers F0 difference in Chinese syllables are stable.

American speakers judge by pitch height, not only pitch height will decrease, but also in uttering Chinese. First they will emphasize on the first high-level tone and pinpoint the Chinese syllable pitch height at a stable level without much tilt rate. Then they come to the second step – to decide F0 height. But they are not sure. So the usual strategy is to narrow F0 difference between narrow and non-narrow focus. Other possible strategies are that they’ll first fix the narrow focus F0 height, then try to extend to a lower F0 height for non-narrow focus syllable. But this time again they are not sure. Or to fix a low F0 height, then extend to a higher F0 for narrow focus correspondent. They are not so good at distinguishing L2 narrow focus because they are not sure of pitch height, despite the fact that all examples in our case is level tone and excluded the tone effect. That is, even for level tones, L2 learners are not sure of pitch height in our case. Their strategy is to resort to a safer smaller range to erase the Chinese specific F0 range. They cannot judge F0 height.

Our research result partly conforms to Yin (2011). He studied Chinese L1 speakers’ performance for a nine syllable length material, with the same CV structure. He reported female F0 range of 80Hz, OQ rising about 7% and SQ decreasing tremendously. Our result conforms to his F0 part and contradicts to OQ and SQ data. Yin also reported males’ performance not that obvious as females. F0 raise but range is smaller than females’. It’s only 25Hz. Males’ OQ raise slightly and less than females. We haven’t observed such. This might owe to the different sentence length. We observed male speakers’ SQ raising slightly which conforms to his study.

4. Acknowledgment

Sincere thanks go to Prof. Kong Jiangping, Phonetics Lab, PKU, who have provided experimental surroundings, and Mr. Ye Zehua, who have developed Matlab programming for this study platform. Without these, this study is impossible.

5. References

[1] Baer, T., Lofquist, A., McGarr, NS. 1983. Laryngeal Vibrations: A Comparison between High-speed Filming and Glottographic Techniques, *Journal of the acoustical society of America* 73: 1304-1308.

[2] Chen, Gwang-tsai. 1974. The pitch range of English and Chinese speakers. in *Journal of Chinese Linguistics*, Vol 2, No 2.

[3] Fant, G. and Kruckenberg, A. 2004. Analysis and synthesis of Swedish prosody with outlooks on production and perception, In G. Fant, H, Fujisaki J Chao and Y. Xu (Eds.), Festschrift Wu Zongji, From traditional phonology to modern speech processing, pp. 73-95. Beijing: Foreign Language Teaching and Research Press

[4] Fourcin, A.J., 1974. Laryngographic Examination of Vocal Fold Vibration, Ventilatory and Phonatary Control Systems (edited by Wyke, B.), Oxford, New York, 315-333.

[5] Gilbert HR, Potter CR, Hoodin R., 1984. The Laryngograph as a Measure of Vocal Fold Contact Area. *Journal of Speech and Hearing Research,* 27, 178-82.

[6] Henrich, N., C.d'Alessandro, B.Doval, and M.Castellengo. 2004. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of the acoustical society of America* 115, No.3, 1321-1332.

[7] Kong, J. 2001. On language phonation. Beijing: Central Ethic Groups University Press.

[8] Ladefoged, Peter. 1982. A Course in Phonetics. San Diego: Harcourt Brace Jovanovich Publishers.

[9] Ye, Z. 2010. A tense and lax phonation study of Xinping Yi language [D]. Beijing: Peking University.

[10] Yin, J. 2011. A phonation study of Chinese prosody [D]. Beijing: Peking University

L1 and L2 Respiratory Patterns of Chinese and English Bilinguals' Read Speech

*YANG Jin, KONG Jiangping**

Abstract

This paper investigated the respiratory patterns of native English and Chinese's bilinguals in the read speech material of both languages. Experiments are conducted by respiratory belts, to measure breast breath signal, abdomen breath signal and the accumulated breath signal of the two. Results showed that speakers tend to adopt different breath patterns in L1 and L2 and tend to borrow their L1 respiratory pattern into their L2 processing strategy. Results also show that this borrowing seems to be positively correlated with their L2 oral proficiency. These findings have implications for respiratory signal processing, prosodic typology, bilingualism and second language acquisition.

Keywords*:* respiratory pattern; rhythm; respiratory belt

1. Introduction

Since the introduction of the Pitch Accent Theory (Bolinger, 1958) [1], most researchers have come to an agreement that pitch is the most important factor in perceiving heaviness. The nowadays popular ToBI labeling system is itself based on this framework. The main trend has been regarding rhythm as LH alternations. In other words, rhythm refers to the regular strong-weak, or, long-short alternation pattern. Language rhythm is the periodical impulses at suprasegmental level. This regular pattern is otherwise one fundamental property of prosody.

Rhythmic patterns vary across languages. Lloyd James (1940, cited by Pike, 1945) [2] attributed the prosodic difference between Spanish or Italian and English or Dutch. He used the metaphor "machine-gun rhythm" for the first group of languages and "Morse code rhythm" for the second. Actually it also depends on different prosodic levels, how they're defined and how they're organized together. In Chinese Putonghua, most researchers have a consensus that there is a rhythmic unit a level above syllable – two-syllable cluster or three-syllable cluster. The earliest studies called feet and suprafeet respectively (Chen, 2001 and Shih, 1986) [3] [4]. Another view holds two syllable cluster as the standard feet, three-syllable cluster as suprafeet, residue feet which contain a neutral tone in a two-syllable cluster while degenerated feet as monofeet (Feng, 1997) [5]. Some other researchers called it "minimal rhythmic unit" (Chen, 2001, Shen, 1985) [3] [6]. Wang Hongjun (2002) [7] summarized Chinese metrical pattern as "Two syllables are normal feet; three syllables are tolerable; while one or three syllables are marked" at the foot level which affirmed Shih Chilin's dynamic foot division rules – IC foot, DM foot and supra f'1. Shih claimed in her PhD. dissertation that IC is immediate component; DM as Duplex Meter where unpaired two syllables are connected from left to right into a di-syllable foot but not two components that are not branching at the same syntactic direction. Supra f' is a supra foot that is a combination of the rest monosyllable and its neighboring disyllable, also in conformity with the syntactic branching direction. Wang's claim indicated that Chinese has different rhythmic pattern from English.

Based on the above-mentioned phonologists' researches, it is possible to infer that speakers will adopt different respiratory patterns in uttering Chinese and English separately.

Breath signals reflect speech respiratory patterns and speakers' plan of speech content. Z. H. Hu in Ming Dynasty claimed in his masterpiece *A Handbook of Tang Dynasty Prosody* that "To minimize to three characters is too vague while to maximize to nine characters is too tight". Since one Chinese character corresponds to one syllable, Hu suggested that a good sentence should contain no more than nine syllables and no less than three syllables [8]. In this sense, five to eight syllables are ideal prosodic structure for a sentence. Prosodic units are quantified as time reflection such as pauses or duration. So segment time index is a major acoustic reference. Phonetic research focuses has been put on duration, mainly on syllable duration at the boundary of levels of prosodic boundaries and pause duration (Y. F. Yang, 1997, B. Wang, Y. F. Yang and S. N.

*To whom correspondence should be addressed: jpkong@pku.edu.cn. This paper is supported by China Social Sciences Funds of the Ministry of Education 10JJD740007, USST Social Science Project 11XSY22, Shanghai Municipality Key Course "Translation" & "Interpretation for Science & Technology" 1K10305001 Project, and National Social Science Junior Project 09CYY017.

Lv, 2004, Z. Y. Xiong, 2003) [9] [10] [11].

Thus an experiment is conducted by using respiratory belts to measure the respiratory patterns of Chinese and English bilinguals either at Chinese or English read speech.

2. Method

2.1 Participants

Four English and Chinese bilinguals participated in this experiment. They are one male and one female native American English speakers and one male and one female native Beijing dialect speakers. The male native American English speaker is a graduate student of Chinese at Peking University and the rest three speakers are undergraduate students of Peking University at the time of the recording. The male native American English speaker has quite high Chinese proficiency and is therefore regarded as advanced learner of Chinese. The female native American English speakers has arrived at Peking University for a year and two months to study Chinese. She is regarded as a preliminary learner of Chinese. The two undergraduate Chinese students are local Beijingers. They speak Beijing Mandarin whose phonetic system is the basis of Standard Chinese, Chinese Putonghua, the official language of China. As for the English proficiency of Chinese undergraduates, since oral proficiency is not surely positively correlated with written language proficiency, we invited Oral English teachers to evaluate their spoken proficiency. The male undergraduate is judged as mid-low level and the female undergraduate mid-high level, hence preliminary and advanced in accordance with the English speakers.

Each subject recorded two sets of data, in Chinese or English. None of subjects reported having any speech disorders. All subjects received payment.

2.2 Material

An English novel excerpt and a Chinese novel excerpt are selected as read speech material. The English material is taken from *Harry Potter “Chapter Six: The Journey From Platform Nine and Three-Quarters”*, 299 words in total.

The Chinese material is taken from *Miscellaneous Stars*, 198 words in total.

2.3 Recording Procedure

All recordings took place in a sound-proof room in the Phonetics Lab, Department of Chinese Language and Literature, Peking University. The main recording equipment is a 16 bit myoelectrigraph and electroencephalograph information gatherer produced by Australia Powerlab Company. The recording software is Chart 5 which is carried by the equipment itself. Recording collects four channels of signals: 1) sound file collected by Sony microphone; 2) graph collected by EGG produced by Kay Corporation; 3) abdomen breath signal collected by MLT1132 respiratory belt; and 4) breast breath signal collected by MLT1132 respiratory belt. The four channels of signals are used for analysis. EGG has no special requirement for sampling rate. In order to enable MATLAB software procedure to run fast, the sampling rate is fixed at 20 kHz and EGG signal is fixed as high-pass above 50Hz to avoid neck artery pulse disturbance.

2.4 Results & Analysis

Exact respiratory contour are shown in the following figures. Each figure contains three channels of signals – breast breath, abdomen breath and the accumulated two. The range of each window is the same 16s.

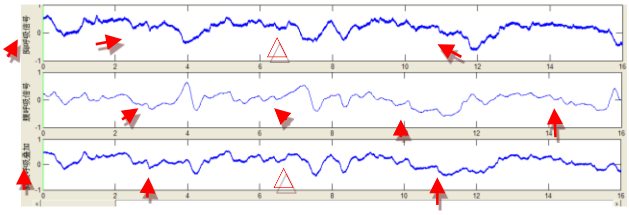


Figure 2. The male native American English speaker read English novel

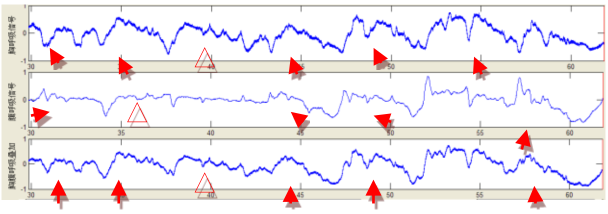


Figure 2. The male native American English speaker read Chinese novel

Fig. 1 and Fig. 2 are English and Chinese novel read by a male native American English speaker. In the same 16s range shown in the pane, Chinese material contain more breath resets, either breast or abdomen ones than its English counterpart. For each breath reset, the Chinese breath tilt rate is more rapid than the English one. The Chinese breath reset peak is also higher than the English one. Both figures contain low resets which indicate strong and sudden exhales.

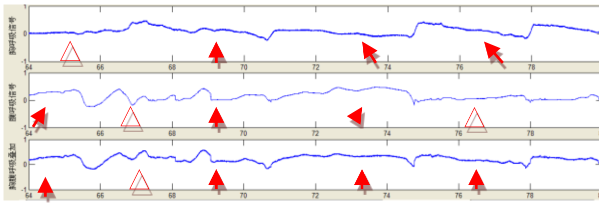


Figure 3. The female native American English speaker read English novel

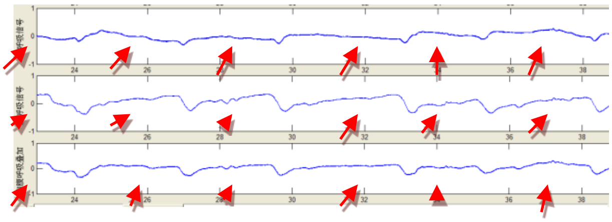


Figure 4. The female native American English speaker read Chinese novel

Fig 3 and Fig. 4 are English and Chinese novel read by a female native American English speaker. The comparison between the two figures showed a similar situation as for breath reset numbers. However, unlike the male native English speaker, the female speaker doesn't show a more rapid tilt rate and higher reset peak. Rather, both Fig. 3 and Fig. 4 are similarly flat. We later found this is due to the poor oral Chinese proficiency of the subject. Since in the same time range of 16s, she uttered less Chinese syllables, hence the slower tempo led to a more moderate tilt rate. Like the previous case, both figures contain low resets which indicate quite strong and sudden exhales.

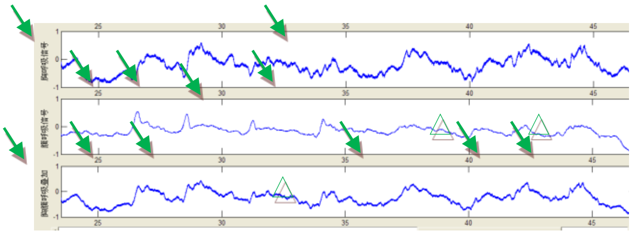


Figure 5: The female native Chinese speaker read Chinese novel

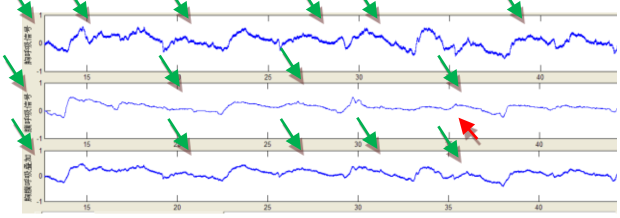


Figure 6: The female native Chinese speaker read English novel

Fig. 5 and Fig. 6 are Chinese and English novel read by a female native Chinese speaker. Again, Chinese material contain more breath resets, either breast or abdomen ones than its English counterpart. For each breath reset, the Chinese breath tilt rate is more rapid than the English one. The Chinese breath reset peak is also higher than the English one. Unlike native English speakers, low resets are not witnessed.

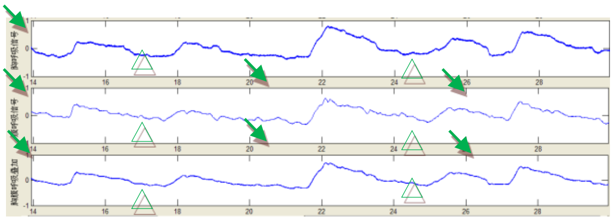


Figure 7. The male native Chinese speaker read Chinese novel

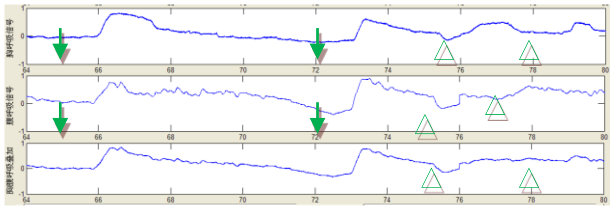


Figure 8. The male native Chinese speaker read English novel

Fig. 7 and Fig. 8 are Chinese and English novel read by a male native Chinese speaker. Again, Chinese material contain more breath resets, either breast or abdomen ones than its English counterpart. Inhale peaks are high and obvious. Very clear triangular-shaped tilt is witnessed. English material breath tilt rate is long and more flat than Chinese material. Only one low dent is witnessed.

3. Analysis and discussion

The respiratory patterns can be summarized in a general statistics shown in the following table, Tab. I. The eight rows record different speakers' signals, identified by the name. For example, "M_AmS_En" means male native American student reading English speech and "F_ChS_Ch" means female native Chinese student read Chinese speech. The three columns stand for signals from the three channels. BBRs stand for Breast Breath Resets, ABRs stand for Abdomen Breath Reset numbers and the third column stand for the accumulation of the two. The numbers in brackets are for minor ups or downs.

TABLE I. BREATH UNITS COUNT

	BBRs	ABRs	accumulation
M_AmS_En	4(1)	4	4(1)
M_AmS_Ch	6(1)	5(1)	6(1)
F_AmS_En	4(1)	5(2)	5(1)
F_AmS_Ch	6	6	6
F_ChS_Ch	7	6(2)	7(1)
F_ChS_En	6	4	5
M_ChS_Ch	5(2)	5(2)	5(2)
M_ChS_En	4(2)	4(2)	4(2)

Based on this table, we can reach the following results.

The first, breath has units. It can be divided into different tiers, from very small to very big, changing from an axis. As earlier researches (J. J. Tan, 2008) [12] has found there are three tiers of breaths in speech in a Chinese news read speech study, we here claim that there might be more tiers depending on different languages and the calculation accuracy.

The second, speakers adopt regular/fixed respiratory patterns in uttering speech of certain languages since numbers in the three columns don’t differentiate much.

The third, major and minor (or even more minor breath levels can be classified since we can broaden the time range in combination with a more detailed calculation of different levels of resents or peaks) breaths coexist, but not in a regular alternation pattern. It is inferred that this is correlated with different tiers of phonological units in a language typology, also decided by syllable numbers within a respiratory unit and tempo.

The fourth, males have more minor breaths than females, or in other words, males adopt more major-minor breath alternation in uttering speech.

The fifth, there are generally more breath units in Chinese read speech, which doesn’t vary across L1 and L2 speakers of either Chinese or English. It is inferred that Chinese is a syllabic language while English is a mora language.

The sixth, Chinese and English has very different respiratory patterns. Chinese is more highlighted by inhale resets (which indicates the inhaling moment) while English is more highlighted by exhale resets (which indicates the exhaling moment).

The seventh, Chinese takes a “Strides & Swift-paced” respiratory pattern while English takes a “Moderate-Equal Step” respiratory pattern. Namely, Chinese breast breath appears to be a skewed big triangle of major breath enclosing several small triangles of minor breaths. English takes almost equal length respiratory steps of normal distribution.

The eighth, speakers adopt a combination of L1 and L2 language respiratory pattern in uttering L2 read speech. This is due to a hybrid of L1 respiratory pattern and an unconscious process confined by L2 phonological and syntactical information.

Sometimes L2 effect transfers as is witnessed in Fig. 6 where the red arrow indicates that a typical exhale reset respiratory feature of English is adopted by the female native Chinese speaker while uttering English speech.

Finally there’s one point worth to mention. Since each level of learners, medium-high and medium low, contains only one speaker, it seems that sex factor is not effectively balanced in this experiment design. However we argue that since male and female speakers, if belong to a same L1 group or L2 group, do not vary much in their respiratory performance, sex factor does not play a significant role in this case.

Results show that speakers tend to borrow their L1 respiratory pattern into their L2 processing strategy. Results also show that this borrowing seems to be negatively correlated with their L2 oral proficiency. The higher the language proficiency, the less borrowing happens. We hope these findings have implications for prosodic typology, bilingualism and second language acquisition.

4. CONCLUSION

This respiratory belt study indicates that Chinese takes a “Strides & Swift-paced” respiratory pattern which contains obvious major and minor breaths while English takes a “Moderate-Equal Step” respiratory pattern. Chinese language reflects high breast breath inhale resets at the left boundary of prosodic segments while English language tends to have more moderate resets and lower exhale resets. Speakers tend to borrow their L1 respiratory pattern into their L2 processing strategy. Results also show that this borrowing seems to be negatively correlated with their L2 oral proficiency. The higher the speakers’ proficiency, the less borrowing there is. Thus it provided a phonation evidence for narrow focus phonetic representation as well as a quantified description of rhythmic patterns of Chinese and English.

However, there are also limitations and further requests for future research. We only include breast breaths and abdomen breaths in a same 16s range within one pane in our work by respiratory belt. Syllable numbers and tempo is not taken into consideration in our current thesis, and they are also significant prosodic referents. Statistics is another consideration in future research so as to verify our findings in a more technical and convincing way. The above-mentioned aspects are to be supplemented in our future research on respiration.

5. Acknowledgment

Sincere thanks go to Prof. Kong Jiangping, Phonetics Lab, PKU, who have provided experimental surroundings, and Dr. Yang Feng, who have developed Matlab programming for this study platform. Without these, this study is impossible.

6. References

[1] D. Bolinger, “A theory of pitch accent in English”, in Word, vol. 14, 1958, pp.109–49.

[2] K. L. Pike, The intonation of American English. Ann Arbor: University of Michigan PressSoquet, 1945.

[3] M. Y. Chen, Tone sandhi: patterns across Chinese dialects. Cambridge: Cambridge University Press, 2001.

[4] C. L. Shih, The Prosodic Domain of Tone Sandhi in Chinese. PhD. disseration. University of California at San Diego, 1986.

[5] S. L. Feng, Prosody, Morphology and Syntax. Beijing: Peking University Press, 1997.

[6] J. Shen, “Register and Intonation in Tones of Beijing Mandarin,” in Experimental Records of Beijing Mandarin Phonetics, T. Lin and L. J. Wang, Eds. Beijing: Peking University Press, 1985, pp. 75-107.

[7] H. J. Wang, “How the Metrical Boundary Relates to Metrical Pattern, Syntax, Pragmatics in Chinese Mandarin,” in On Linguistics, vol. 26, Beijing: Commercial Publishing House, 2002, pp.279-300.

[8] Z. H. Hu, A Handbook of Tang Dynasty Prosody. Classical Literature Press, 1957.

[9] Y. F. Yang, “Prosodic Representation at Syntactic Boundaries,” in Journal of Acoustics, vol. 5, 1997, pp.414-421.

[10] B. Wang, Y. F. Yang and S. N. Lv, “Acoustic Analysis of Chinese Prosodic Hierarchies’ Boundaries,” in Journal of Acoustics, vol. 29, 2004, pp.29-36.

[11] Z. Y. Xiong, Prosodic Features and Communicative Functions of Spontaneous Prosodic Boundaries. PhD. dissertation, China Academy of Social Science, 2003.

[12] J. J. Tan, “On Breath Resets of Mandarin Chinese News,” in Chinese Phonetics Journal, vol. 1, Beijing: Commercial Publishing House, 2008, pp. 20-24.

基于藏缅语的音位结构和负担量计算方法研究*

孔江平

1. 引言

随着人们利用基因研究人类演化的进展，已有比较充分的证据说明人类起源于非洲，如果这个结论正确，目前世界上说不同语言的民族就具有同一个祖先，从这个角度看，人类语言的演化经历了相同的时间和同样的演化路线，因此我们可以提出一个问题，语言的本质和演化的本质是什么？

从古人类学的研究成果看，相对于人类生理的进化，语言形成和演化的时间要短得多。虽然从目前古人类学的研究成果看语言产生的时间还很难确定，但现代言语科学技术的进步，利用古人类化石经过声道复原，最终合成出语音已经成为可能。因而，怎样从语言学的角度利用现代活的语言来研究语言进化的基本性质则是语言学家面对的重要课题。

郑锦全先生的“词涯八千”（郑锦全，1999,2006）很巧妙地证明了人类掌握一种语言的基本能力。如果排除掉五千年文字的影响，现在比较封闭的社会中，语音的音位通常有几十个，基本语素在八百到到一千左右，常用词汇在三千左右，基本句法结构在二百左右。然而人类的发音器官能发出一两千个用于语言的音素，但人类只选了几十个能区别语义的音位，这说明了人类大脑目前处理语言音位的能力和水平。因此怎样从音位数量、音位结构、音位负担量研究语言的本质是本文的出发点。

过去对音位负担量的研究主要是通过大文本的计算，由于世界上大多数语言没有文字，因此，研究只能在个别有文字的语言中进行，这无疑很大地限制了这一领域的发展。根据多年的研究，我们发展了一种计算词汇层面音位负担量的方法，从而可以对每一种语言进行信息量和结构的计算。本文以藏缅语为例，介绍基于小词汇量的音位负担量、信息量和结构的基本计算方法，同时从音位负担量角度讨论的藏缅语语音结构、音位负担量及其演变的本质。

2. 音位负担量研究

在语音息量的研究方面，该研究可以追溯到早期的布拉格学派时期，当时主要注重于音位学的二元对立。50 年代的研究主要有霍凯特(Hockett,1955,1951) 的研究和格林博格(Greenberg,1959)的研究。霍凯特认为：功能负担的重要性在于它对描写音韵系统有重要的价值，从而使我们可以有一个尺度来认识语言信息、语言冗余度和言语识别。格林博格认为：功能负担以通用的方式反映了一组音位或一组对立特征各成员之间的对有区别意义信号的贡献。60 年代的研究，论文主要介绍了赫厄希斯瓦尔德（Hoenigswald，1960）关于功能负担和音变的研究，他认为：功能负担和语言的音变有关，并提出了一个假说，即，在一种语言里，如果一种对立用的很少，它的消失对系统造成的危害要小于功能负担大的对立。京•罗伯特(King R. D.,1965)将音变和功能负担一同进行研究，并着重研究了音位功能和语音音变的关系，发现在日耳曼语中，功能负担和历史音变的关系不大。

在 60 年代，王士元教授有两篇重要的论文（王士元，1960，1967）。第一篇针对美国英语辅音出现频率的统计差异进行了研究，研究结果表明美国英语辅音的频率受文献风格、方言差异和样本数量的影响不大。其差异主要是来自于不同的统计词表（电子字典）和版本。第二排篇关于“音位功能信息量”研究的经典文章，论文首先讨论了音位功能负担的概念，在前人研究的基础上，王士元先生首次实现了功能负担的计算和指出了计量功能负担的困难，并给出了解决这些困难的方法。首先他讨论了音位系统中常见的三种分布、霍凯特与格林博格的测量方法以及这些方法和香农（Shannon，1948,1951）的通信理论及各种语言学概念的关系。其次在这些背景知识的基础上，王士元教授讨论了功能负担计量必须满足的五个条件。最后他系统地发展了四种计量功能负担的方法。另外，王士元先

*本项研究得到了国家社会科学重大项目“中国有声语言及口传文化保护与传承的数字化方法研究和基础理论研究”的支持，批准号：10&ZD125。

生还指出：“音变严重受到其他许多因素的影响，如音位之间语音的相似度和语言的接触等，但正如许多历史语言学家相信的那样，如果功能负担在音变中确实起作用的话，那么用量化的解释至少可以从一个方面阐明音变这一难题”。关于音位系统的分布，王士元教授指出：“对任意一个语音序列，有三种相互关联的分布，即相似性分布、交叉性分布和互补性分布。当语音序列中每一个音位的排列都共有一组相同的环境时，它就处于相似性分布中；当音位的排列共有一些相同的环境，而不是所有的环境时，该序列处于交叉性分布中；当音位排列没有任何共有环境时，该序列处于互补性分布中”。王士元教授的研究为后来功能负担的研究建立了一个理论上的基本框架。实际上现代语音识别技术中常用的双音子和三音子的概念就起源于音位功能负担量的研究。

3. 研究方法

根据以上两节的讨论，我们首先确定在本研究中只用单音节语素，主要是单音节词。计算采用声、韵和调为音位单位。两个单音节词之间信息量设定为一个信息量单位。声、韵和调在区别两个词时各自的负担为三分之一。根据这些定义，下面介绍一下基本语素音位负担量的计算方法。

我们以彝语喜德话为例，选了 6 个单音节词和其相关数据，见表一。表一第一列是序号，第二列是汉译，第三列是国际音标，第四列是声母的国际音标，第五列是韵母的国际音标，第六列是声调的调值。从表一可以看出，声调有 3 个，声母有 5 个，韵母有 3 个。

表一、喜德彝语例词表

序号	汉义	彝（喜德）	声母	韵母	声调
1	天	mu33	m	u	33
2	山	bo33	b	o	33
3	肝	si21	s	i	21
4	胃	hi55	h	i	55
5	汗	ku21	k	u	21
6	士兵	mo55	m	o	55

在计算时取此表中的第一个词，将其声韵母分别和其他所有词对比，如果是最小对立对就得一分，具体方法是：如果是声母对立就给这个词的声母加一分，如果是韵母对立就给韵母加一分，如果是声调对立就给声调加一分，为了方便计算，一分的数值是 6，这样可以避免小数。表二是只 6 个词声韵调的分数，本文将这种对立称为单项对立。从表二可以看出，在单项对立上，声母的得分较高，韵母和声调较低。

表二、单项对立表

声母对立	韵母对立	声调对立
192	90	12
408	30	24
246	30	12
246	18	42
120	24	24
216	18	36

众所周知，在语言的音位系统中冗余度是普遍存在的，在词汇这一层面往往体现为二项对立和三项对立。二项对立本文定义为两个词之间，对立是由声母、韵母和声调中的两项来完成，如两个词是靠声韵来完成对立就分别给声母和韵母各加二分之一分，其数值为 3。如果是由声调两项完成对立，就分别给声母和声调各加二分之一分，其数值为 3。如果是由韵调两项完成对立，就分别给韵母和声调各加二分之一分，其数值为 3。见表三。从表三可以看出，声韵对立的数值最大，声调对立的数值比较小，韵调对立的数值最小。

表三、声韵调二项对立表

声韵(声)	声韵(韵)	声调(声)	声调(调)	韵调(韵)	韵调(调)
1161	1161	171	171	15	15
1077	1077	195	195	33	33
297	297	330	330	54	54
459	459	300	300	42	42
357	357	201	201	66	66
489	489	297	297	45	45

第三种情况是三项对立，即两个词之间声韵调都不同，对比词的声韵调各加三分之一分，其数值是 2。因此在三项对立中，声韵调的得分完全相同，见表四。从表四可以看出，6 个词条声母、韵母和声调的得分数值虽然有差别，但每一词条的数值是相同的。

表四、声韵调三项对立表

声韵调（声）	声韵调（韵）	声韵调（调）
594	594	594
562	562	562
1044	1044	1044
948	948	948
1118	1118	1118
950	950	950

从表三可以看出，而相对立的情况比较复杂，二单项对立和三相对立的情况比较简单，为了更加清楚的查看结果，我们而相对立的六中情况合为声韵调三种情况，见表五。从表五可以看出，表中只有声、韵和调三列数值。从具体的数值来看，声母的得分数值最大，韵母次之，二声调的得分数值最小。因此我们可以得知，在两项对立中，声母的音位负担是最大了，韵母的音位负担次之，而声调的音位负担最小。

表五、声韵调二项对立综合表

二声母总值	二韵母总值	二声调总值
1332	1176	186
1272	1110	228
627	351	384
759	501	342
558	423	267
786	534	342

将以上的各种情况简单总结一下可以看出，以声、韵和调为音位单位的对立在单音节词层面，总共有 7 中情况，分别是：1）声母对立；2）韵母对立；3）声调对立；4）声韵对立；5）声调对立；6）韵调对立；7）声韵调对立。从彝语者六个例词的数据来看，三项对立的数值最大，其次是两项对立，最小是单项对立。从本文使用的藏缅语所有数据来看，情况也是如此。从这一分析的结果可以看出，音位学中的对立原则在汉藏语以单音节和声韵调为基本语音和音位单位的语言中，实际上是以三相对立和而相对立为主，最小对立对气的作用很小，并不是音位对立的主体。

表六、声韵调音位负担量总数值表

声母总值	韵母总值	声调总值
2118	1860	792
2242	1702	814
1917	1425	1440
1953	1467	1332
1796	1565	1409
1952	1502	1328

为了方便分析，表六给出了声母、韵母和声调音位负担量总数值，声母为 11978、韵母为 9521 和声调为 7115。从这些数值可以看出彝语声韵调的音位负担量是由差别的，而且声母最大，声调最小。这一结果使得我们可以来研究整个藏缅语声韵调的音位负担量，因而为定量评价藏缅语声韵调各自的功能和类型提供证据。也开辟了定量分析不同语言音位系统和语言演化程度的方法。

4. 藏缅语的音位负担量

根据以上对音位负担量的定义和计算方法，我们计算了部分藏缅语的音位负担量，藏缅语的数据库是根据《藏缅语族语言词汇》（黄布凡，1992）建立，在本项研究中只用了其中的单音节词，由于每种语言的单音节数量不同，本文的计算结果除了实际的数值外，还计算了声韵调比值，这样就可以对所有语言进行对比研究，见表七。表中第一列是序号，序号是根据声调音位负担量的大小排序而成，数值小的序号小，数值大的序号大；第二列是语言或某语言的方言名称；第三列是声母音位负担量总值；第四列是韵母音位负担量总值；第五列是声调音位负担量总值；第六列是声韵调音位负担量总值；第七列是声母音位负担量比值；第八列是韵母音位负担量总值；第九列式声调音位负担量总值。其中第三列至第六列的数据要处理词汇的总数才可以使用在比较研究方面，不能简单地单独使用。每种语言具体声韵调音位负担量的研究结果将另文要论文。

表七、藏缅语声韵调音位负担量及声韵调比值表

序号	方言点	声母总值	韵母总值	声调总值	声韵调总值	声母比例	韵母比例	声调比例
1	嘉戎	122112	112248	0	234360	0.52	0.48	0.00
2	藏（夏河）	2157414	2043690	0	4201104	0.51	0.49	0.00
3	藏（书面语）	2576946	2491050	0	5067996	0.51	0.49	0.00
4	羌	1686158	1668194	3020	3357372	0.50	0.50	0.00
5	藏（阿力克）	1059990	1006410	2376	2068776	0.51	0.49	0.00
6	博嘎尔珞巴	848218	904144	2278	1754640	0.48	0.52	0.00
7	墨脱门巴	2565864	2576694	1024866	6167424	0.42	0.42	0.17
8	独龙	1150068	1160004	462984	2773056	0.41	0.42	0.17
9	义都珞巴	101120	89210	42794	233124	0.43	0.38	0.18
10	却域	2460136	2368204	1208860	6037200	0.41	0.39	0.20
11	阿依怒	87334	84742	45316	217392	0.40	0.39	0.21
12	彝（南华）	1630160	1564388	847544	4042092	0.40	0.39	0.21
13	彝（喜德）	1618072	1371046	827170	3816288	0.42	0.36	0.22
14	克伦	474610	465448	263122	1203180	0.39	0.39	0.22
15	贵琼	371858	352190	202964	927012	0.40	0.38	0.22
16	仙岛	1994042	2039090	1133312	5166444	0.39	0.39	0.22
17	阿昌	2517026	2583032	1440146	6540204	0.38	0.39	0.22
18	格曼僜	435402	440184	250686	1126272	0.39	0.39	0.22
19	浪速	2297168	2384534	1353014	6034716	0.38	0.40	0.22
20	波拉	2088520	2173318	1231678	5493516	0.38	0.40	0.22
21	错那门巴_	665470	676474	396448	1738392	0.38	0.39	0.23
22	载瓦	2242938	2331534	1353168	5927640	0.38	0.39	0.23
23	达让僜	198802	193300	116794	508896	0.39	0.38	0.23
24	哈尼（绿春）	1058744	998612	617960	2675316	0.40	0.37	0.23
25	景颇	1000370	1043324	618854	2662548	0.38	0.39	0.23
26	藏（巴塘）	1490360	1434542	895682	3820584	0.39	0.38	0.23
27	彝（巍山）	1713606	1579020	1009098	4301724	0.40	0.37	0.23
28	史兴	495886	466402	296812	1259100	0.39	0.37	0.24
29	勒期	1677070	1772518	1072228	4521816	0.37	0.39	0.24
30	哈尼（墨江）	1021232	975956	622436	2619624	0.39	0.37	0.24
31	傈僳	1479748	1387264	951208	3818220	0.39	0.36	0.25
32	白	1185304	1231198	833434	3249936	0.36	0.38	0.26
33	纳木兹	4148886	4108596	2855610	11113092	0.37	0.37	0.26
34	缅（书面语）	2345622	2272662	1599816	6218100	0.38	0.37	0.26
35	藏（拉萨）	1428336	1421790	1017570	3867696	0.37	0.37	0.26
36	彝（武定）	2809544	2553182	1921742	7284468	0.39	0.35	0.26
37	土家	664624	620026	470794	1755444	0.38	0.35	0.27
38	扎坝	4729342	4709830	3527248	12966420	0.36	0.36	0.27
39	吕苏	2481422	2414234	1885064	6780720	0.37	0.36	0.28
40	彝（撒尼）	1695308	1447904	1219292	4362504	0.39	0.33	0.28
41	基诺	767482	694054	571300	2032836	0.38	0.34	0.28

42	缅（仰光）	4555552	4465258	3541774	12562584	0.36	0.36	0.28
43	普米（兰坪）	6272854	6186160	4963510	17422524	0.36	0.36	0.28
44	纳西	6410460	6170514	5016138	17597112	0.36	0.35	0.29
45	怒苏怒	6144024	5974662	4955754	17074440	0.36	0.35	0.29
46	普米（九龙）	6123764	6051278	5104298	17279340	0.35	0.35	0.30
47	拉祜	1420182	1253196	1202202	3875580	0.37	0.32	0.31
48	嘎卓	886158	804126	765120	2455404	0.36	0.33	0.31
49	木雅	3599292	3606120	4308216	11513628	0.31	0.31	0.37

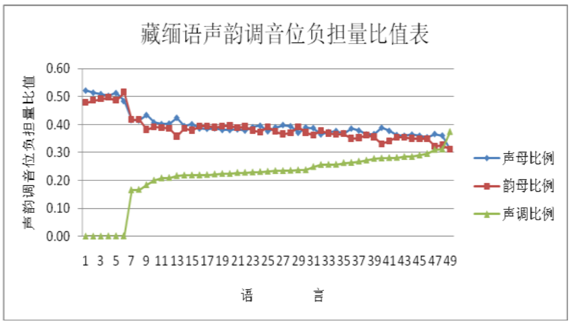
数据表是按照声调的音位负担量进行的排序，从数据可以看出，前六个语言或方言是：1）嘉戎；2）藏夏河话；3）藏书面语；4）羌；5）藏阿力克话和 6）博嘎尔珞巴。这六个语言或方言没有声调，因此声调的音位负担量为零，排在表的最前面。这六种语言或方言声母和韵母的音位负担量比较大，其和等于 1。声调音位负担量最大的六个语言是：1）纳西；2）怒苏怒；3）普米九龙话；4）拉祜；5）嘎卓和 6）木雅，其比值分别是 0.29、0.29、0.30、0.31、0.31 和 0.37。从这些基本数据可以看出，藏缅语声韵调的音位负担量是不同的，他们反映了各自语言的音位系统的差别和发展的不同程度，这为我们研究藏缅语音位体系、音位机构和音位负担量和语言信息之间的关系奠定了基础。本项爱那个研究的计算结果也可以按声母或韵来排序，从而进一步研究声母和韵母的特性。

5. 研究结果

根据本文对词汇层面音位负担量的定义和对藏缅语 49 个语言和方言的计算，可以得到许多很结果和结论，限于篇幅本文在此先从宏观的角度就一些主要的结果进行讨论。有关声母、韵母和声调音位负担量的内部结构、分布以及和语言系数的关系将放在另外的文章中讨论。

首先，从图一可以看出：1）声调从无到有，体现为声母和韵母音位负担量的下降。2）在声调的音位负担量为 0 时，声母和韵母的负担量比有声调的声韵母负担量要高许多，两者之间有一个跳跃。3）在有声调的语言和方言中，声调的信息量和声韵母的信息量成反比关系，也就说一种语言里，如果声调承载的信息量大，声母和韵母能承载的信息量就是小。4）大多数语言声韵母的负担量在一个数量级上，相对来说比较大，而声调的音位负担量要比声韵母小很多。从这 49 个语言和方言上来看，声调音位负担量在有些语言中已经很接近声母和韵母的数值，者体现了藏缅语声调音位负担量发展成都的不同。见图一。

图一、藏缅语声韵调音位负担量比值表

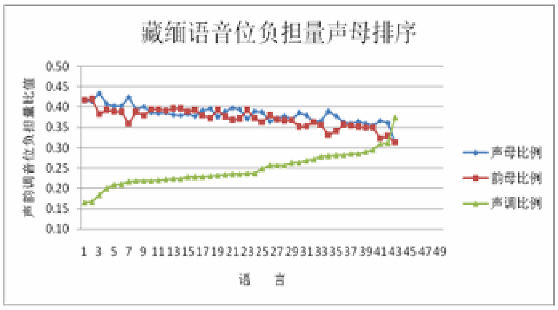


很显然，在藏缅语中，声调的音位负担量是很不同的，这反映了藏缅语声调在不同语言中发展和演化的不同阶段。最小的只有 0.17，而最大的为 0.37，这个数值大过该语言的声母和韵母。

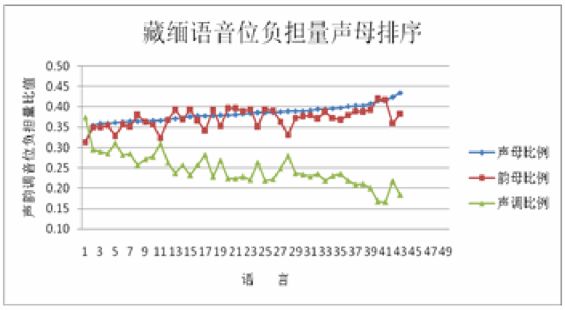
第二，宏观上声母和韵母的信息量相对于声调来说基本相等，也就是说，一种语言里，声母能承载的信息量和韵母基本相同，但和声调有很大的差别。因此从音位结构、功能和层次上可以看出，在藏缅语的的发展过程中，声母和韵母在结构和功能上是一个层次，随着声调的发展和演化，声调的功能逐渐增强，而声母和韵母的功能相对减弱。

根据本文的计算方法，如果按声调排序可以发现声韵母的音位负担量成镜像分布，见图二；如果按声母排序可以发现声调和韵母的音位负担量成镜像分布，见图三；如果按韵母排序可以发现声调和声母的音位负担量呈镜像分布。但三者在数值上有较大差别。

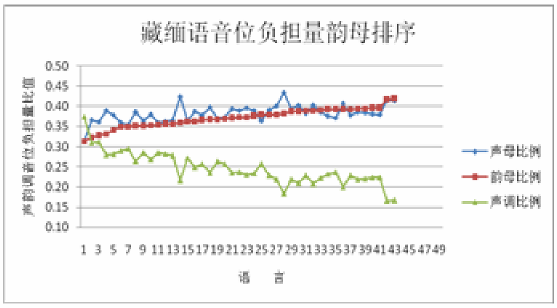
图二、藏缅语音位负担量声调排序图



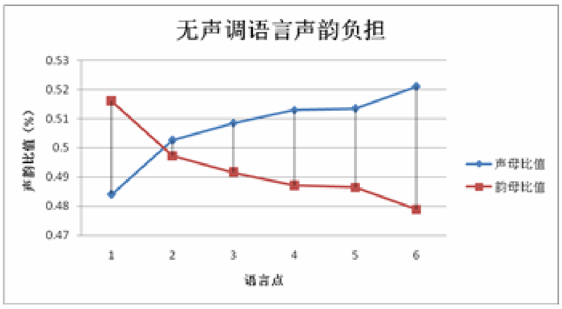
图三、藏缅语音位负担量声母排序图



图四、藏缅语音位负担量韵母排序图



图五、无声调语言声韵母音位负担量表



第三，在无声调的语言里，声母和韵母在信息量的常在上成反比关系，也就说声母承载的信息量大，韵母承载的信息量就小，否则相反，见图二和表七。这些语言呈现出镜像的分布，但总的来说，声母和韵母的数值差别不太大，基本在 0.5 左右。这一点可能和我们采用了声母和韵母做基本单位有关。可以看出前面 6 种语言是无声调的语言，其声调的音位负担量为 0，所以音位负担量有声母和韵母承载，声韵母的负担量比有声调的声韵母要高许多，因为其总和等于一。

第四，对于声母和韵母来说，音位数量的增减不太影响声韵母总的音位负担量。以藏语为例，从古藏语到现在大多数方言，复辅音声母大量脱落，但声母总的音位负担量并没有减少，而是由剩余的声母承载，直到声调产生，整体音位负担量下降。这无疑从一个语言内部的音位负担量为研究语言的演化提供了线索。

6. 讨论

本文的研究表明，语言音位负担的语言学研究能解释许多语言信息量的结构、承载和演变的性质和规律。从 49 个藏缅语语言和方言的情况看，有声调语言和无声调语言之间存在有一个跳跃，声韵调音位负担量的曲线不是平滑的，这存在两种可能：1）我们在调查语言的过程中，声调产生的过程没有调查出来。如，藏语有许多方言正处于声调的产生过程中，而我们目前采用的结构主义音位学的调查方法不能很好地或者说精确地描写声调产生的过程；2）声调的产生确实是突变产生的。另外，在计算方法上，由于不同音位对立类型的数量级差别较大，也有可能掩盖了一些内部的规律，因此在计算上做进一步的加权算法研究是很必要的。最后希望这项研究能成为研究语言学研究和信息量研究的一个很好的切合点，形成科学地研究语言演化的一个新的方向。

7. 参考文献

[1] 黄布凡主编，1992，《藏缅语族语言词汇》，中央民族学院出版社，北京。

[2] 郑锦全，2006，从词语八千到学海无涯，国立中山大学中文系，高雄，讲座。

[3] Greenberg, H.H.，1959，A method of measuring functional yield as applied to tone in Mrican languages. Georgetown University Monograph Series on Languages and Linguistics 12: 7-16.

[4] Hockett, C.F., 1955, A Manual of Phonology, p.218, Baltimore Waverly Press.

[5] Hockett, C.F., 1961, The quantification of functional load. Rand Report, p.2338, Santa Monica.

[6] Hoenigswald.H.M., 1960, Language change and linguistic reconstruction, pp.79-80, University of Chicago Press, Chicago.

[7] King, R.D., 1955, Functional load: its measure and its role in sound change, University of Wisconsin PhD dissertation. A version of this will appear in 'Language'.

[8] Shannon, C.E., 1951, Prediction and entropy of printed English. Bell System Technical Journal, 30: 50--64.

[9] Shannon,C.E. and Weaver, W., 1949, The mathematical theory of communication, University of Illinois Press, Urbana 1949.

[10] Wang, W. S-Y., 1967, Phonemic Theory A (with Application to Midwestern English), Doctoral dissertation. The University of Michigan, -Stress in English. Language

[11] Wang, W. S-Y. and Crawford, J., 1960, Frequency studies of English consonants, Language and Speech 3: 131-139.

[12] Cheng, C. C., 1999, 词涯八千 (Active vocabulary upper limit 8000), Graduate Institute of Linguistics, National Taiwan University, Taipei, May 10. Invited lecture.

蒲西霍尔语软颚化的语音对立[*]

林幼菁[1]，孙天心[2]，陈正贤[3]

（1 北京大学　2 中央研究院　3 台湾大学　中央研究院）

摘要：

软颚化的发音原理是将舌根朝软颚一带向上隆起，做出发元音*ɯ*的动作。这个语音现象在文献上一向被划分为辅音特征，而且在世界语言中很少具对立作用。在这样的背景下，蒲西霍尔语软颚化现象可谓独树一帜，不仅具备区别语意对立的功能，而且确定不是辅音特征。本文从声学语音学、音韵学原理、以及历史演化的角度，讨论分析蒲西霍尔语软颚化的语音对立。

1. 引言

软颚化(velarization)的发音原理是将舌根朝软颚一带向上隆起，做出发元音*ɯ*的动作。最常见的例子当属英语的软颚化边音[ɫ]（俗称「dark *l*」）。这个语音现象在文献上一向被划分为辅音特征，而且在世界语言中很少具对立作用。本文第二作者长期在四川境内调查藏缅语的过程中，在嘉戎语组语言（包括嘉戎语修梧方言、拉坞戎语小依里方言、霍尔语上寨方言）以及北部羌语中都发现有大量的软颚化现象，這些語言的软颚化现象独树一帜，不仅具备区别詞義的功能，而且确定不是辅音特征。本文以四川省阿坝藏族羌族自治州壤塘县蒲西乡霍尔语上寨方言为例(以下简称「蒲西话」)，从声学语音学的角度出发，确认本语声学表现符合文献所描述之软颚化语音特征（第二节）；并以语音观察及音韵学原理为基础，确定蒲西话软颚化并非辅音特征（第三节）；最后，从历史比较的角度，探讨蒲西话软颚化語音可能的来源（第四节）。

2. 蒲西话软颚化的声学测量

2.1. 软颚化的语音及声学特征

关于软颚化的语音特征，相关的研究大多聚焦在英语的软颚化边音「dark *l*」([ɫ])；若要举出具跨语言观察、又明确提出声学对应特征的指标性研究，则首推 Brosnahan and Malmberg 1970、Fant 1975、和 Ladefoged and Maddieson 1996。针对相关的发音机制，Brosnahan and Malmberg (1970: 67)指出软颚化的发音动作主要是舌位向后、在软颚与咽喉后壁一带形成压缩，至于舌位高低并没有直接关联。Fant (1975)则主张软颚化会使舌位降低、向后，并且在软颚区域形成压缩。Ladefoged and Maddieson (1996: 361)在检视了世界数种语言的软颚化语音之后，则发现舌位高低、软颚一带是否成阻都不是关键，与软颚化密切相关、且最稳定的生理特征是舌位向后。

由于对软颚化的定义与发音机制的判断不尽相同，三组研究所提出的对应声学讯号也呈现异同皆有的情况。其中，三组研究皆有共识的，是第二共振峰的变化。舌位向后反映在声学上是第二共振峰下降，由于三组研究皆主张舌位向后是软颚化发音的主要动作，因此他们都提出第二共振峰下降这个相应声学特征。至于第一共振峰的对应则只有 Fant 1975 提出，他主张软颚化的发音会使舌位下降，所以第一共振峰会升高。第三共振峰的对应特征 Brosnahan and Malmberg 1970 和 Fant 1975 都在讨论中提及，但两者的出入较大。Fant 1975 以电路原理及摄动原理为本，主张舌根与软颚之间若成阻，第三共振峰会上升。但是 Brosnahan and Malmberg 1970 却相反地指出第三共振峰会微微下降。后者之所以提出如此迴异的声学结果，有一个可能是他们对软颚化的定义与 Fant 1975 本来就有出入。Fant 1975 在下定义时，明确地指定软颚化

[*] 本文材料得以顺利收集，必须感谢蒲西霍尔语的发音老师钟诚女士以及蒲西乡的藏族朋友们在录音的过程中热忱协助与配合。此研究的初步成果已于 2010 年四川境内藏缅语研讨会宣读，在此向提供意见与反馈的与会者致上谢意。中央研究院语言学研究所的萧素英助研究员与余文生副研究员亦在本文撰写期间提供有益建言，在此一并致谢。

成阻位置是软颚与舌根，而这样的发音动作反映在声学上是第三共振峰的升高。反观 Brosnahan and Malmberg 1970 所描述的软颚化发音方式，压缩的区域则不限于软颚，而是延伸到咽喉的后壁。但事实上，如果在咽喉后壁一带成阻，发出的语音其实不是软颚化，应该是咽喉化(pharyngealized)(Ladefoged and Maddieson 1996: 306-310; Laver 1994: 326-327)，而第三共振峰的下降正好是咽喉化的一个声学对应特征(Ladefoged and Maddieson 1996: 307-310)。

2.2. 研究方法

2.2.1. 语料的收集

我们请来五位发音合作人进行录音，这五位发音人分别是三女二男，年龄在 44 至 59 岁之间。他们都是蒲西乡本地藏族，以蒲西话为母语。

我们设计的朗读语料内含 21 组呈现软颚化有无的最小对立组[3]，选用的字都是单音节、无韵尾的，如此可避免元音发音不到位(target undershoot)的现象(元音共振峰的变化若受韵尾辅音影响，取点测量较困难)。在录音的时候，每位发音合作人将所有的语料朗读六遍。录音系于壤塘县蒲西乡进行，使用的录音设备是 SONY ECM959A 单指向麦克风以及 SONY MZ-R50 MD 录音机。

2.2.2.声学测量

我们将录音的取样频率设定在 22050 Hz，依据语音分析软件 Praat 所计算出的声谱图，首先人工切划出元音的测量范围，最左边的点是声母辅音到元音的过渡结束的位置，最右边则是元音结束的位置。接下来，就由 Praat 自动量测选取范围中点的 F1、F2、F3 这三个共振峰的值。所得出的数据又进一步进行人工检验，修正错误率大约是 29%。

2.2.3. 测量结果与分析

我们将得到的数据以相依样本单因子变异数分析 (repeated measure one-way ANOVA)，分别针对三个共振峰作统计分析，将软颚化视为「受试者内设计」之因子，以排除个别差异造成的误差。我们前后作了两次统计分析。第一次我们将所有发音人的数据一起进行计算，得到的结果如下：

共振峰	第一共振峰	第二共振峰	第三共振峰
F 值	5.98	48.09	7.78
p 值	0.07 > 0.01	0.002 < 0.01	0.049 > 0.01

表 1. 蒲西话软颚化共振峰变化分析(第一次计算：所有发音人数据)

此分析结果显示，若将所有发音人的数据皆纳入分析，蒲西话的软颚化现象符合 Ladefoged and Maddieson 1996 跨语言的观察，即第一与第三共振峰的变化不显著，这表示舌位高低与软颚区域是否成阻并不是蒲西话软颚化的关键生理语音特征；而第二共振峰的显著下降则显示整体而言，舌位向后是蒲西话软颚化语音的主要发音动作。

不过，早在录音过程及以专业听感分析音档语料的阶段，我们就已发现第四位发音人(简称 S4)在朗读软颚化例字时，舌根与软颚成阻的音质几不可闻；虽然能清楚区分对立字的语义差异，但与其他四位发音人对比之下，其软颚化音质有明显的不同，而这样的不同也反映在声学测量的原始数据上。基于这些考量，我们将 S4 的数据排除，进行第二次统计分析。所得出来的结果如下：

共振峰	第一共振峰	第二共振峰	第三共振峰
F 值	6.1	54.9	40.2
p 值	0.089 > 0.01	0.005 < 0.01	0.007 < 0.01

表 2. 蒲西话软颚化共振峰变化分析(第二次计算：排除发音人 S4 数据)

此分析结果显示，发音人 S4 的数据排除后，蒲西话的软颚化语音呈现第二共振峰显著下降、第三共振峰显著上升的情形。以 Fant 1975 提出的语音声学对应为本，我们可以说除了 S4 之外，其他四位发音人在发软颚化语音时，会将舌位向后、并

[3] 此 21 组最小对立组分别呈现三组软颚化与非软颚化元音的对立：/ ə-ə^{ɣ}/、/ ʌ-ʌ^{ɣ}/、/ o-o^{ɣ}/，每组对立有七组例字。最后一组元音对立/ u-u^{ɣ}/因例字太少(仅有一对)，不利统计分析，因此不纳入测量与计算。

在软颚区域形成明显的压缩。除了第一共振峰无显著一致变化以外，其他四位发音人的语音表现算是相当符合 Fant 1975 以及相关文献对软颚化语音提出的经典定义。此结果也表示，纵然舌位向后是蒲西话软颚化最主要的发音动作，但这个语言的使用人口中，有相当比例的人在发软颚化语音时，除了让舌位向后，还会同时在软颚区域压缩成阻。

3. 蒲西话的软颚化属于辅音、元音、还是超音段特征?

根据 Ladefoged and Maddieson 1996 的跨语言研究，软颚化只在极少数的语言中起对立作用(1996: 361)，比如俄语中软颚化与硬颚化(palatalized)的边音是有对立作用的(Ladefoged and Maddieson 1996: 361)； Ikiribati 语的 Tarawa 方言则区分软颚化与一般(非软颚化)鼻音(Laver 1994: 326)；此外 Marshallese 语会在鼻音、流音、和双唇塞音呈现软颚化与非软颚化的对立(Ladefoged and Maddieson 1996: 362-363; Laver 1994: 326)。其他有软颚化语音但不起对立作用的案例还包括了英语和 Catalan 语(Recasens 1991; Recasens et al. 1995; 1996)的软颚化边音[ɫ]、以及泰语曼谷方音中软颚化辅音的副言辞(paralinguistic)用法 (Laver 1994: 326)。这些研究全都在辅音上观察到软颚化的语音，因此截至目前语言学界都将软颚化定位为一种辅音的征性(Brosnahan and Malmberg 1970: 67; Ladefoged and Maddieson 1996: 328, 354, 360-361; Ladefoged 1993: 230-231; Laver 1994, 325-326; Ball and Rahilly 1999: 125-127)。

不过，根据我们对蒲西话软颚化语音的观察，我们认为这个语言的软颚化并不适合被划分为辅音征性。对此我们提出两个理由来论证。第一个理由是音韵分析的学理原则。蒲西话有四十七个辅音音位 (J. Sun 2000: 214-215)，而且所有的辅音都可以出现在软颚化的环境中。如果将软颚化处理成辅音特征，等于让蒲西话为数已相当庞大的辅音数量暴增为两倍(47 x 2 = 94)，迄今世界上尚未有任何语言拥有数量如此惊人的辅音系统。换句话说，将蒲西话的软颚化分析为辅音特征不符合音韵学的经济原则 (principle of economy)。

另外一个理由则是来自声学的观察。在这份研究中，我们标记量测的都是元音的部分，而且排除了辅音声母到元音之间的过渡，测量结果清楚呈现出软颚化的声学特征。再者，如果我们将 Marshallese 语的软颚化辅音声谱图(图 1)与蒲西话软颚化现象的声谱图(图 2)两相比较的话，就会发现蒲西话的软颚化现象与典型的辅音软颚化有明显的差别。在图 1 Marshallese 语的例子(*elał lałe* ‘he’s a down-to-earth person’)中，我们可以在 *elał*的对应声谱图里，看到 *e* 这个元音在一般(非软颚化)边音前，第二共振峰大概在 2000 Hz 左右，而在 *lałe* 这个字里，最后的元音 *e* 紧邻软颚化边音 *ł*，在边音结束来到 *e* 时，第二共振峰从软颚化边音较低的第二共振峰带朝元音 *e* 的第二共振峰带(2000 Hz)陡升，这显示软颚化的语音主要在辅音 *ł*上呈现，紧接的元音则不具软颚化音质。也就是说，如果软颚化是辅音的特征，我们可以预期在紧临的元音上观察到第二共振峰陡升(若元音在辅音后)或陡降(若元音在辅音前)的情形。

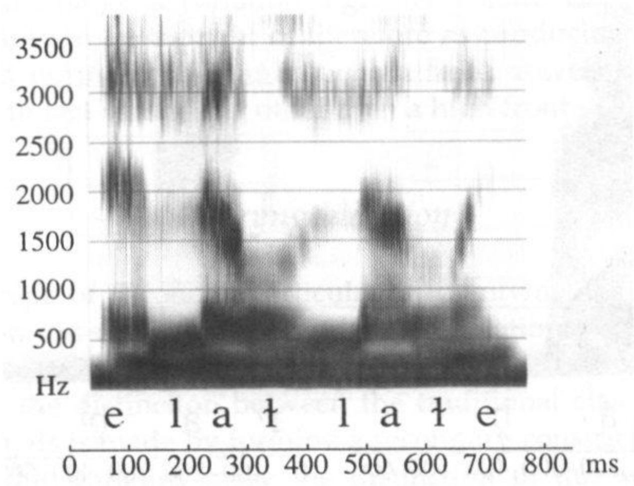


图 1. Marshallese 语 *elał lałe* ‘he’s a down-to-earth person’

(摘自 Ladefoged and Maddieson 1996: Fig. 10.22)

但是我们并没有在蒲西话的软颚化语音观察到类似的状况。一如图 2 所呈现的，在这个蒲西话的接近最小对立组里，两个字(*vlʌ*「派」和 *slʌ*^{ɣ}「翻」)都有一个边音 *l* 加*ʌ*元音的组合。在非软颚化的 *vlʌ*这个例子中，边音的第二共振峰落在 2300 Hz 一带，而*ʌ*元音的第二共振峰比较低，所以从 *l* 过渡到*ʌ*的时候，会看到第二共振峰从 2300 Hz 下降到*ʌ*元音第二共振峰的高度。如果蒲西话的软颚化语音是辅音特征的话，当我们观察 *slʌ*^{ɣ}「翻」这

个例子时，应该会在声母过渡到元音处看到第二共振峰陡升的现象(因为软颚化边音的第二共振峰比非软颚化ʌ的第二共振峰低)。然而语音事实表明，在 *slʌˠ*「翻」这个软颚化的例字里，第二共振峰完全没有从声母过渡到元音的变化。这个例字自边音声母起一直到元音结束，第二共振峰都处于频率相对较低的固定范围，几乎没有任何变动：

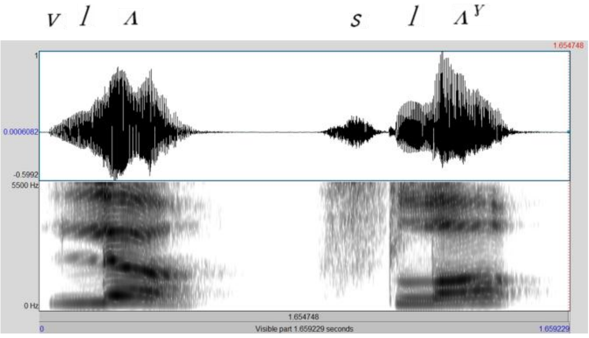


图 2. 蒲西话发音人 S2 所朗读的 *vlʌ*「派」和 *slʌˠ*「翻」的对比

根据以上的音韵原则与声学证据，我们可以基本排除蒲西话的软颚化是辅音征性的可能。不过，上面声谱图所呈现的声学特性也带出另一个问题：蒲西话的软颚化是元音的征性，还是超音段的征性呢？毕竟，早在田野调查审音的阶段，我们就观察到蒲西话的软颚化不单出现在元音上，而是从声母就开始了。换句话说，在我们所收录的例子里，软颚化是贯穿全字的。现阶段我们并不排除蒲西话的软颚化是超音段征性的可能，只是目前暂以元音上加符号[ˠ]来标示。未来我们将搜集更多关键性的语料来厘清这个问题。现在已经确定的，是蒲西话的软颚化并非辅音的特征，这在语音类型学上是非常特殊的现象，也是重要的发现。

4. 蒲西话软颚化语音的起源

蒲西话软颚化语音的起源是霍尔语历史音韵学的重要课题。若要深入探究，必须细致讨论古霍尔语，甚至古嘉戎语组祖语音系，超过本文的研讨范围。以下仅援引现有之比较材料，介绍蒲西话软颚化元音的几种可能的来源。

部分霍尔语软颚化语音应属后起创新。其一，藏语原不存在软颚化语音，而蒲西话中若干藏借词却带有此类音質，例词如 *rʌˠvə*ˆ˙「牛栏」(藏语⟨ra⟩)、*vlʌ́ˠma*「喇嘛」(藏语⟨bla.ma⟩)。

其二，少数蒲西话软颚化語音在亲属语言里与软腭浊擦音*ɣ*对应，有可能源自复辅音中之软颚辅音脱落后的代偿现象，例词如 *rə_ˠ* 「买」(小依里拉坞戎语*ɣdə*ˆ；业隆拉坞戎语*ɣruʔ*)、*ri_ˠ*「马」(道孚霍尔语 *rɣi*)。

然而，蒲西话不少本土词中的软颚化语音可能属于存古特征。最有力的证据便是目前三种存在音位性软颚化语音的现代嘉戎语组语言（蒲西霍尔语、小依里拉坞戎语、修梧嘉戎语）的对应词均有软颚化语音，可能源自古嘉戎语组祖语一般元音与软颚化元音之对立，例词如下：[4]

	蒲西霍尔语	小依里拉坞戎语	修梧嘉戎语
冰	lvôˠ	rpʰəˠm	ta-lvāˠm
（男人之）姐妹	sn ôˠ	---	tə-snāˠm
宽	lo̱ˠ	lə̱ˠm	kə-lāˠm
颈瘤	zvâˠv	zvʌ̱ˠv	tə-zbâˠv
辣	ltsʰʌ̂ˠv	ltsʰa̱ˠv	kə-vartsâˠv
深	nʌˠv	nʌ̂ˠv	kə-nôˠv

5. 参考文献

[1] Ball, Martin J. & Joan Rahilly. 1999. Phonetics: the Science of Speech. London: Arnold.

[2] Brosnahan, Leonard F. & Bertill Malmberg. 1970. Introduction to Phonetics. Cambridge (USA): Cambridge University Press.

[3] Fant, Gunnar. 1975. Vocal-tract area and length perturbations. Quarterly Progress and Status Report 4, 1-14. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

[4] Ladefoged, Peter. 1993. A Course in Phonetics. Orlando: Harcourt Brace and Company.

[5] Ladefoged, Peter & Ian Maddieson. 1996. The Sounds of the World's Languages. Oxford: Blackwell.

[6] Laver, John. 1994. Principles of Phonetics. Cambridge

[4] 以下所引均为本文第二作者亲自调查之材料。

(UK): Cambridge University Press.

[7] Recasens, Daniel. 1991. An electropalatographic and acoustic study of consonant-to-vowels coarticulation. Journal of Phonetics 19: 177-192.

[8] Recasens, Daniel, Jordi Fontdevila & Maria Dolors Pallarès. 1995. Velarization degree and coarticulatory resistance for /l/ in Catalan and German. Jounral of Phonetics 23: 37-52.

[9] Recasens, Daniel, Jordi Fontdevila & Maria Dolors Pallarès. 1996. Linguopalatal coarticulation and alveolar-palatal correlations for velarized and non-velarized /l/. Journal of Phonetics 24: 165-185.

[10] Sun, Jackson T.-S. 2000. Stem alternation in Puxi verb inflection: Toward validating the rGyalrongic Subgroup in Qiangic. Language and Linguistics 1: 211-232.

桂南横县（陶圩）平话阴阳入声分化的声学机制*

关英伟　梁晓丽

提要

本文主要采用谐波分析、基频分析、音节时长分析和共振峰分析等方法，探讨桂南横县陶圩平话入声内部分化的声学机制。研究结果显示：桂南陶圩平话上下类入声音节的元音共振峰在调音音色上没有明显差别，其入声分化机制是由时长因素和不同的喉头机制共同作用的结果。上类入声时长均短于下类入声时长，阴入又短于阳入，入声的平均时长由短到长顺序为：上阴入<上阳入<下阴入<下阳入；入从喉头机制看，上下类入声属于不同的发声方式：上类入声为紧音，下类入声为松音。在发声类型上，紧音属于能使声调升高的紧嗓音，紧音的调值要高于松音，松音为正常嗓音。

关键词：

陶圩平话　入声　发声　紧嗓音　正常嗓音

1. 引言

1.1

横县平话是桂南平话邕江支一个独具特色的代表点之一。调类数量多是桂南平话的特点之一（杨焕典等，1985）。从已发表的材料看，桂南平话调类的复杂性主要体现在入声的分化上。覃远雄（2004）以桂南八处方言的平话材料为依据，将桂南平话的入声归为三种类型：（1）清入一类，浊入两类；（2）清入两类、浊入一类；（3）清入两类，浊入两类。其中横县、宜州平话就是属于第三种类型。闭思明（1998）、黎曙光（2004）、黄海瑶（2008）对横县那阳镇、横州镇、百合镇的调查结果以及我们对横县陶圩镇平话的调查结果与覃远雄的调查结果基本相符：横县四镇的平话都是 10 个声调，平上去入各分阴阳两调。其中阴入和阳入又各分上下两调。

对横县入声分化的原因和条件此前的研究主有三种描写：一是以元音的长短为分化条件，上阴入和上阳入的元音为短元音，下阴入和下阳入的元音为长元音（闭克朝，1985）；二是以古今韵母对比为分化条件，但有交叉（黎曙光，2004）；三是以今读韵母为分化条件，即甲类韵母一个调，乙类韵母另一个调，今读逢甲类韵母归上类，逢乙类韵母归下类（覃远雄，2004）。这三种描写都是从不同的角度对入声分化做的分类。

以往的研究主要是从音系描写和调音层面的分析入手，没有涉及发声层面的研究，发声类型的研究主要集中在喉头的发声方法。从言语产生的声学理论上讲，声调属于声源部分，言语产生模型中的声源对应于生理上声带各种不同的振动频率和振动方式。即体现为（1）声带振动的快慢，在声学上体现为基频的高低；(2）声带的不同振动方式，在声学上体现为频率域的特性不同，前者属于调音，后者属于发声。陶圩平话声调数量众多，入声类型复杂，在调音层面和发声层面都会表现出不同的声学特征。本研究采用声学的方法考察陶圩平话声调的调音和发声两个层面的声学特征，探讨阴阳入声内部的分化机制。

1.2 横县平话概况

横县位于桂东南，地处东经 108°48′～109°37′，北纬 22°08′～23°30′。东邻贵港，西连邕宁，南与灵山县接壤，西与邕宁县为界，北边是宾阳县。境内语言主要有汉语和壮语。汉语方言有平话（当地俗称客话）、新民话和白话；壮语分为横南壮话和横北壮话。横县平话主要分布于百合、马山、附城、那阳、板路、峦城、平朗、马岭、莲塘、云表、陶圩、横州等乡镇。流行在不同区域的平话大致相同，基本能互相通话。但也存在一些差别，其中以语音差别最为明显。其中百合、马山、那阳、附城等 4 个乡镇全部使用平话。在横县以平话为母语的约有 66 万人。平话是横县的第一大汉语方言，是各民族之间交流沟通的主要语言工具。

陶圩镇位居横县西北部，距县城横州 30 公里，距首府南宁 90 公里，与石塘、校椅、莲塘、平马等镇相接壤，总人口 8 万。镇内分布有汉语平话和壮话两种语言，绝大部分人说平话，当地人称之为“陶圩话”，只有少数一些靠近平马、石塘、校椅等壮话集中镇的村落是说壮话。其中说双语的情况

*本研究得到国家社科基金“广西汉语方言的发声类型研究”（No：10XYY003）的资助。

比较多见，即“见客讲客，见壮讲壮”。

2. 横县陶圩平话的声调

2.1 实验说明

主要考察陶圩平话声调基频高低变化的情况，语音学是用五度值来定义声调的，在声学上主要是从基频来量化。为了符合听感，我们将基频转换成五度值来分析。

2.1.1 实验材料

录音软件为 Praat，采样率为 22kHz，单通道，采样精度为 16 位。全部数据使用 Excel 电子表格进行统计和分析。

语音材料为横县陶圩镇平话，从十个声调中选取实验字，其中舒声调每个调类选 8 个例字，每个字读三遍，每个调类共得到 24 个样本，六个舒声调共 144 个样本（6×8×3）。促声调每个调选 10 个例字，每个字读 3 遍，四个促声调共得到 120 个样本（4×10×3），全部样本量为 244 个。

本研究发音人为梁晓丽，1985 年出生，陶圩镇人，母语为陶圩话，硕士研究生毕业。

2.1.2 数据处理

首先对基频进行归一化处理，用“音高提取程序”提取每个声调的时长和每个声调 10 个时刻点的基频数据，计算每个音节在 10 个采样点上的原始基频数据的平均值。 再将每个发音人基频数据的平均值转换成对数，最后转换成五度值，得到相对化和归一化的数据，并作出陶圩平话五度音高曲线图，见图 2.1。

2.2 陶圩平话声调的音高曲线特征

图 2.1 是对陶圩平话的十个调类的声学数据归一并转换成五度值后得到的五度音高曲线图。图中纵轴数字 1-5 分别与五度值对应。1、2 度在低音区，4、5 度在高音区，3 度在中音区，是低音区和高音区的交界点。

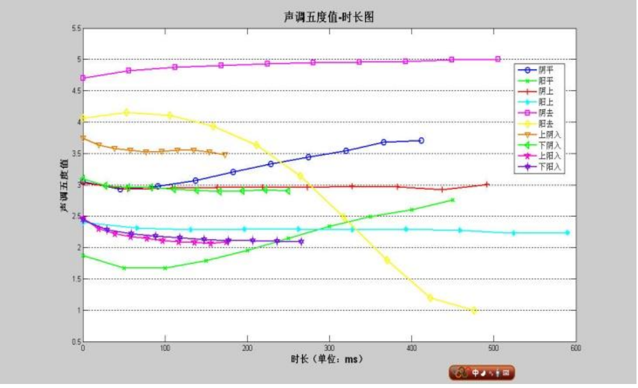


图 2.1　陶圩平话声调五度音高曲线图

从图 2.1 看，阴平调是一个中升调，调值为[34]；阳平调是个低升调，调值为[13]；阴上调是个平调，整条曲线都在 3 度区间内，整条曲线呈现出“平”的特点，调值定为[33]；阳上调是一个低平调，音高曲线位于 2 度区间内，也呈现出“平”的特点，调值定为[22]；阴去调是一个高平调，音高曲线位于 5 度区间内，尽管整条曲线有微微上升的趋势，但总体呈现出“平”的特征，调值为[55]。阳去调是一个高降调，起点在四度区间，终点落在 1 度区间，调值为[41]。阴入和阳入都为平调，其音高曲线表现为“平”的特征。其中上下阴入调值相差一度，分别为[44]和[33]，下阴入与阴上音高曲线合并，调值相同；上下阳入与阳上的音高曲线基本合并，调值同为[22]。

从时长看，上类入声要短于下类入声。

根据上图，我们将陶圩平话的调类和调值列表如下：

表 2.1　陶圩平话调类调值

调类	阴平	阳平	阴上	阳上	阴去	阳去	上阴入	下阴入	上阳入	下阳入
调值	34	13	33	22	55	41	4	33	2	22
调型	升	升	平	平	平	降	平	平	平	平

上下类入声分别为促调，所以我们在调值下划横线，以区别舒声调，同时上类入声从听感上要比下类入声短，所以我们只用一个数字表示，以区别上下两类入声。

2.3 陶圩平话声调的特点

（1)今横县陶圩平话有 10 个声调,古平声字、上声、去声和入声字按古声母的清浊今各分为阴阳两调。其中阴入和阳入又各分为上下两小类。

（2）十个调类中 7 个声调的调型均为平调，只有阴平和阳平是升调，阳去为降调。

（3）入声与舒声相配， 入声的调值与舒声调值相同或相近，而且阴调类配阴调类，阳调类配阳调类。具体表现为：上下阳入的调值与阳上的调值相同，都是[22]，只是促声和舒声的差别；上下阴入的调型与阴上相同，都为平调，也是促声与舒声的差别。其中下阴入与阴上的调值相同，均为[33]。

（4）入声均以[-p、-t、-k]收尾，都是短促调。从听感上可以感觉到上类入声较下类入声更为短促。

（5）上下阳入音高曲线重合，上下阴与阴上基频曲线重合，上下阴入基频曲线相差一度。说明陶圩平话不完全是以声调的对立来区别音节的。

3. 陶圩平话的发声音质

陶圩平话声调的基频曲线分析结果显示，上下类入声都是平调，上下阴入调值相近，分别为 4 和 33；上下阳入基频重合，分别为 2 和 22。这说明陶圩平话入声的基频音高变化没有太大的区别，不完全是以声调的对立来区别音节的，可以假设，入声的发声有其独特的发声音质。

发声音质是指通过不同调声发声产生的不同音色。调声发声在声学上对应于嗓音发声类型的频率域特征。(孔江平 2001 年 285,288 页）我们采用谐波差值分析的方法，进一步分析陶圩平话入声的发声音质，从发声机制上探析陶圩平话入声的性质。

3.1 谐波差值分析方法

谐波差值分析，主要是测量第一谐波 h1 和第二谐波 h2 的振幅，从其差值或比值可以反映嗓音发声类型的性质，是语音学中用来研究发声类型的性质的最常用的方法之一（孔江平 2001,23 页）。

谐波振幅的差值在一般情况下能反映声带振动时的紧张程度。h1-h2 的数值越大，

嗓音高频的能量就越小，即声源谱的能量就衰减的越快，声带表现为松或漏气，数值越小，嗓音高频的能量就越高，即声源谱的能量就衰减的越慢，声带体现为越紧。

本实验提取的参数主要有：第一、二谐波 h1、h2 的振幅（dB）和第一、二共振峰 F1、F2 的能量值（dB），这一类参数主要是测量声带的松紧，反映嗓音的发声性质。

3.2 实验材料

3.2.1 陶圩平话的入声韵

陶圩平话入声韵共 12 个，阴阳入声以今读入声韵母为条件各分化为上下两类。其中有对立的 5 对,还有 6 个韵母是下类特有的。入声韵母见表 3.1。本文的研究基于有对立的上下类入声韵母。

表 3.1 陶圩平话入声韵母

上类		下类	
上阴入	上阳入	下阴入	下阳入
A组:ap at ak ek ok		A组:ap at ak ek ok	
		B组:ep et op ot ip it ut	

3.2.2 实验字表设计

实验字表分别选取上下类入声有对立的 5 对韵母：ap:ap｜at:at｜ak:ak｜ek:ek ｜ok:ok，每个韵母上下类各选两个词，四个声调共计 40 个词（5×2×4）。在词表设计上充分考虑到最小对立对原则。词表中除了“特/白”不是最小对立对，其余都为最小对立对。每个音节读三遍，40 个音节共得到 120 个样本。实验字表见表 3.2。

表 3.2　陶圩平话入声实验字表

韵母	阴入		阳入		备注
	上类	下类	上类	下类	
ap:ap	k急	k夹	ts集	ts杂	
	k级	k甲	ts习	ts闸	
at:at	k骨	k刮	f佛	f罚	
	p笔	p八	t突	t达	
ak:ak	p北	p百	ts贼	ts砸	
	ts则	ts窄	t特	p白	
ek:ek	ts积	ts着$_1$	ts直	ts着$_2$	着$_1$：穿
	k击	k脚	x易	x药	着$_2$：睡着
ok:ok	k谷	k郭	ts昨	ts浊	
	ts竹	ts桌	t读	t铎	

3.3 实验结果分析

谐波振幅值的测量方法是，取每个音节韵母段开始后 30-60 毫秒的一个点，分别测量 h1 和 h2 的振幅值，并将数据进行平均。得到表 3.2 的谐波差值。

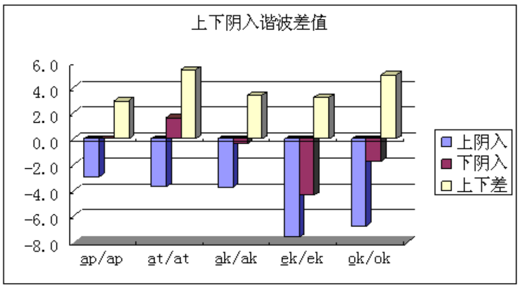


图 3.1　陶圩平话上下阴入谐波差值示意图

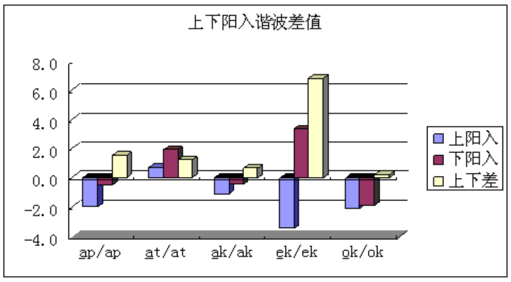


图 3.2　陶圩平话上下阳入谐波差值示意图

从参数表和示意图上看，上类阴入和阳入谐波差的平均值均为负值，即 h1-h2 的数值较小，h2 均大于 h1。下类阴入也表现为负值，但是均小于上类阴入，而且上下两类阴入 h1-h2 的差值较大，为 3.95dB，下类阳入表现为正值，两类谐波差值为 2.4dB。因此可以说上类入声在发音时声带确实比下类入声紧张，体现出一种紧嗓音的发声性质；下类入声第一谐波能量较强，表现为正常嗓音的发声类型。

表 3.3　陶圩平话入声元音谐波差值参数（单位：dB）

韵母	上类			下类	
	声调	例字	h1-h2	例字	h1-h2
ap:ap	阴入	急	-3.4	夹	0.6
		级	-2.5	甲	-0.6
	阳入	集	-2.8	杂	0.6
		习	-1.2	闸	0.9
at:at	阴入	笔	-1.3	八	0.8
		骨	-6.1	刮	2.5
	阳入	突	2.1	达	1.0
		佛	-0.7	罚	2.8
ak:ak	阴入	北	-3.2	百	-1.1
		则	-4.3	窄	0.4
	阳入	贼	-2.8	砸	-0.9
		特	0.6	白	0.2
ek:ek	阴入	击	-6.0	脚	-5.5
		积	-9.2	着	-3.2
	阳入	直	-5.3	着	-1.2
		易	-1.6	药	8.2
ok:ok	阴入	谷	-5.2	郭	-0.3
		竹	-8.3	桌	-3.2
	阳入	昨	-3.1	浊	-2.0
		读	-1.1	铎	-1.4
平均	阴入		-4.95		-1
平均	阳入		-1.6		0.8
松-紧	阴入		3.95		2.4

从能量截面图也可以看出上下类入声的差异。

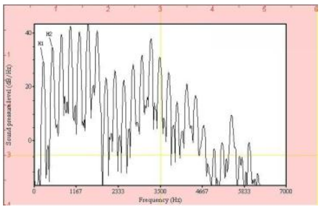


图3.3 上阴入“质”[a]的元音功率谱

（h1- h2=0.7dB）

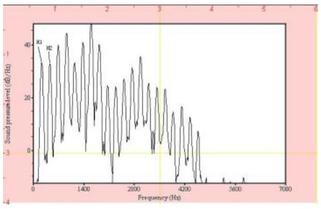


图3.4 下阴入“札”[a] 元音功率谱

（h1- h2= -5.3dB）

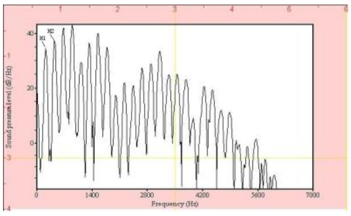


图3.5上阳入“集”[a]的元音功率谱

（h1- h2=-2.8dB）

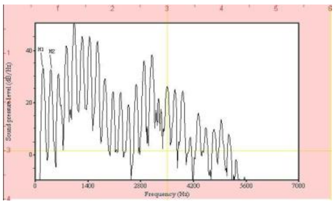


图3.6下阳入“杂”[a]的元音功率谱

（h1- h2=0.6dB）

上图中，上类入声的 h1 均小于 h2，h1-h2 谐波差均为负值，下类入声的 h1 均大于 h2，h1-h2 谐波差均为正值。

我们进一步对上下类入声的 h1-h2 的谐波差作了 T 检验，检验结果见表 3.3：

表 3.4 陶圩平话入声谐波差 T-test

	标准差	df	Sig.(双侧)
上阴入	2.52	18	0.02
下阴入	2.37		
上阳入	2.07	18	0.051
下阳入	2.96		

从检验结果看，上下阴入的标准差分别为 2.52 和 2.37，T 检验 结果显示，sig 值为 0.02，小于<0.05，两者有显著区别。上下阳入的标准差分别为 2.07 和 2.96，这说明下阳入的数据相对分散，T 检验结果显示，sig 值为 0.051，接近 0.05，差异不太显著。

可以肯定，上下阴入存在不同的发声类型。为了确定上下阳入的 5 对元音在音质上到底有多大差别，我们用了第二共振峰 F2 与第一谐波 h1 的振幅值差值以及第二共振峰 F2 与第一共振峰 F1 的振幅值差值来进一步考察。数据见表 3.5。

从表 3.5 数据看，下类阴入 F2-h1 的平均值除了[ek、ok]韵外，均小于上类入声，为 5＜7，下类阴入的 F2-F1 的平均值也小于上类入声，为 -17.5＜-14.6，除了 [at]韵外，其余韵母的 F2-F1 值均小于上阴入。下类阳入的 F2-F1 平均值均同样也小于上类，为-12.7＜-11.3，除了[ek]韵外，其余韵母的 F2-F1 值均小于上类，但是 F2-h1 平均值稍大于上类，为 8.4>7.0，主要是[ap]和[ak] 韵的差值较大导致的，除了[ap] 和[ak]韵之外，其余韵母的 F2- h1 值均小于上类。从上述参数可以看出，发上类入声时，整个声带处于较紧张的状态，嗓音能量衰减的较慢，高频能量较强，使声音听起来较为洪亮，形成紧元音音色，下类入声则相反，在音色上没有紧元音的音色，与其他嗓音发声状态相比，各种量能比基本上处于正常发声状态。因此，可以说横县陶圩平话上类入声的紧元音是由声带紧缩而产生的一种紧元音，在发声类型上属于紧喉音，这种紧嗓音是一种能使声调升高的紧音。从调值上看，上类入声无论阴阳，其调值都比下类入声要高，上阴入和上阳入的调值分别为[4]和[33]，下阴入和下阳入的调值分别为[2]和[22]。由此我们认为，陶圩平话入声的内部分化是由于发声机制的不同造成的。

表 3.5 陶圩平话入声各项振幅值差值参数（单位:dB）

上类				下类		
韵母	调类	AF2-h1	AF2-AF1	调类	AF2-h1	AF2-AF1
ap:ap	阴入	11.2	-10.1	阴入	4.2	-15.4
	阳入	8.8	-11.2	阳入	14.3	-11.6
at:at	阴入	7.1	-20.0	阴入	6.0	-19.3
	阳入	9.6	-11.6	阳入	9.0	-18.3
ak:ak	阴入	8.4	-16.0	阴入	2.3	-22.3
	阳入	8.0	-11.7	阳入	12.3	-17.5
ek:ek	阴入	3.5	-8.3	阴入	5.9	-13.4
	阳入	1.4	-10.7	阳入	-0.3	-5.5
ok:ok	阴入	4.1	-20.7	阴入	7.7	-17.1
	阳入	7.5	-11.7	阳入	6.8	-9.7
平均	阴入	7.0	-14.6	阴入	5.0	-17.5
	阳入	7.0	-11.3	阳入	8.4	-12.7

4. 陶圩平话的音节时长和元音音质

4.1. 音节时长分析

音节时长反映声调的长短，从听感上可以听出陶圩平话上下类入声的时长不同，我们测量陶圩平话上下类有对立的 5 对韵母的 278 个入声音节时长，目的是考察上下类入声音节时长的差值并进行统计学上的分析。数据见表 4.1，并对每个调类的时长数据进行平均化处理。

表 4.1 陶圩平话入声时长 （单位：ms）

韵母	上类		下类	
	调类	时长	调类	时长
ap/ ap	阴入	159.47	阴入	218.67
	阳入	170.85	阳入	258.19
at/ at	阴入	174.26	阴入	274.37
	阳入	142.30	阳入	269.82
ak/ ak	阴入	155.23	阴入	222.72
	阳入	198.63	阳入	250.67
ek/ ek	阴入	191.29	阴入	239.35
	阳入	184.59	阳入	294.40
ok/ ok	阴入	170.76	阴入	318.28
	阳入	193.55	阳入	357.71
平均	阴入	170.20	阴入	254.68
	阳入	177.99	阳入	286.16

从表 4.1 看，上类入声时长均短于下类入声时长，同类入声音节时长阴入又短于阳入。上类入声韵母的时长均小于 200 毫秒，上阴入平均时长为 170.2 毫秒，上阳入的平均时长为 177.99 毫秒。下类入声韵母的时长均大于 200 毫秒，下阴入的平均时长为 254.68 毫秒，下阳入平均时长为 286.16 毫秒。我们将其平均时长由长到短进行排序，顺序为：下阳入>下阴入>上阳入>上阴入。

时长数据的 T 检验结果为：上类入声韵母时长

短于下类元音时长差异显著，Sig. 值均小于 0.000。

4.2 元音的调音音质

语音学是从舌位的高低和唇形的圆展来定义调音音质的，在声学上主要是从共振峰结构上来量化。本部分主要声学上考察陶圩平话元音的调音音质音质，方法是提取第一共振峰 F1 和第二共振峰 F2 的频率，见表 4.2，并以此画出横县陶圩平话的声学元音图，见图 4.1。

表 4.2 陶圩平话元音共振峰参数（单位：Hz）

	韵母	F1	F2
上阴入	ap	917.5	1489.8
	at	978.9	1563.0
	ak	952.0	1465.5
	ek	703.2	2214.4
	ok	712.7	1249.4
上阳入	ap	878.4	1463.1
	at	932.2	1539.0
	ak	957.5	1435.3
	ek	665.0	2122.4
	ok	678.5	1158.7
下阴入	ap	943.4	1541.4
	at	1006.6	1601.1
	ak	946.1	1514.3
	ek	747.6	2117.4
	ok	741.1	1190.0
下阳入	ap	1028.8	1482.9
	at	1099.4	1616.1
	ak	1013.9	1549.3
	ek	806.2	1864.3
	ok	706.6	1141.6

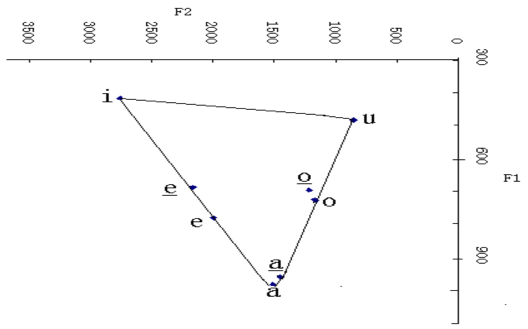


图 4.1 陶圩平话上下类入声元音声学元音图

根据数据和声学元音图可以看出，上下类入声的每对元音在音质上差异不大。总的来说，上类入声的 F1 都要小于下类入声的 F1，F2 差别不大。也就是说，上类入声元音都比下类入声元音要高，即紧元音的开口度都比松元音要小。可以认为陶圩平话中入声的内部分化不完全是调音的差异为区别特征，而是元音的松紧和时长共同作用的结果。紧音相当于紧嗓音，松音相当于正常嗓音。

5. 结论

通过以上研究，对横县陶圩平话上下类入声的发声类型及其声学特征可以得出以下声学上的结论：

1、横县陶圩平话上下类入声的内部分化是由时长因素和不同的喉头机制互为伴随来承担音位功能。也就是说，横县平话入声的分化机制是由时长因素和不同的喉头机制共同作用的结果。

2、从喉头机制看，上下类入声属于不同的发声方式。上类入声为紧音，下类入声为松音，下类入声的 h1- h2 值均大于上类入声。在发声类型上，紧音属于能使声调升高的紧嗓音，紧音的调值要高于松音，松音为正常嗓音。

3、上下类入声的音节时长有显著差异，上类入声的音节时长比下类入声的音节时长短，平均时长顺序为：下阳入>下阴入>上阳入>上阴入，上阴入平均时长为 170 Hz，下阴入平均时长为 255 毫秒，平均时长差为 84.5 毫秒；上阳入平均时长为 178 毫秒，下阳入平均时长为 286 毫秒，平均时长差为 108.2 毫秒。

4、上下类入声音节的元音共振峰在调音音色上没有明显差别。但上下类入声的元音共振峰有明显的规律性，主要表现为，紧元音的 F1 比松元音的 F1 略小，在生理上体现为紧元音的开口度比松元音的要小些，舌位略高些，但在调音音色上没有明显差别。

6. 参考文献

[1] 闭克朝 1985 《桂南平话的入声》，《方言》第 4 期，290 页

[2] 闭思明 1998 《横县那阳平话的语音特点》，《右江民族师专学报》第 3 期。

[3] 黄海瑶 2008 《广西横县百合平话音系》，《桂林师范高等专科学校学报》第 2 期，15-24 页。

[4] 孔江平 2001 《论语言发声》，北京：中央民族大学

出版社。

[5] 黎曙光　2004　《略论横县平话语音特点》,《广西民族学院学报》第 3 期。

[6] 覃远雄　2004　《桂南平话的声调及其演变》,《方言》第 3 期。

[7] 杨焕典等　1985　《广西的汉语方言（稿）》,《方言》第 3 期。

高级阶段韩国学习者的阳平加工方式

吴韩娜

1. 引言

随着学习汉语的韩国学生数量的增多, 近年来针对韩国留学生汉语语音习得的研究成果占了很大的比例。韩国学生学习汉语和欧美学生比起来, 困难要少得多。但韩语是非声调语言, 即音节层里的不同音高在韩语中没有特殊意义。因此, 从来没接触过“声调”这个概念的韩国学生, 在掌握声调的过程中会遇到很多问题。他们反映学习中最难的是“声调”, 这在学界具有一定共识。因此, 在对外汉语教学界, 这方面的研究一直没有停止过。

根据前人的研究结果, 韩国学生的四声偏误中占的比例最大的是阳平和上声 (高玉娟、李宝贵, 2006; 马燕华, 1994; 孟柱亿, 1992; 宋春阳, 1998; 朱川, 1997 等) 。可见阳平和上声是韩国学生的难点。

其中本文主要探讨的是“阳平”问题。归纳前人的具体研究, 零起点韩国学生阳平中调型的偏误多于调域的偏误; 最主要调型错误表现为在升调之前加上短暂降调, 出现降升调的发音曲线(冯丽萍、胡秀梅, 2005) ; 而调域的偏误则主要表现为上升的幅度不够高而终点偏低,其跨度空间不明显,起点过低等,在很多情况下念成位于下半域的微升调,因此与上声的声调曲线极其接近(冯丽萍、胡秀梅, 2005; 朱川, 1997)。其他研究虽然大多数没有明确地说明被试的详情(如, 汉语水平, 语言背景等), 但总体地说, 韩国学生的阳平偏误中最为明显的特征是升的趋势不很明显的现象, 即把阳平念成低平调或有微升的低调。

中介语不断发展, 由低级到高级, 逐渐离开第一语言向目的语靠拢(刘珣, 2002; 鲁健骥, 1993)。但从前人的研究, 我们只能了解到初级阶段韩国学生的情况, 而至于高级阶段学生的情况以及以后的发展趋势却很不明确。这很可能和语音教学集中在初级阶段的现状有关系。

针对这些情况, 本文根据中介语特征和阳平发音的实际情况, 以声学分析的方法考察高级阶段韩国学习者的阳平习得情况并通过汉语母语者的听辨实验, 学习者对阳平的加工方式以及认知策略进行讨论。从而试考虑所定为“偏误”的中介语特征是否非得要去纠正的错误。

2. 研究方法

2.1 设备

设备包括话筒(Aiwa pin mic.)、笔记本计算机以及语音分析软件(Pratt 和 Matlab) 和计算软件(Excel)。

2.2 被试

(1) 对照群: 说普通话的北京大学中文系研究生 2 名(1 男 1 女)。

(2) 实验群: 本实验选择母语背景为首尔标准语的汉语水平高级阶段的韩国学生, 被试男 6 名女 4 名, 一共 10 名, 分别予以考察。高级汉语水平学习者均为学习汉语时间为两年以上的北京大学学生, 均有汉语水平考试(HSK) 高级证书。为方便起见, 在本文中中国发音人标为 C; 韩国发音人标为 K; 男声标为 M; 女声标为 F。

2.3 录音材料

词表包括双音节 20 组(阴平+阴平、阴平+阳平、阴平+上声、阴平+去声等), 共形成 20 组实验材料 (参看附录 1), 并加两组分别放在前后以防录音效果的阻碍, 分析时删掉这两组。20 组实验材料的呈现顺序经过随机化处理。本文其中只考察“前字阳平”的习得情况, 以后描写阳平的表现形式及讨论原因时我们要以其他双音节调的分析结果为参考数据。

2.4 录音过程和数据处理

(1) 录音: 被试朗读词表。

(2) 字组切分: 把被试的录音材料输入计算机文件中, 使用 Praat 把样本切分为单/双音节。

(3) 分析: 再以自编的 Matlab 程序来提取基频。

再将各个样本的声调曲线频率值进行时长的归一化处理，归一化后一共有 5 个数据点，这是为了能够更加准确地反映出声调的调型特征。然后根据不同发音人各自的调域进行归一整理，就可以得到 5 度制的参考数值。声调音高频率数据和 5 度值之间的对应关系可以采用如下的公式来计算，把各个样本的音高数值变换为 5 度制：

$$5\text{度值} = ((\log x - \log b)/(\log a - \log b)) \times 4 + 1$$

其中，a 为调域上限频率，b 为调域下限频率，x 为测量点频率。得出的值就是 x 点的五度值参考标度。

3. 实验结果

3.1 中国学生(参照群)的前字阳平

下面图 1 给出的是从两位中国学生（CM,CF）所发阳平双字调样本中提取并计算所得的 5 度值数据。下文中韩国学习者的实验结果都以图 1 的数值为参考值。

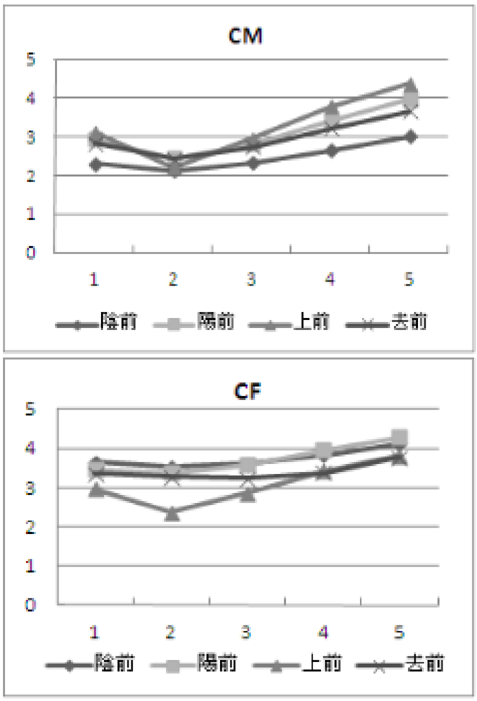


图 1 中国学生的普通话前字阳平曲线 (5 度值)

根据前人的实验结果，不管后字是什么声调，前字阳平 F0 曲线大多数呈中降升型（林茂灿，1987）。而后来郭锦桴（1991）指出，除了个别字外，大多数阳平调呈升调型，没有中降现象。

从图 1 看出，两位中国发音人的阳平曲线和上述的研究结果基本相同，即双字组前字阳平曲线呈中降升型。值得注意的是男女发音人的阳平曲线中我们没有发现调值为 35 的所谓典型的阳平，即男声的阳平曲线主要集中在 2-4 度之间，而女声的有集中在 3-4 度的倾向，与一般的阳平标准有所出入。

至于韩国学生的阳平，据前人的描写，他们的主要偏误为“高音上不去”，即大多数把它发成低平调型。那韩国学习者到了高级阶段之后的“阳平”习得情况是否与初级阶段或其他研究结果一致？

3.2 高级阶段韩国学习者(参照群)的普通话前字阳平

下图 2 显示 10 位高级阶段韩国学习者的双字组阳平的实验结果。为直观起见，先没有以平均值来表示，而 10 位学生的实际结果直接给出。

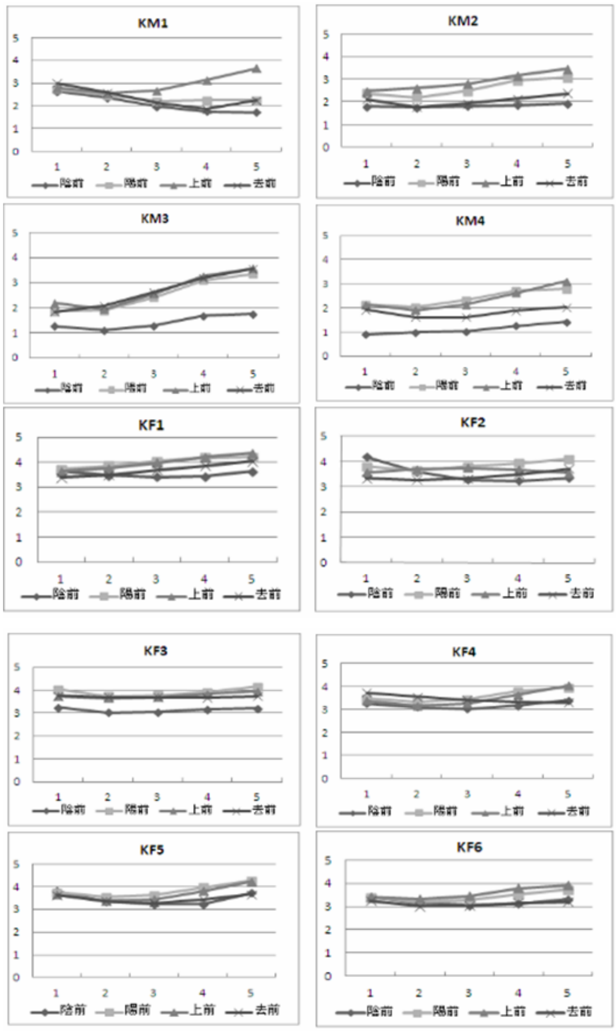


图 2 高级阶段韩国学习者的普通话前字阳平曲线(5 度值)

从上图我们可以发现，男女之间存在着不小的差异：整体来看，韩国女声的阳平曲线集中在3-4 度，

和中国女声一样调域非常窄,自然升的幅度也很小。与此相比,韩国男声的阳平曲线的变化比较大;整个调值的范围也从 1 度到 4 度,自然其调域比女发音人要宽得多。下面我们看一下他们的调型分布情况。

表 1 高级阶段韩国学习者的前字阳平调型分布

(注:为区别起见,降升型中,凡是始点高于终点的调型都标为“曲折型”)

		降升型	升型	曲折型	降型	其他
阴前	男	2	1		1	
阴前	女	1		5		
阳前	男	2	1	1		
阳前	女	5	1			
上前	男	3	1			
上前	女	4	1			1
去前	男	2	1	1		
去前	女	2	1	2	1	
总合		21	7	9	2	1

据上文的分析结果,中国发音人的阳平调型一律是降升型,而上表 1 的调型分布表示,高级阶段韩国学生的阳平调类型有降升型 21 个、升型 7 个,一般归为偏误的曲折型和降型分别有 9 个和 2 个。可以说韩国发音人的阳平调型中 70%基本接近普通话阳平的调型。考虑到他们的汉语水平,此结果还是可以预料到的。在他们的偏误类型中,“曲折型”偏误是最明显的特征,此现象符合前人的研究结果:即,终点“升不到位”的偏误现象。值得关注的是,此现象只在女发音人身上比较普遍,尤其是阴平前的阳平,将“降型”偏吴也归到此类的话,去声前的阳平也可以加上。从此我们可以知道,声调习得方式以及化石化程度男女之间有所差异。

下面表 2 和表 3 分别给出高级阶段韩国学习者的阳平调值与调域平均分析结果。

表 2 高级阶段韩国学习者的前字阳平调值平均 (5 度值)

	CM	KM	CF	KF
始点	2.79	2.73	3.36	3.88
折点	2.31	2.48	3.13	3.71
终点	3.75	3.14	4.00	4.03

首先,从表 2 我们再次发现中国发音人的调值与所谓的标准调值不太一致。林焘(1996)先生曾指出:“从现代语音学实验分析的结果看,在同一个调域里单说这四个调类,并不总是这四种调值。如果用五度值表示,阴平也可以是[44],上声也可以是[212]或[312],去声也可以是[41]。阳平的变化最多,既可以读成[25]或[24],也可以读成[325]或[425]”。而 CM 的平均调值为[223],CF 的则是一种不同的类型[334]。难道中国发音人的阳平也有问题吗?当然不是。到此,我们应该质疑阳平的标准调值到底是多少?其界限是在哪里?这都是有待进一步研究的问题。至于 KM 和 KF 的调值也跟刚才提到的调值类型有所不同,KM 和 KF 的调值平均值分别为[223]和[334]。从调值的平均值来看,韩国男发音人呈现出终点稍微低的降升型阳平,而女发音人的阳平是一种虽然保持着降升型,但其曲线的变化极小,更接近于高平调型。

下面表 3 给出的是韩国学习者的前字阳平调域分析结果。

表 3 高级阶段韩国学习者的前字阳平调域平均 (5 度值)

	CM	KM	CF	KF
阴前	0.89	0.45	0.59	0.32
阳前	1.51	0.73	0.91	0.43
上前	2.15	0.99	1.45	0.47
去前	1.22	0.77	0.54	0.28
平均	1.44	0.73	0.87	0.38

表 3 的结果显示各群组之间的异同:首先,不管中国发音人还是韩国发音人,男发音人的调域比女发音人的调域宽一倍左右。另外由于协同发音的影响,阴平和去声之前的调域比阳平和上声之前的调域要窄一些。就是说,此现象具有普遍性,并不是特定国家发音人的特征。两国发音人之间的差异则在于,无论男女,韩国发音人的调域只有中国发音人的一半,平均调域都不到 1 度。

归纳上文的几次调查结果可知,高级阶段韩国

学生中,男发音人的阳平调值更接近于标准阳平。虽然终点还不到位,但具有降升型的调型,调域也相对来说比较宽。相比,女发音人的阳平由于调域太窄,以 5 度值来换算的话,哪怕是一个正确的调型也只能标成接近阴平的高平调。

综上所述,本次实验呈现以下几个高级阶段韩国学习者的阳平特征:

(1)调型正确率比较高,偏误类型中“曲折型”最为突出,尤其是女发音人的阴平前的阳平;

(2)在调值上,男发音人的阳平有终点稍微低的倾向,而在女发音人的表现上,全调值过于高,并且幅度极小的现象非常普遍;

(3)无论什么性别,调域非常窄,大部分都五度值为 1 度以下。

考虑到被试的汉语水平和学习时间,这些特征可能是从初级阶段开始慢慢形成的语音习惯。从中介语理论的角度说,也可能是“化石化”的表现。那为什么到了高级阶段还出现这种现象呢?下面我们继续讨论一下其原因是什么。

4. 讨论

韩国学习者声调格局的形成原因可能来自于很多方面,这在学界具有一定共识,其中主要的原因有:

(1)声调习得的普遍性。Ohala 等(1973)研究表明,“升调难,降调容易”是人类共同的普遍特征。对“阳平调”的难度,从中国儿童和其他母语背景的外国学生的汉语声调习得研究中也可以得到验证(王韫佳,1997);

(2)有的与普通话的声调格局有关。如混淆调型相似的声调(阳平和上声);

(3)难以控制声带。从生理的角度说,语音的高低,即声调的变化是由声带振动的快慢所调解的结果。对母语是非声调语言的学习者来说,发音机制、振动频率的快慢调节是需要经过不断训练才能形成的;

(4)母语的负迁移。由于韩语是非声调语言,即音节层里的不同音高在韩语中没有特殊意义。因此,从来没接触过“声调”这个概念的韩国学生,在掌握声调的过程中会遇到很多问题(Ohala、W.Ewan,1973;冯丽萍、胡秀梅,2005;王韫佳,1997)。

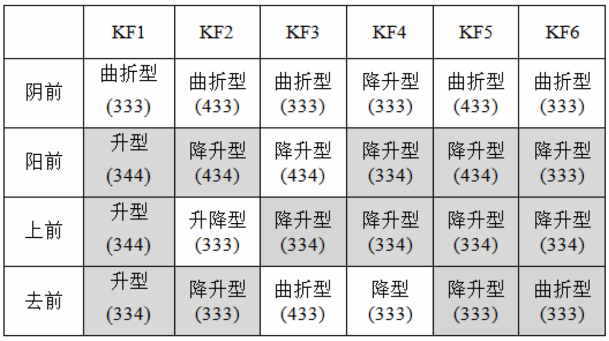
在前人研究的基础上,本文要再加上另外两个方面的原因,从而探讨根据声学分析判断为偏误的声调特征是否可以断定为非得纠正的“偏误”现象。

4.1 汉语母语者感知的影响

林焘、王士元(1984)曾经对汉语母语者做过合成语音的感知实验,实验结果表明,随着后字的音高变化,原来平调的前字有可能听成升调。即,前字的音高和后字的音高之间有一定的距离就会产生感知上的调类变化,此现象所形成的就是“一低一高”模式。从该实验结果以及上文的中国母语者的阳平表现可以推断,外国学习者的前字阳平发成“平调”的现象与汉语母语者的实际发音情况和“听错觉”有一定的关系。因此,我们以同样的阳平样本对两位汉语母语者进行了听辨实验,本实验采用随机播放的方式进行。从这次实验得出的结果如下。下表 4 中填心的部分为两位汉语母语者都听成“阳平”的样本:

表 4 韩国学习者的普通话前字阳平调型和汉语母语者的感知(5 度值)

	KM1	KM2	KM3	KM4
阴前	降型 (322)	降升型 (222)	降升型 (212)	升型 (112)
阳前	曲折型 (322)	降升型 (223)	升型 (223)	降升型 (223)
上前	降升型 (333)	升型 (233)	降升型 (223)	降升型 (223)
去前	曲折型 (322)	降升型 (222)	升型 (233)	降升型 (222)



	KF1	KF2	KF3	KF4	KF5	KF6
阴前	曲折型 (333)	曲折型 (433)	曲折型 (333)	降升型 (333)	曲折型 (433)	曲折型 (333)
阳前	升型 (344)	降升型 (434)	降升型 (434)	降升型 (334)	降升型 (434)	降升型 (333)
上前	升型 (344)	升降型 (333)	降升型 (334)	降升型 (334)	降升型 (334)	降升型 (334)
去前	升型 (334)	降升型 (333)	曲折型 (433)	降型 (333)	降升型 (333)	曲折型 (333)

表 4 的结果显示,不管发音人的调值(5 度值)是多少,调域宽与窄,只要调型为升调或者是降升调就绝大多数被听成阳平调。出现的几个例外可能是由音高差和折点的位置不同而造成的结果。另外,按照标准调值来判断的话正确的几乎为零,而在汉

语母语者听辨实验其正确率则上升到50%了。这说明高级阶段的学生并不严格按照标准调值(35)发阳平,但母语者的实际感知效果反而增加。由此我们值得再考的问题是,在实际交际当中或者对母语者而言,所定为偏误的现象往往并不是使学习者非得纠正的“错误”。虽然本次实验比较简单,听辨人也只有两个人,但结果足以说明高级阶段韩国学习者的阳平表现和汉语母语者阳平感知有一定的关系,对汉语声调的实际范畴和感知对学习策略的影响有待继续研究。

4.2 高级阶段学生的学习策略

在中介语研究中,我们是否一直忽略学习者的学习能力?在学习第二语言的过程中,由于成人外语学习者有了比较完备的认知条件,自己去处理是完全可能的(董燕萍,2005)。一般来说,高级阶段学生说话又要省力也要流利,因此,随着学习者语言水平的提高,自己已确立的学习策略和发音方式也在不断调整。自然,对学习者而言,汉语母语者的回馈极其重要,这直接影响到学生的习得过程。也就是说,我们在上文分析的高级阶段学生的阳平表现很可能是已经经过无数次的反馈和调整而确立的最省力,最有效的交际方式。上文的听辨实验结果已初步证实了高级阶段学习者所采取的“一低一高”策略是有一定效果的。

5. 小结

由于被试的母语(韩语)缺乏丰富的调值变化,因此对韩国学生来说声调习得确实是一件很困难的事。在各项调查里,我们基本了解到高级阶段韩国学习者的声调习得特点。总体上看,他们在调型上正确率比较高,发成降升型和升型的达到70%,最明显的特点为女发音人的阴平前的阳平发成曲折型的现象比较普遍。在调值上,男发音人的阳平相对来说好一些,除了终点稍微低之外,比较接近标准调值。而女发音人阳平显示整个调值过于高,并且升的幅度非常小的倾向。男女发音人的共同特点为调域都特别窄,均为中国人的一半。

进一步,本文还通过母语者的听辨实验探讨了产生高级阶段学习者将阳平发成平调的原因。我们考虑到外语学习过程中的的双向性(学习者和母语者的交际)和学习策略,认为高级阶段汉语学习者的阳平表现,除生理、声调格局和母语的影响之外,还会受到“汉语母语者的听错觉”和“高级阶段学生的认知策略”的影响。从此我们发现,在教学过程中一直被视为偏误的现象,往往和实际母语者的表现相当一致:如,高调前的阳平表现为平调的现象,在母语者的感知上也不一定认为是偏误。可以说,通过本次实验高级阶段学习者的阳平加工方式及学习策略的有效性得到初步证实。总之,从母语者的实际表现和感知的角度看,所谓“偏误”的范围问题还是值得商榷。

由于母语类型各异,中介音也呈现出不同的特征。找出不同母语留学生中介音特征,可以针对不同国家的留学生设计对策,促使中介音更快地向目的音转化。但,本次实验也再次提醒我们,首先要明确目的语的实际标准和偏误的定义,才能够进行准确的分析,从而实现合适的纠偏。

6. 参考文献

[1] 董燕萍(2005)«心里语言学与外语教学»,外语教学与研究出版社,北京

[2] 冯丽萍,胡秀梅(2005)零起点韩国学生阳平二字组声调格局研究,«汉语学习»第4期

[3] 高玉娟,李宝贵(2006) 韩国留学生汉语声调习得偏误的声学研究,云南师范大学学报

[4] 林涛、王士元(1984)调感知问题,«中国语言学报»,第二期

[5] 马燕华(1994)初级汉语水平留学生的普通话声调误区,«北京师范大学学报»,第3期

[6] 吴宗济、林茂灿 主编(1989)«实验语音学概要»,高等教育出版社,北京

[7] 石锋(1990)«语音学探微»,北京大学出版社,北京

[8] 石逢,廖荣蓉(1994)«语音丛稿»,北京语言学院出版社,北京

[9] 宋春阳(1998)谈对韩国学生的语音教学-难音及对策,«南开学报»,第3期

[10] 孙德坤(1993)中介语理论与汉语习得研究,«语言文字应用»,第4期

[11] 王韫佳(1995)也谈美国人学习汉语声调,«语言教学与研究»,第3期

[12] 王韫佳(1997)阳平的协同发音与外国人学习阳平,«语言教学与研究»,第4期

[13] 张红(2003)«日本留学生汉语声调习得及偏误分析»福建师范大学 硕士学位论文

[14] 朱川(1997)«外国学生汉语语音学习对策»,语文出版社,北京

[15] 孟柱亿 (1992)«汉中中间语言研究:以音韵论为中心»,

韩国外国语大学院 博士学位论文

[16] Ohala, J. and W. Ewan (1973). Speed of pitch change, «Journal of Acoustical Society of America »

附录 1. 词表

后字 前字	**阴平**	**阳平**	**上声**	**去声**	**轻声**
阴平	高低	孤独	跌倒	包庇	勾搭
阳平	白搭	独白	毒打	达到	鼻头
上声	古都	抵达	打倒	北部	搞的
去声	地瓜	告别	递给	地步	地道

实际分析时删掉的前后两字: 大 打

长时共振峰分布特征在声纹鉴定中的应用

曹洪林 [1] 孔江平 [2]

（1 北京大学中文系 caohonglin@pku.edu.cn；2 北京大学中文系 kongjp@gmail.com）

摘要：

利用长时共振峰分布特征区分不同发音人是近年来新兴的一种鉴定方法，本文对此进行了介绍和分析。本文以汉语普通话为语料，对 20 位男性发音人和 10 位女性发音人进行了研究，分析了所有发音人前四条共振峰的长时分布情况，发现各条长时共振峰分布的均值、中位数、众数、峰数、峰度和倾斜度等参数能够反映出不同发音人的个性特征，而且稳定性较强。前四条长时共振峰的分布结构比较均匀，由此推测在语料足够长的情况下，所有元音平均后的结果应该是一个类似央元音的“音”。本文还对长时共振峰分布特征在声纹鉴定中的具体应用进行了一些讨论。

关键词：

普通话；长时平均分布；共振峰；声纹鉴定。

1. 引言

共振峰是声纹鉴定中最重要的特征之一，它能够提供很多发音人的个性特征。目前，对于共振峰的利用，鉴定人员经常从定量和定性两个角度进行分析。定量分析共振峰频率的方法有很多，其中最为经典的还是测量不同元音（稳定段）共振峰中心频率值的方法。最近 McDougall[1-2]提出了一种定量分析共振峰动态特性的新方法，即对复合元音和响音的共振峰进行多项式拟合，引入判别式对共振峰和拟合系数进行判别分析；李敬阳等[3]利用类似的方法，对汉语普通话复合元音共振峰的动态特性进行了定量分析研究，而在此之前许多有关音节内和音节间共振峰动态特性的研究一直停留在定性比较层面（有时候也会结合测量元音共振峰最低点和最高点的频率值进行分析）。Nolan 与 Grigoras[4]提出了第三种测量共振峰频率的方法，即长时共振峰分布测量法（Long-Term Formant Distributions，缩写为 LTF）。与前两种定量分析方法不同，该方法不是分析具体的目标元音，而是提取一整段语音的全部信息进行分析，得出每条共振峰的整体分布情况，该分布特征可以用于区分不同发音人。该方法已在德国联邦调查局 BKA 的语音分析实验室得到广泛应用[5]。此外，Jessen 与 Becker[6]研究发现，LTF 方法还有很多优点：LTF 数值（如 LTF2 和 LTF3 的均值，另见[5]）与发音人的身高呈负相关关系、不同鉴定人员之间测量同一语料的一致性较高以及 LTF 数值在不同语言之间的差异性较小等。Becker 等[7-8]还发现，LTF 分布可以作为说话人自动识别的一个新特征。Moos[9-10]则对 71 位德语男性发音人的 LTF 分布进行了测量分析，建立了一个较有价值的参考数据库，同时发现朗读时的 LTF 数值要比自然说话时的稍高。

上述有关 LTF 方法的研究主要是针对英语和德语等语言的男性发音人展开的，迄今未见有关汉语的研究。本文将以汉语普通话为对象，介绍并分析 LTF 方法在声纹鉴定中的应用。

2. 实验方法

2.1 发音人

20 位男性发音人，年龄在 19-36 岁之间（均值 27.9，方差 4.8），10 位女性发音人，年龄在 21-30 岁之间（均值 24.5，方差 2.91），所有发音人均能说比较标准的普通话，录音时身体健康，无嗓音疾病、感冒等症状。

2.2 语料

录音材料分为两个部分，一部分是普通话的 4 个单元音[a]、[i]、[u]和[ə]，另一部分是长篇语料《北风和太阳》。

2.3 录音

在北京大学语言学专业录音室中，使用 SONY ECM-44B 领夹式麦克风录音，采样频率为 22kHz，

精度为 16 位。在录音过程中，首先让发音人熟悉语料，并朗读一遍，然后再使用正常的语速和音量将语料自然说出，尽量避免使用朗读的方式。发单元音时，要求发音人持续发音 3 秒钟以上，两部分录音材料均录制两遍。

2.4 测量

使用 WaveSurfer [11]软件进行切音、提取语音共振峰，提取过程见图 1。具体设置如下：哈明窗，LPC 系数：12，共振峰数量：4，截止采样频率：10 或 8kHz（男声），10 或 11kHz（女声）[1]。提取单元音的共振峰时，选择中间 1-2 秒的稳定段进行测量，取两次发音的平均值进行分析。分析长篇语料时，选择第二次录音做研究使用，为了获得清晰准确的共振峰结构，笔者只对语料中共振峰结构明显的元音、边音[l]、语流中间浊化的[h]等浊音（以及部分发音人的填声停顿和犹豫词）进行了分析，而将原始语料中的无声段、呼吸声、含鼻音的音节以及共振峰结构不明显的元音部分剪切掉，保存为新的 wav 文件作为分析对象。语料剪切前后的时长信息见表 1。针对自动提取共振峰不正确的语音，采用手动调整加以修正。

对分析得到的各条共振峰（F1-4）数据进行统计分析，做出直方图（分箱间隔为 25Hz），使用 Matlab 软件中的 Fourier 6 阶拟合函数进行曲线拟合。

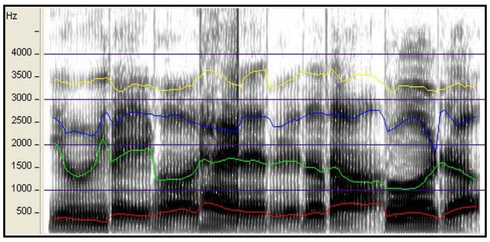


图 1：使用 WaveSurfer 软件提取一段元音的共振峰。其中，自下往上的红线、绿线、蓝线和黄线分别表示 F1-4 的数据。

表 1：篇章《北风和太阳》剪切前后时长对比。

时长(秒)	分类	最大值	最小值	均值	方差
原始	男	46.3	29.1	34.9	4.3
	女	36.7	29.6	32.7	2.0
剪切后	男	16.0	9.3	11.7	1.8
	女	13.8	10.8	12.3	1.0

3. 结果

3.1 不同发音人的 LTF 分布特征

5 位男性发音人和 5 位女性发音人的长时共振峰分布情况见图 2。图中的蓝色部分是第 1 至 4 条共振峰（F1-4）每个频率段的累积直方图，图中从左至右的黄色、红色、绿色和紫色 4 条曲线分别是对 F1-4 直方图中各顶点位置对应数字（达到一定频率范围的次数）进行拟合得到的曲线，该拟合曲线清晰地反映了各条共振峰的整体分布情况。由于语料中存在不同的元音，其共振峰结构不同，拟合曲线之间存在重叠现象。每条共振峰分布曲线的峰值基本上代表了该共振峰分布的众数（即出现次数最多的频率段。例外情况见图2中M5和Fe5的LTF4）。

比较图 2 中的 5 位男性发音人（M1-5），可以发现他们LTF1-4分布曲线的峰值均不相同。比如，M1 的 LTF3 峰值为 2437Hz（众数 2475Hz），M3 的 LTF3 峰值却相对较高为 2656Hz（众数 2650Hz）；M2 的 LTF4 峰值为 3517Hz（众数 3525Hz），比 M4 的 LTF4 峰值（3817Hz，众数 3800Hz）要低 300Hz。同时，各条曲线的形状也有很大差异，有的出现单峰，有的出现双峰或 3 个峰。比如，M1 的 LTF4 表现为双峰，而 M4 的 LTF4 则为单峰，M1 的 LTF2 有 3 个强度（对应的纵轴数值，即出现次数）较低的峰，而 M4 的 LTF2 则有两个比较强的尖峰。

整体而言，LTF1 的分布范围最小，也最为集中，均有比较突出的“尖峰”出现；LTF2 的分布则较为分散，少有比较突出的“尖峰”出现，常常表现出比较“扁平”的特点；LTF3 的分布范围也比较大，但与 LTF2 不同，LTF3 常常出现“尖峰”，强度常略小于 LTF1；LTF4 与 LTF3 的情况比较类似。

不同人的 LTF 分布的峰度（kurtosis）存在较大差异，当某条共振峰的变化范围较大时，其 LTF 的分布形状就比较扁平（platykurtic），相反，有些人的高次共振峰比较稳定，变化较小，其分布形状就多会出现“尖峰”（leptokurtic）。从图 2 中 10 位发音人的 LTF 分布可知，分布曲线接近正态分布的较少（如 Fe2 的 LTF3），多数都是非对称性分布，其倾斜度（skewness）参数之间也有较大差异。当 LTF 的峰值（众数）较大，高频数据出现数量多于低频数据时，曲线往往向高频方向倾斜（如 M4 的 LTF4、Fe5 的 LTF3），相反，曲线就会偏向低频方向（如 M4 的 LTF3、Fe4 的 LTF3）。图 2 中男女性发音人 LTF1-4 的分布情况类似。不同的是，相比之下，女性发音人 LTF1-4 的峰值数据和分布范围均有明显的提高。这与女性发音人的声道长度一般比男性发音人的短，女声的共振峰频率更高等特点是相吻合的。

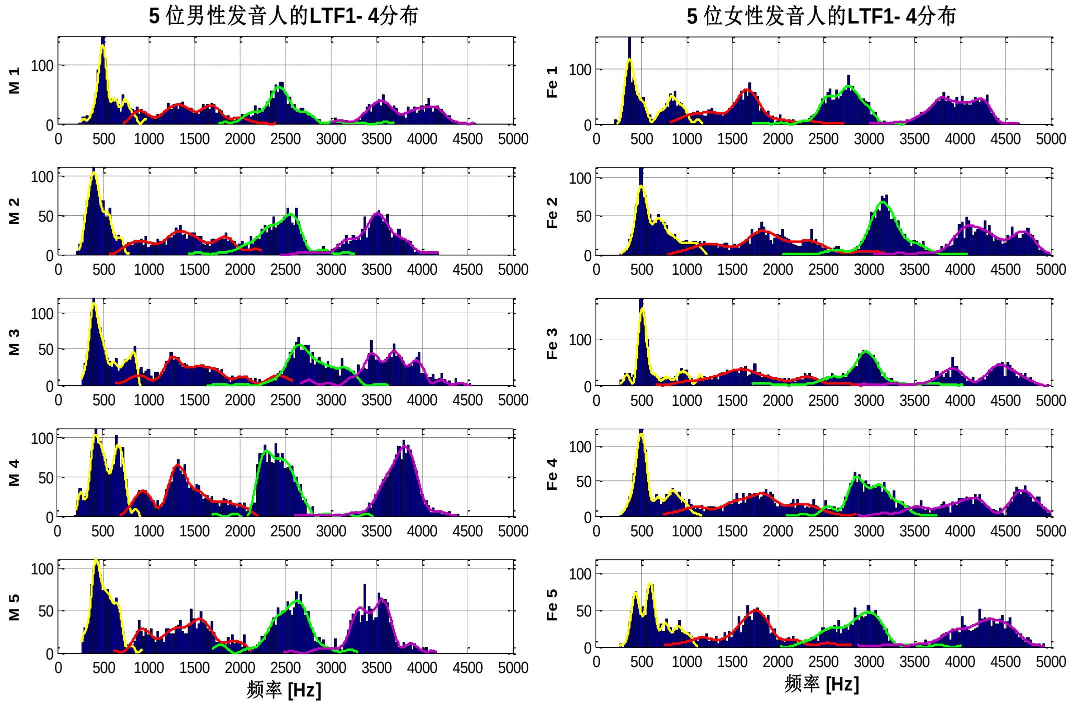


图 2：5 位男性发音人（M1-5）和 5 位女性发音人（Fe1-5）的 LTF1-4 分布。图中的黄线、红线、绿线和紫线分别表示 F1、F2、F3 和 F4 的长时分布情况。

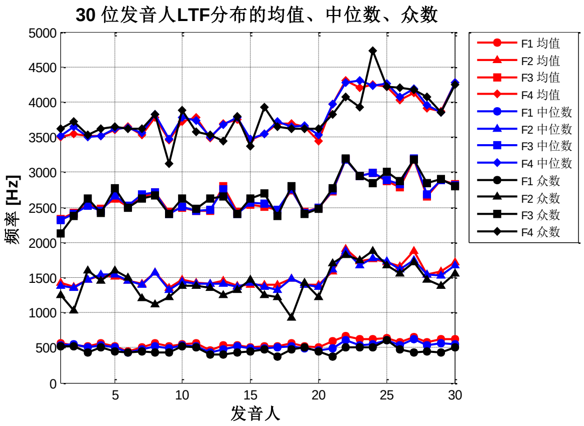


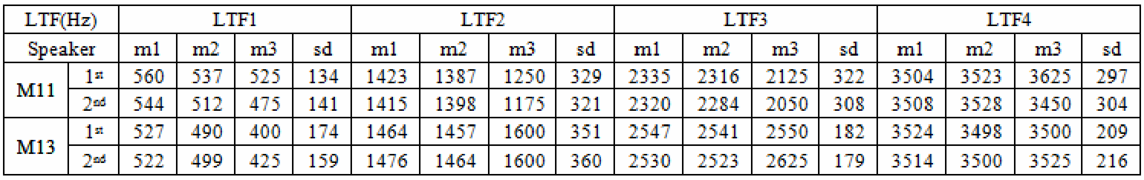
图 3：20 位男性发音人（1-20）与 10 位女性发音人（21-30）LTF 均值、中位数与众数比较。

图 3 显示的是 30 位发音人 LTF 分布数值的均值、中位数和众数（如前文所述，该参数多数情况下与 LTF 拟合曲线的“峰值”相对应）。由于 LTF1-4 均非正态分布，上述 3 个参数值并不重合，三者越接近，其对应的 LTF 分布（或其中的一部分）就越接近正态分布。比较图 2 中发音人 Fe2 、Fe3 的 LTF3 分布情况与图 3 中对应女性发音人（22、23，黑色圆圈内）LTF3 的 3 个参数的重合情况。整体观察可知，图 3 中 LTF 的均值和中位数比较接近（重合性较好），而众数与前两者的差异较大（重合性较差），这种差异性的存在正说明了不同发音人 LTF 分布的独特性。

3.2 同一发音人的 LTF 分布特征

图 4 显示的是两位男性发音人两次说长篇语料的 LTF 分布情况。很明显，在 M13 的两次发音中，F3 和 F4 比较稳定，对应 LTF3 和 LTF4 均出现单个比较突出的“尖峰”，而 M11 的 LTF3 和 LTF4 的变化范围则更大一些，从曲线形状上也一致表现的比较“扁平”，出现多个不太突出的“小峰”。两人两次发音 LTF1-4 的均值、中位数、众数和标准差见表 2。由图 4 和表 2 可以看出，同一人相同状态下发音的 LTF 分布变化较小，较为接近。

表 2：两位男性发音人两次发音 LTF1-4 的均值（m1）、中位数（m2）、众数（m3）和标准差（sd）



LTF(Hz)		LTF1				LTF2				LTF3				LTF4			
Speaker		m1	m2	m3	sd	m1	m2	m3	sd	m1	m2	m3	sd	m1	m2	m3	sd
M11	1st	560	537	525	134	1423	1387	1250	329	2335	2316	2125	322	3504	3523	3625	297
	2nd	544	512	475	141	1415	1398	1175	321	2320	2284	2050	308	3508	3528	3450	304
M13	1st	527	490	400	174	1464	1457	1600	351	2547	2541	2550	182	3524	3498	3500	209
	2nd	522	499	425	159	1476	1464	1600	360	2530	2523	2625	179	3514	3500	3525	216

4 讨论

4.1 LTF 与央元音[ə]的关系

顾名思义，长时共振峰分布是对一段语料中元音的各条共振峰（F1-4）的所有数值分别进行平均，查看其整体的分布情况。从图 3 的结果来看，发音人 LTF1-4 相邻共振峰之间的距离比较平均，如 20 位男性发音人 LTF1-4 均值参数的平均值分别为 524Hz、1436 Hz、2523 Hz 及 3600 Hz，与成年男性理想状态下简单均匀声管模型的央元音的共振峰结构相似（第 n 条共振峰的频率为 500*（2n-1）Hz，参考[12]）。图 5 比较了 20 位男性发音人的 LTF 均值、中位数和众数三个参数的平均值在声学元音图中的位置。由图 5 可以看出，LTF1-2 的均值、中位数和众数均位于元音三角形内部，分布在比较中间的位置，与央元音[ə]比较接近，与[ə]的 F1 和 F2 相比，LTF1 稍低，LTF2 稍高。众所周知，不同元音的共振峰结构不同，发元音[i]时，前腔面积小、后腔面积大，[i]的 F1 最低、F2 最高，发元音[a]时，前腔面积大、后腔面积小，[a]的 F1 与 F2 最为接近。在连续语流中，如果时长足够长，语料中出现大量的单元音、复元音时，尽管有的元音能够达到其目标值，有的不能，但总体平均后的结果都应该是一个类似央元音的“音”，其前腔、后腔的面积相当，声道形状应该与自然状态下发[ə]时的声道形状类似。

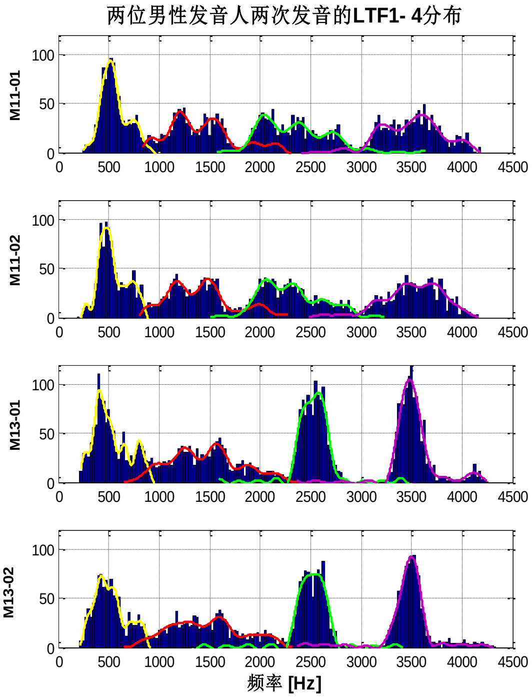


图 4：两位男性发音人两次发音的 LTF 分布情况。

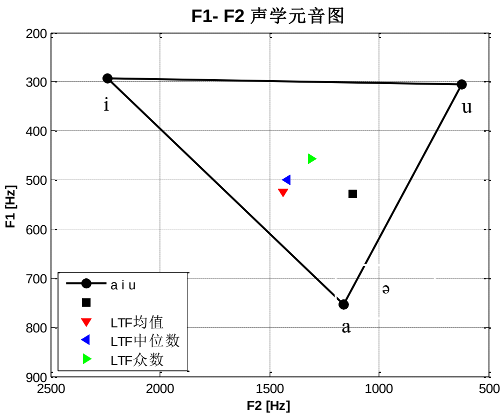


图 5：LTF 在声学元音图中的位置（基于 20 位男性发音人的平均数据）。

4.2 LTF 方法的应用

在声纹鉴定中，任何长时特征的利用，都有几个共同的问题需要加以明确，LTF 方法也不例外。

第一，要合理利用该特征，最少需要多长的语料？Catalina 推荐的时长标准是不少于 10 秒[13]，需要注意的是，他提到的时长是原始检材和样本的时长，语料中包含了辅音等非元音的成分。Moos 研究认为，根据共振峰和说话状态的不同，对于只包含元音的语料，比较合理的时长下限是 5-8 秒左右，作者同时还强调，在不同人之间，该标准会发生变化[9]。尽管本文语料长度在 11-12 秒左右（详见表 1），能够满足 Moos 的标准要求，但鉴于两项研究的语种不同，应该对基于汉语普通话的时长标准做进一步的分析。第二，是否与文本内容有关？本文与 Moos 的研究语料均为《北风和太阳》的寓言故事（后者研究的是德语），内容都是固定的（与文本有关）。尽管 Nolan 与 Grigoras 在提出 LTF 方法的时候，曾将案件的样本语音拆分成两部分（将样本 K 分为两半：K1 与 K2）进行分析[4]，发现原始样本与拆分后的前后两部分的 LTF1-4 有很好的一致性，但作者并未交代前后两部分的内容是否一致，即使不一致，也只有单个人的数据，若要增加可信度，则需要进行更多发音人的测试。可以设想当一段语料足够长，期间出现各个元音的几率相当时（与文本无关），LTF 分布特征应该会趋于稳定，当然这还需要作进一步的研究。第三，是否与语言/方言有关？LTF 分布主要反映了发音人声道的整体共鸣特性，如果语料足够长，同一人说不同语言/方言时，由于其 LTF 反映的是同一人的整体声道特性，这些被体现出来的声道特性之间的差异性应该不大。当然，不同语言/方言之间音系结构上的差别有可能会成为较大的影响因素。如果 LTF 在不同语言/方言之间的差异性足够小的话，那么由一种语言/方言得到的 LTF 数据便可以推广利用到另外一种语言/方言中去，这便为实践中不同语言之间、方言与普通话之间的语音比对提供了新的途径。Jessen 与 Becker[6]的研究已经发现,LTF 在德语、俄语和阿尔巴尼亚语之间的差异性较小。有必要针对汉语普通话与方言及其他语言之间 LTF 的差异性展开进一步的研究。

在实际办案中大多数案件语料都是电话（或手机）录音，由于电话信道的带宽限制（300-3400Hz 左右），多数情况下只能显示前 3 条共振峰（在录音质量较差的情况下，有时只能显示两条共振峰）。从图 2-4 中也能看出多数 F4 的数值超过了这个的范围，不能或不能完全显示出来（可以导致 F4 的提取不准确），因此，实践中可以测量前 3 条共振峰的长时分布特征（LTF1-3）。但是对于一些由性能较好的录音设备（如安全机关使用的部分专业监听设备或部分民事案件中当事人使用的录音笔）录制的采样频率较高的语料，可以测量到 LTF4。当然该方法也可以用来进行普通语音学研究，测量的共振峰数量可视研究目的而定。

由于 F3、F4 高次共振峰的稳定性最强，在鉴定中更有价值，因此 LTF3 与 LTF4 的数值及分布特征应该能更好地体现发音人之间的个性差异。至于哪一条共振峰分布特征的价值更高，则需要进行更深入的研究。当然，在鉴定中比较明智的做法是对 LTF1-3/4 进行综合分析，而不是只看某条共振峰的分布特征。

如上文所述，使用 LTF 方法的前提是被研究语料的录音质量较好，能够反映出比较清晰的共振峰结构，在此基础上对各条共振峰的分布进行统计分析，才能够显示出共振峰整体分布的情况。由此得到的 LTF 分布不仅在峰数和形状（峰度、倾斜度）上能够较好地体现出发音人的个性特征，而且还能提供各条共振峰整体分布的多维数据，像均值、中位数、众数、标准差等，这些数据为研究发音人的声道特性提供了新的素材。正如 Jessen 对该方法的评价：“此方法的优点是使用相对高效省时，适用范围广，甚至连鉴定专家自己都不会讲的语言都能够适用（因为该方法无需对元音的音系范畴进行辨别和切分，仅需要具备较好的普通声学语音学的知识即可）；缺点是，由于该方法集中了所有的元音信息，与分析单个元音相比，对结果进行解释变得更加困难。对不同的元音进行集中分析，很可能会忽略一些更有价值的个性特征。”[14]的确如此，由于 LTF 只是一个静态特征，因此在语音鉴定时不应该单独使用它，而是要结合其他特征进行综合分析，特别是要与单元音的共振峰频率、音节内及音节间共振峰的动态特征结合起来一起使用。

5 结论

本文对 20 位男性发音人和 10 位女性发音人的普通话语料进行了研究，通过对其元音共振峰频率的统计分析，得出了第 1 至 4 条共振峰（F1-4）的长时分布情况（LTF1-4）。相比男性发音人而言，女性发音人的 LTF 数据和分布范围均有明显的提高，这与女性发音人的声道长度较短、共振峰频率较高等特征是相吻合的。利用 LTF 分布的均值、中位数、众数（相当于峰值）、峰数和形状（峰度和倾斜度）

等参数可以较好地区分不同发音人。

通过比较 LTF 与央元音[ə]的共振峰的关系发现，相邻 LTF 之间的距离比较平均，LTF1-4 的整体分布结构与央元音的共振峰结构类似。可以推测在连续语流中，如果时长足够长，其所有元音平均后的结果都应该是一个类似央元音的“音”。

尽管 LTF 方法可以用来区分个人，但不可否认的是，LTF 分布只是一个静态参数，鉴定中不宜单独使用，而应该与其他动态特征结合起来做综合分析。本文仅对 LTF 分布特征做了概括性的介绍，今后有必要对 LTF 分布特征与时长、文本内容及语言的关系等问题进行进一步的探讨。

6 致谢

本研究得到国家自然科学基金项目“汉语普通话生理模型研究”的支持（批准号：61073085）。感谢 Michael Jessen 博士对 LTF 方法在 BKA 实验室中使用情况的介绍，感谢王英利老师、李英浩博士、董理博士对本文初稿提出的部分修改意见。

7. 参考文献

[1] McDougall K. Speaker-specific formant dynamics: an experiment on Australian English /ai/[J]. International Journal of Speech, Language and the Law, 2004, 11(1): 103-130.

[2] McDougall K. Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies [J]. The International Journal of Speech, Language and the Law. 2006, 13(1): 89-126.

[3] 李敬阳，王莉，崔杰等. 发音人汉语普通话复合元音共振峰动态特征分析 [C]. 第一届全国声像资料检验鉴定技术交流会论文选, 公安部物证鉴定中心，北京：中国人民公安大学出版社，2011, 612-615.

[4] Nolan F, Grigoras C. A case for formant analysis in forensic speaker identification [J]. The International Journal of Speech, Language and the Law. 2005, 12 (2): 143-173.

[5] Jessen M. The forensic phonetician forensic speaker identification by experts. In: Coulthard M, Johnson A. (eds), The Routledge Handbook of Forensic Linguistics, 2010, 378-394.

[6] Jessen M, Becker T. Long-term formant distribution as a forensic-phonetic feature [J]. The Journal of the Acoustical Society of America. 2010, 128: 2378.

[7] Becker T, Jessen M, Grigoras C. Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models [C]. In proceeding of Interspeech, Brisbane, 2008: 1505-1508.

[8] Becker T, Jessen M, Grigoras C. Speaker Verification Based on Formants Using Gaussian Mixture Models[C]. In proceeding of NAG/DAGA International Conference on Acoustics, Rotterdam, 2009.

[9] Moos A. Long-Term Formant Distribution (LTF) based on German spontaneous and read speech [C]. In proceeding of IAFPA, Lausanne, 2008: 5-6.

[10] Moos A. Forensische Sprechererkennung mit der Messmethode LTF (long-term formant distribution) [D]. MA thesis, Universitӓt des Saarlandes. 2008.

[11] The WaveSurfer software. (2011-12-31). http://www.speech.kth.se/wavesurfer/.

[12] Ladefoged, P. Elements of Acoustic Phonetics, 2nd ed [M]. University of Chicago Press, Chicago, 1996.

[13] Catalina Forensic Audio Toolbox. (2011-12-31). http://www.forensicav.ro/download/CatalinaManual3h.pdf

[14] Jessen M.（曹洪林, 王英利译）. 法庭语音学 [J]. 证据科学, 2010, 6: 712-738.

昆曲老生和小生的语音和嗓音声学初探

董理

摘要

本文以昆曲为研究对象，分析同一个人用老生唱法演唱、用小生唱法演唱和按昆曲字音朗读[1]这三种形式的语音和嗓音差异。语音方面，本文计算了三种形式的长时平均谱和共振峰，主要得出以下结论：1，昆曲老生和小生唱法不存在歌唱共振峰，而存在演讲者共振峰；2，昆曲演唱的发音比朗读的发音更中立化，昆曲演唱的口腔位置接近，与朗读的差异显著；3，朗读时声道长度最短，老生次之，小生声道长度最长。嗓音方面，本文提取了EGG信号的基频、开商和速度商，比较了三种发音形式下三个参数的差异，并计算了三个参数的相关性，主要得出以下结论：1，同一个人朗读、老生唱法和小生唱法的基频、开商和速度商在声学空间内的分布区域不同；2，朗读的基频最低，老生唱法次之，小生唱法基频最高；3，老生唱法的开商和速度商都大于朗读，可推测老生唱法使用紧的气嗓音；4，小生唱法的开商大于朗读，而速度商小于朗读，可推测小生唱法使用松的气嗓音；5，在三维空间内，朗读的分布最集中，小生和老生相对分散，小生是由于基频范围广，老生则因为开商、速度商和基频的范围都广；6，老生的基频与速度商呈显著线性负相关，基频与开商、开商与速度商呈中度线性负相关；小生的基频与速度商呈显著线性负相关，基频与开商呈中度线性负相关；朗读的基频与开商呈低度线性正相关。

关键词：

昆曲 老生 小生 朗读 语音 嗓音

引言

近年来对昆曲唱法的研究多已经从主观描述渐渐向使用科学术语描述过渡。如刘海燕的《试论昆曲<牡丹亭.游园>闺门旦的演唱艺术》介绍了闺门旦的发音用嗓特点，认为其不仅用假声，同时还有混合声和真声参与。作者还探讨了凤音、云音、鬼音等不同发音的共鸣腔特点和共鸣位置，以及口形对呼吸及咬字的重要性。该文虽使用了看似科学的术语，但仍是评演唱人经验及主观判断来进行描述的。于善英和池万刚的《京昆艺术嗓音声音形态分析与研究》对京剧和昆曲演员声音的频谱进行了分析，认为好的京剧和昆曲演员的声谱规则，而不好的则杂乱无章，同时认为京剧和昆曲各行当中都存在歌唱共振峰。该文章虽进行了频谱分析，但并未提及分析方法，所用术语也非常模糊，结论并不可靠。Sundberg 等在"Acoustical Study of Classical Peking Opera Singing"一文探讨了京剧中老生和花脸的长时平均谱，发现二者并没有歌唱共振峰。这与于善英等的研究结果相矛盾，所以在昆曲各行当是否存在歌唱共振峰的问题上要再做研究。焦磊在《昆曲演唱的发声调音研究——文献的实验语音学解释》中也对昆曲进行了频谱分析，发现小生、老生中也存在歌唱共振峰[2]，但其仅对单个元音、单一时间片段数据进行分析，不能证明歌唱共振峰确实存在。焦磊在文中还探讨了小生“小膛调”的嗓音特点，指出不存在一个介于真、假声之间的发音状态，其听感效果是由喉下气压增大和声道共鸣共同作用产生的。这可以说是开了昆曲实验语音学研究的先河，其研究出发点非常值得肯定。

虽说“隔行如隔山”，部分杰出的昆曲曲友可以改变其发声及共鸣方式，来完成多个行当的演唱。使用同一个人的声音材料来研究不同行当之间的嗓音和共鸣差异，可以排除不同被试之间的生理条件差异，进而得到更准确的实验数据。本文正是利用同一个人使用老生唱法、小生唱法演唱及按昆曲字音朗读的语音和嗓音数据，来进行实验研究的。本文分别对语音信号和嗓音信号进行了分析。第二章对语音信号进行了分析，分为两部分：第一部分对三种发音形式[3]进行了长时平均谱的研究，探讨不同情况下的谱特征；第二部分比较了单元音的共振

[1] 并非朗诵，而是如正常说话一般将材料念出来

[2] 原文表述为“额外的共振峰”，是引用Sundberg的描述，事实上就指代歌唱共振峰。

[3] 这里暂时使用“三种发音形式”来指代老生唱法、小生唱法和正常说话，其中的“发音形式”并不代表共鸣腔类型或嗓音类型。

峰差异。第三章对嗓音信号进行了分析，分为两部分：第一部分比较了三种发音形式的基频、开商、速度商的大小及分布情况；第二部分对基频、开商、速度商的相关性进行了分析。

1 实验说明

1.1 录音及分析软件

本次实验的录音设备是 Powerlab 肌电脑电仪及配套软件 ADInstruments 公司的 Powerlab Chart5 for Windows。本次实验采集了四路信号：第一通道是声音信号，第二通道是电子声门仪（EGG）信号 ，第三和第四通道是分别通过两条 MLT1132 呼吸带传感器采集的胸呼吸和腹呼吸信号，采样频率是 20kHz。本文只使用了前两路信号。

语音信号的分析使用的是 KTH 开发的语音分析软件 WaveSurfer-1.8.8p3，其使用 snack 技术来提取共振峰。嗓音信号的分析使用的是叶泽华编写的嗓音分析软件 voicelab，其以声门闭合点为周期标志，用尺度法找出声门闭合点，尺度为 0.25。统计分析使用的是 SPSS 16.0。

1.2 发音人及录音材料

发音人是一位 34 岁的男性，1998 年开始接触昆曲，2005 年开始学唱，先后师从北京曲家朱复和楼宇烈两位先生学唱，能唱老生和小生两个行当。

本次实验录制了老生曲目《长生殿・弹词》的《一枝花》，要求被试用老生唱法和小生唱法各演唱一遍。老生是按照原调演唱，小生则让被试适当提高声调演唱，以充分发挥小生演唱特性。最后，被试再按照昆曲的字音朗读一遍曲词。

2 语音信号分析

2.1 长时平均谱

长时平均谱是一种常用而有效的分析连续语流或歌唱的工具。通常截取 30-40s 的语音数据，用其声压平均值来再现其频谱特性。由于演唱的时间较长，本文将小生唱法和老生唱法的录音分成四段求其长时平均谱，然后再将这四段进行平均，得出小生和老生的均值，最后与朗读的长时平均谱进行比较。在计算长时平均谱之前，笔者已将辅音和空白段全部切出。语音的采样频率为 20kHz，分析带宽为 375Hz，FFT 点数为 128 点，分析频率域为 0-6000Hz。图 1 为三种发音形式的长时平均谱。表 1 给出了三个长时平均谱的Pearson 双尾检验结果，显著水平为 0.01，相关临界值为 0.403。

图 1 三种发音形式的长时平均谱

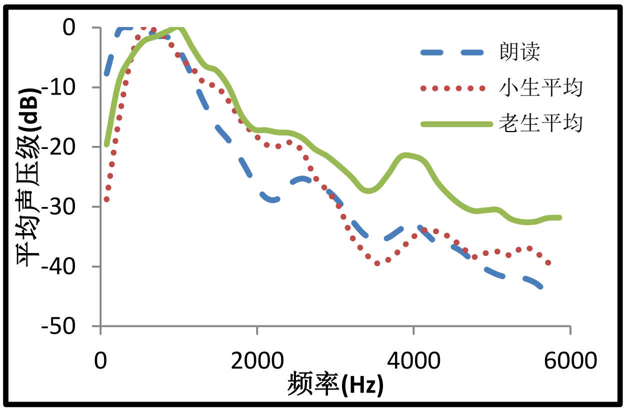


表 1 三种长时平均谱之间的相关性

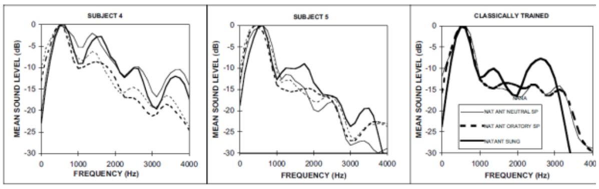
	朗读 vs. 小生	朗读 vs. 老生	小生 vs. 老生
r 值	0.93577	0.936779	0.95986

从图 1 中可以看出，老生的最高峰值对应频率最大，在 1000Hz 附近，小生的最高峰值对应频率居中，在 500Hz 左右，朗读的最小，在 400Hz 左右。从频谱走势上看，在 2500Hz 以下，小生和老生的谱斜率绝对值要小于朗读，即朗读的能量随着频率的提高要下降得更快。在 2500Hz 附近，朗读和小生的谱曲线出现一个小的峰值。从 2500Hz 到 3500Hz，小生的谱斜率绝对值忽然增大，能量迅速下降。在 4000Hz 附近，三条曲线都有一个小的峰值，其中老生的峰值最明显。从 4000Hz 到 6000Hz，小生的谱斜率绝对值又变小，曲线与朗读曲线逐渐分开。从整体来看，最高峰值过后，老生的声压级一直大于小生和朗读，小生的声压级与朗读相互交错，先大、后小、再大。从表 1 中可以看到，朗读、小生和老生两两之间的相关系数都大于 0.9，属于高度相关，其中，小生与老生之间的相关性要略大一些。

那么，是否可以下结论说昆曲老生和小生唱法不存在歌唱共振峰呢？参考一下 Sundberg 对乡村歌手的说话和唱歌的长时平均谱的分析。图 2 是乡村歌手与古典歌手的长时平均谱比较，其中实线为演唱国歌和其他歌曲的数据，虚线为朗读相同内容的数据。前两张图为两个乡村歌手的数据，第三张图为一个古典歌手的数据。可以看出，古典歌手的谱曲线在 2000-3000Hz 之间存在明显的歌唱共振峰，

而乡村歌手的谱中则没有此共振峰，但也和本文得出的数据一样，在 4000Hz 附近有一个小的峰值——演讲者共振峰[4]，这类共振峰同样存在于演员和电台播音员的语流中。通过与 Sundberg 的文章所得数据的比较，可以得出结论：昆曲老生和小生演唱时不存在歌唱共振峰，而存在演讲者共振峰。

图 2 乡村歌手与古典歌手的长时平均谱比较[5]

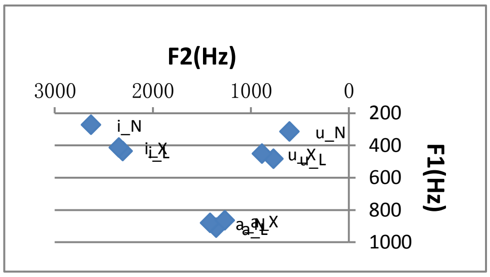


通过本小节的研究，可以得出以下结论：1，在 3000Hz 以下区域，昆曲演唱的频谱能量要大于朗读，小生与老生的能量相近，在 3000Hz 以上，小生能量衰减，居于老生和朗读之间；2，昆曲老生和小生唱法不存在歌唱共振峰，而存在演讲者共振峰。

2.2 共振峰

本文对三种发音形式的录音中所有以单元音/a/,/i/,/u/为韵母的音进行了比较。在昆曲演唱中，常出现声多字少的情况，即一个字要唱几个音。本文将同一个字按照基频稳定段的多少切成相应的几个片段，以减小基频对共振峰的影响。对每个语音片段，选取 0.1s 的基频稳定段（包含 10 个采样点的共振峰数据），提取出前三条共振峰的均值。图 3 是三种发音形式的声学元音图，图中点右侧的标记下划线前表示元音，下划线后的 N、X 和 L 分别代表朗读、小生和老生。图 3 中的数据值是由全部同类音节的共振峰数据平均而来的。

图 3 三种发音形式的声学元音图



从图 3 中可以看出，小生唱法和老生唱法的数据比较接近，二者与朗读的距离较远。整体来看，昆曲演唱中的元音位置要更集中，更中性化，尤其是/i/和/u/：演唱中的/i/比朗读中的/i/开口度更大，发音位置更靠后；演唱中的/u/比朗读中的/u/开口度更大，发音位置更靠前。F2 不仅关系到发音位置的前后，还与唇形面积有关，面积缩小，F2 就会降低。昆曲演唱时，/i/和/a/的第二共振峰要小于朗读，/u/的大于朗读，可以推测在发展唇元音时，昆曲的口型更圆，在发圆唇元音时，昆曲的口型更展。比较老生唱法和小生唱法的数据可以发现，/i/和/u/在小生唱法中要比老生唱法中要偏上偏前、唇形面积更大，/a/的小生唱法则比老生唱法要偏上偏后、唇形面积更小。

本文还对三种发音形式的三个元音分别作了 T 检验，用以比较其差异的显著性，显著水平均为 0.05。由于老生和小生的数据个数、基频条件一致，选用配对样本的 T 检验来检测，结果发现，只有/i/的 F3 有显著差异。比较老生与朗读、小生与朗读之间的差异，选用独立样本的 T 检验来检测，结果发现：二者与朗读之间，/i/的 F1、F2 和 F3、/u/的 F1 有显著差异。

下面仅就具备显著差异的数据进行讨论，主要参考鲍怀翘先生的研究[6]。/i/的收紧点在口腔前部，前腔小，后腔大。老生/i/的 F3 大于小生，所以老生发/i/时，前腔小于小生，后腔有小于小生的趋势。于是，可以推测老生的/i/舌头更紧张，更靠近硬腭。老生和小生的/i/的 F1 都大于朗读，F2、F3 都小于朗读。朗读时/i/的前腔小于演唱时的，后腔则大于演唱时的，可以推测朗读时整个舌头更靠前。/u/的收紧点在口腔后部，前腔大，后腔小。老生和小生的 F1 大于朗读。F1 随后腔的缩小而上升，因此，演唱时的后腔要比朗读时小，。对比声学元音图，演唱时舌位偏下偏前，就是说，演唱时，舌头前伸（与后缩相对）了，后腔的面积反而减少了。一个可能的元音是演唱时舌根部位紧张，更靠向后咽壁，使得后腔面积减少。

老生、小生和朗读的/i/的 F3 都存在显著差异，这也非常值得关注，其数值朗读最大，老生次之，小生最小。F3 的大小与声道长度成反比，所以，朗读时声道长度最短，老生次之，小生最长。这与演唱经验相吻合，使用老生唱法时要压喉，所以声道长度更长。没有受过声乐训练的人在喉头随着基频

[4] 原文为“speakers formant”

[5] 引自 “Long-Term-Average Spectrum Characteristics of Country Singers During Speaking and Singing”

[6] 见于《实验语音学概要》第五章

的升高而升高，所以低沉的声音对应的喉头位置更低，声道长度更长。这里说的低沉包括音高和音质两方面，喉头的降低会使元音音质变得低沉。老生为老年男性的声音，理应更低沉，喉头的降低正好恰当的表现了其艺术原型特征。低沉对应着尖细，按理来说，小生应该喉头提升，但其喉头并未提升，反而下降。这与美声唱法是一致的，音高越高，越不能提高喉头，缩小喉咽距离，越应该降低喉头。美声唱法与正常说话的喉头运动方向恰好相反。这就提示我们，小生的嗓音特点与说话是不同的。

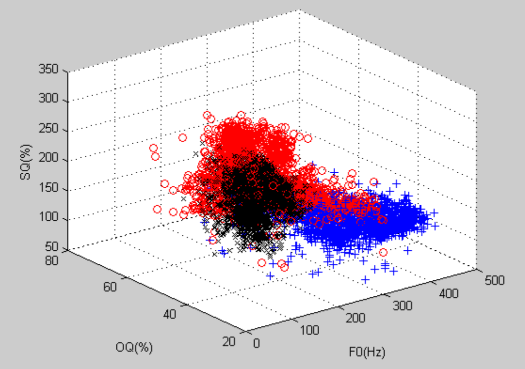
通过本小节的研究，可以得出以下结论：1，昆曲演唱的发音比朗读的发音更中立化；2，昆曲演唱的口腔位置接近，与朗读的差异显著；3，朗读时声道长度最短，老生次之，小生声道长度最长。

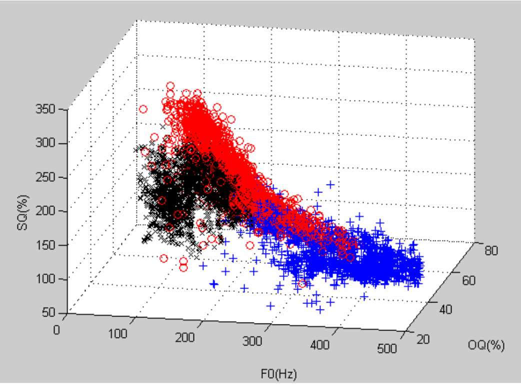
3 嗓音信号分析

3.1 基频、开商、速度商比较

本文分别提取了三种发音形式的 EGG 信号的基频、开商和速度商，并用 MATLAB 程序将所有数据画在了由基频、开商和速度商组成的声学空间图当中，如图 4 所示。图 4 中是同一张三维图从两个角度看得结果，其中红色圆圈开标老生数据，蓝色加号代表小生数据，黑色乘号代表朗读数据。

图 4 声学空间图





从图中可以看出，三种发音形式的分布所占区域不同，形状也不同。

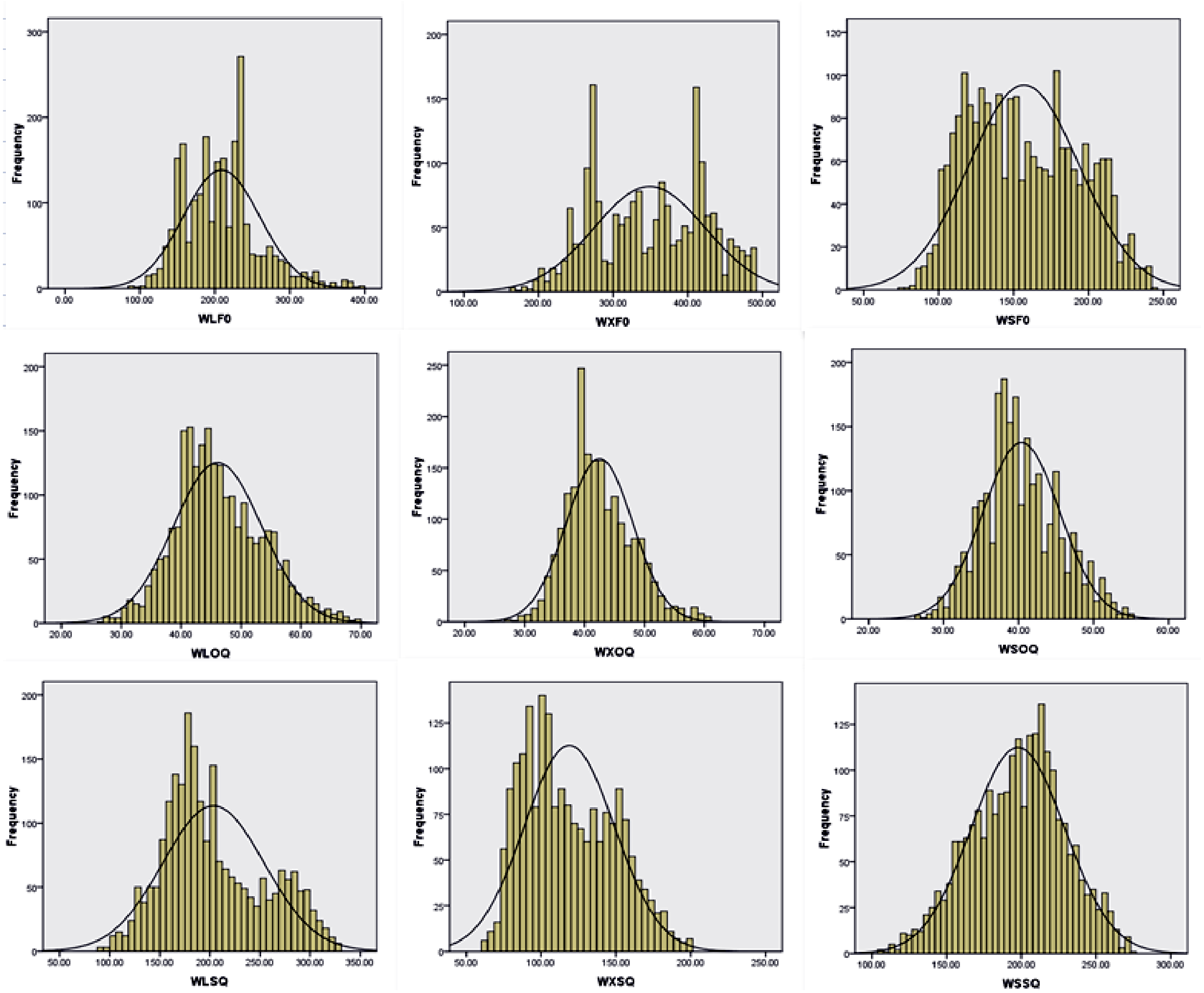
先看分布位置：朗读的基频最低，老生次之，小生基频最高；朗读的开商最小，小生次之，老生开商最大；小生的速度商最小，朗读次之，老生的速度商最大。速度商大意味着声带紧。开商大意味着漏气多。据此，如果假定朗读使用的是正常嗓音，就可以做出以下推测：老生的开商和速度商都大于朗读，可推测老生使用紧的气嗓音；小生的开商大于朗读，而速度商小于朗读，可推测小生使用松的气嗓音。根据《论语言发声》的研究，男性的基频小于女性，而速度商大于女性，二者的开商差异不大。这与老生和小生的差异非常相似。

再看数据分布形状：朗读接近球形，老生和小生则接近扁片。这就暗示着，老生和小生的各参数之间存在线性关系。老生的数据看似覆盖在朗读和小生的数据上，小生的数据似乎是老生在基频方向的延续。朗读的分布最集中，小生和老生相对分散。究其原因，小生是由于基频范围广，老生则因为开商、速度商和基频的范围都广。基频、开商和速度商的分布情况见图 5。图 5 中每行代表一个参数，第一行为基频，第二行为开商，第三行为速度商；每列代表一种发音形式，第一列为老生，第二列为小生，第三列为朗读。表 2 为分布图所对应的均值和标准差数据。

表 2 基频、开商和速度商的均值和方差

	Mean	Std. Dev		Mean	Std. Dev		Mean	Std. Dev
WLF0	208.88	51.031	**WLOQ**	46.06	7.337	**WLSQ**	203.53	50.588
WXF0	348.92	74.604	**WXOQ**	42.49	5.539	**WXSQ**	118.82	30.163
WSF0	156.98	36.966	**WSOQ**	40.37	5.132	**WSSQ**	197.71	31.71

图 5 基频、开商和速度商的分布图



对照图 5 和表 2 可以发现，朗读的数据方差最小，数据最集中。小生的基频方差最大，其基频分布最广。老生的开商和速度商的方差都是最大的，基频的方差也大于朗读的，其在三个维度上的分布都很广。数据中反映的与三维图中观察到的相一致。

通过本小节的研究，可以得出以下结论：1，在声学空间图中，老生、小生和朗读的数据分布区域不同；2，根据老生、小生与朗读之间三个参数的大小关系，可以推测出，老生唱法使用的是紧的气嗓音，小生唱法使用的是送的气嗓音；3，朗读的数据分布集中，老生和小生的数据相对分散。

3.2 基频、开商、速度商的相关性

本节对基频、开商、速度商两两之间的相关性进行了讨论。实验中每一组有 2000 多个数据，由于数据量太大会使相关系数过小，所以在计算之前，先就数据进行局部平均。如在计算基频和开商的相关系数时，将基频从大到小排序，开商也随之排序，然后每十个数据取一个平均值。最终对每组剩下的 200 多数据进行 Pearson 双尾检验，相关系数如表 3，显著水平为 0.01，相关临界值为 0.182。

表 3 基频、开商、速度商两两之间的相关系数

	F0&OQ	F0&SQ	OQ&SQ
老生	-0.4697	-0.90232	-0.56809
小生	-0.54088	-0.81062	-0.02785
朗读	0.256154	-0.04791	0.036869

从表中可以看出：老生的基频与速度商呈显著线性负相关，基频与开商、开商与速度商呈中度线性负相关；小生的基频与速度商呈显著线性负相关，基频与开商呈中度线性负相关；朗读的基频与开商呈低度线性正相关。这与上一小节中通过观察分布的形状得出的结论是一致的，即老生和小生的各参数之间存在线性关系。在相关性

这一点上，老生与小生的结论更一致，与朗读的结论相距甚远。

但这与以往的研究似乎相矛盾。《论语言发声》中基频、开商和速度商的关系为：速度商与音调高低成反比，开商与音调高低成反比。这与老生和小生的结论相一致，而最应该一致的朗读却没有这种相关关系。一种可能的解释是，在不同的基频范围内，开商和速度商的走势是不同的，如果选取的基频段内，恰好开商和速度商都随基频上升而下降，那么相关关系明显；如果恰好截取了开商或速度商转折的位置，则没有相关关系。笔者将三组数据排在一起，以增加基频的范围，结果发现基频和开商不存在线性关系了。这增强了这一解释的可能性。

通过本小节的研究，可以得出以下结论：1，老生的三个参数之间两两负相关；2，小生的基频与速度商、基频与开商负相关；3，朗读的基频与开商正相关。

4. 结论和讨论

本文的研究只选用了一个被试，因此不能反映出不同被试之间的共性与个性。尽管如此，由于被试得到广泛的认可，其数据结果是有一定普遍价值的。

本文对昆曲老生唱法、小生唱法和按昆曲字音朗读三种发音形式的语音和嗓音信号进行了多角度的分析，得到了颇多结论。语音方面：1，昆曲老生和小生唱法不存在歌唱共振峰，而存在演讲者共振峰；2，昆曲演唱的发音比朗读的发音更中立化，昆曲演唱的口腔位置接近，与朗读的差异显著；3，朗读时声道长度最短，老生次之，小生声道长度最长。嗓音方面：1，朗读、老生唱法和小生唱法的基频、开商和速度商在声学空间内的分布区域不同；2，朗读的基频最低，老生唱法次之，小生唱法基频最高；3，老生唱法的开商和速度商都大于朗读，可推测老生唱法使用紧的气嗓音；4，小生唱法的开商大于朗读，而速度商小于朗读，可推测小生唱法使用松的气嗓音；5，在三维空间内，朗读的分布最集中，小生和老生相对分散，小生是由于基频范围广，老生则因为开商、速度商和基频的范围都广；6，老生的基频与速度商呈显著线性负相关，基频与开商、开商与速度商呈中度线性负相关；小生的基频与速度商呈显著线性负相关，基频与开商呈中度线性负相关；朗读的基频与开商呈低度线性正相关。

综合以上数据发现，在用小生唱法演唱时，声带比较放松，喉头的升降模式与美声唱法一致，即音调上升，喉头下降；在用老生唱法演唱时，声带比较紧张，喉头的升降模式与说话一致，即音调下降，喉头下降。虽然优秀的昆曲演员的声音听起来都比较集中，有“亮心”，但小生和老生都不存在歌唱共振峰。根据 Sundberg 的研究，歌唱共振峰的存在使的美声独唱者的声音可以穿透巨大的交响乐伴奏，进而让观众听清。其他不需要达此目的的西方唱法则不存在歌唱共振峰。声音有“亮心”并不意味着一定存在歌唱共振峰，因为优秀的播音员的音质也非常集中，其所存在的演讲者共振峰可以达到同样的效果。比较昆曲和美声唱法的咬字可以发现，前者非常准确，而后者则非常模糊，常常听不出唱的元音到底是什么，这也是歌唱共振峰的一个副作用。演讲者共振峰并不影响元音的感知。这从另一个角度辅证了老生和小生不存在歌唱共振峰。

回顾声学空间图，每种发音形式对应的区域都是三维的，也就是说，在两个参数值相同的情况下，第三个参数不是固定的，有多种可能性。这就告诉我们，昆曲的演唱不是死板的、一成不变的，可以在大规律下进行小的自由发挥。不同人的演唱处理得不同，这增加了听曲的乐趣。

5. 参考文献

[1] 刘海燕 《试论昆曲<牡丹亭.游园>闺门旦的演唱艺术》中国音乐学 2006 年第 2 期

[2] 于善英、池万刚 《京昆艺术嗓音声音形态分析与研究》音乐探索 2009 年第 2 期

[3] Johan Sundberg, Lide Gu, Qiang Huang, and Ping Huang “Acoustical Study of Classical Peking Opera Singing” Journal of Voice 2011 May 26

[4] 焦磊 陈春苗《昆曲演唱的发声调音研究——文献的实验语音学解释》第六届国际吴方言学术研讨会 2010 年 11 月

[5] Thomas F. Cleveland, Johan Sundberg, R. E. (Ed) Stone “Long-Term-Average Spectrum Characteristics of Country Singers During Speaking and Singing” Journal of Voice Volume 15, Issue 1, March 2001, Pages 54-60

[6] 李绍山 《语言研究中的统计学》 2001 年 西安交通大学出版社

[7] 吴宗济、林茂灿 《实验语音学概要》 1989 年 高等教育出版社

[8] Sundberg J. “The science of the singing voice” 1987 Northern Illinois University Press

[9] 孔江平《论语言发声》 2001 年 中央民族大学出版社

普通话双音节 V1#C2V2 音节间的逆向协同发音

李英浩[1, 2]，*孔江平*[1,3]

(1.北京大学 中国语言文学系，北京 100871；2.延边大学 外国语学院，延吉 133001；
3.北京大学 中国语言学研究中心，北京 100871)

摘 要

本文研究普通话双音节 V1#C2V2 中 C2 和 V2 对 V1 后过渡段的腭位以及 F2 轨迹的方向和变化幅度的影响。分别使用动态电子腭位和线性预测编码的方法获得 V1 后过渡段的腭位参数和 F2 轨迹。实验结果发现：1) V1 的 F2 轨迹的变化与后腭接触面积的变化显著相关。2) C2 为不同发音方式的唇音、舌尖中音和舌面后音时对 V1 的腭位和/或 F2 轨迹有显著影响。3)元音间逆向协同发音有 2 种情况：C2 为唇音、舌尖中音和舌面后音的条件下，V2 显著影响 V1 的腭位和/或 F2 轨迹；C2 为其他发音部位的条件下，V2 的圆唇特征显著影响 V1 的 F2 轨迹。实验结果表明：C2 对 V1 以及 V2 对 V1 的影响受到 C2 舌体发音限制条件的制约，但是 V2 的圆唇特征对 V1 的影响不受 C2 发音限制条件的制约。

关键词：

动态电子腭位；逆向协同发音；双音节

协同发音不仅是言语产出研究的核心内容，也是语音合成和识别中的难点问题。由于发音器官动作的连续性以及不同发音器官的运动相位关系受多种因素的影响，语流中音段的发音动作往往偏离目标动作，从而导致音段的声学参量存在变异现象[1]。

已有的发音生理和声学研究结果表明：逆向协同发音在发音动作编码过程中就已经实现[2]，因此音段间的逆向协同发音受规则支配。对普通话双音节 V1#C2V2 音联的声学研究表明，C2 和 V2 对 V1 的声学目标以及 V1 的 *F*2 轨迹都有影响[3]。相同发音部位、不同发音方式的C2对V1的*F*2轨迹影响一致，V2 对 V1 的影响只出现在 C2 为唇音的条件下[4]。基于电磁发音仪(EMMA)的研究结果[5]发现：V2 对 V1 舌位动作的影响不能超越 C2(该研究中的 C2 为齿龈和软腭塞音)。普通话二维唇形研究结果[6]发现：V2 的圆唇动作从 V1 的后过渡段就开始启动(V1≠圆唇元音)，因此 V2 圆唇动作有可能影响 V1 后过渡段的 *F*2 轨迹。

本文使用 EPG 分析普通话双音节 V1#C2V2 中 V1 后过渡段的腭位变化，同时考察对应时段的*F*2轨迹，研究 C2 发音部位、发音方式以及 V2 对 V1 后过渡段腭位和 *F*2 轨迹的变化方向、幅度的影响。

1. 语料和标注

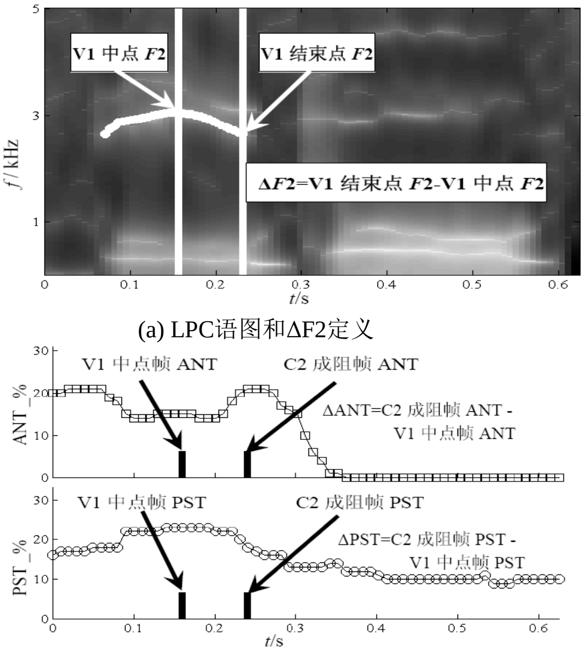
语料来自北京大学语音学实验室的普通话动态电子腭位数据库中女性发音人的982个双音节样本。大多数双音节样本首字声母为不送气舌尖中音/t/；V1 为单元音/i, a, u, i1, i2/(i1 为舌尖前元音，i2 为舌尖后元音)；C2 为 21 个声母，发音部位包括唇音(LA)、舌尖前音(AP)、舌尖中音(AL)、舌尖后音(RE)、舌面前音(PA)和舌面后音(VA)，发音方式包括塞音(S)、塞擦音(A)、擦音(F)、鼻音(N)和边音(L)；V2 包括/i, a, u, i1, i2, ei, y/。C2 和 V2 的组合受音位配列制约。C2 为 PA 的条件下，加入复合元音/iu, ia/考察/u, a/对 V1 的逆向协同发音影响。

双音节用正常语速朗读 2~3 遍。样本信号包括 62 电极的腭位信号(采样频率为 100 Hz)，喉头仪信号和声学信号(采样频率为 22 kHz)。V1 的声学起点由 EGG 信号自动标注，声学结束点结合语图和 EPG 信号手动标注。V1 的 *F*2 轨迹先由 PRAAT 批量获得，然后在基于 Matlab 的普通话腭位分析系统中依据 LPC(linear prediction coding)语图(见图 1a)中共振峰的走向手动校准。

2. 方法

把假腭分为前腭和后腭 2 个区域，分别计算这 2 个腭区的接触面积。前腭接触面积(ANT)和后腭接触面积(PST)分别为假腭前和后 4 行电极中接触电极数目与这个区域电极总数的比值。这 2 个腭位参数能

反映舌前部和舌体运动的过程。本文分析 V1 声学结束点后一帧(C2 成阻帧)与 V1 中点帧的腭位参数的差，即 ΔANT 和 ΔPST。V1 后过渡段的 *F*2 轨迹变化为 V1 结束点 *F*2 和 V1 中点 *F*2 的差，即 Δ*F*2。Δ*F*2 的符号表示 *F*2 轨迹的方向，其绝对值表示 *F*2 轨迹的变化幅度。3 个分析参数的定义方法如图 1 所示。



(a) LPC语图和ΔF2定义

(b) ΔANT 和 ΔPST 定义

图 1 声学参数和腭位参数定义图示

3. 结果

3.1 C2 发音部位对 V1 的逆向协同发音影响

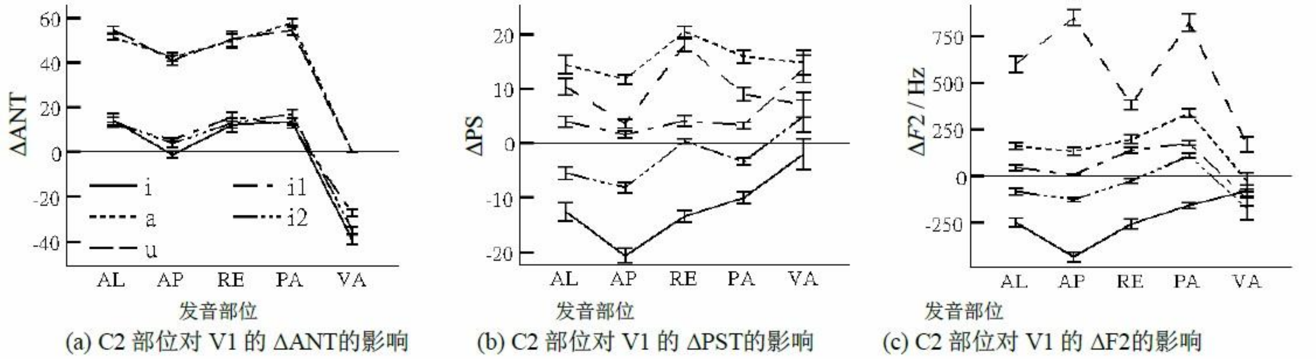
已有研究结果发现:C2 发音部位为 LA 的条件下，V1 后过渡段 F2 轨迹只受 V2 的影响[3-4]。因此，把 C2 为 LA 的情况放在节 3.3 讨论。

图 2 是 5 个发音部位的 C2 对 V1 腭位参数和 F2 轨迹影响的示意图。从图 2a 可知，V1 的 ΔANT 分为 2 种情况：V1/a, u/的 ΔANT 相对一致，C2 发音部位为 VA 的条件下 ANT 基本无变化;C2 发音部位为其他的条件下，ANT 的增幅超过 40%，这与 C2 在齿龈或硬腭前区域的成阻动作相关。V1/i, i1, i2/的 ΔANT 比较相似，C2 发音部位为 AL、RE 和 PA 的条件下，ANT 的增幅一般低于 20%；C2 发音部位为 AP 条件下舌尖元音的 ANT 增幅较小，前高元音的 ANT 基本不变;C2 发音部位为VA的条件下,ANT均下降。上述结果说明：C2 发音部位为 AL、AP、RE 和 PA 的条件下,V1 后过渡段舌前部的动作向 C2 发音目标接近,ANT 的增加与 V1 和 C2 在舌前部的发音目标的距离相关；C2 发音部位为 VA 的条件下，由于 V1/a, u/发音目标的 ANT 为零,因此 ΔANT 也为零,V1/i, i1, i2/在后过渡段向软腭成阻的发音部位过渡，到 V1 声学结束点之后 ANT 降为零。

图 2b 在一定程度上反映了 V1 后过渡段舌体运动的特征。V1/i/在 5 种 C2 发音部位条件下的 PST 均下降，PST 下降的幅度与 C2 对舌体位置和舌形的要求有关。如 C2 发音部位为 AP 的条件下，V1/i/的 PST 降幅最大,这与 AP 成阻过程中舌体靠后[7]、舌面下降成槽的动作有关；而 C2 发音部位为 RE 的条件下，虽然舌面也有成槽的动作，但是舌面下降的幅度要小于 C2 发音部位为 AP 时的情况。V1/i2/的舌体动作过渡有 3 种情况:1)C2 发音部位为 RE 的条件下的 PST 基本无变化,2)C2 发音部位为 VA 的条件下 PST 上升，3)C2 发音部位为其他的条件下的 PST 均下降。3 种情况下 PST 的变化反映了舌体位置和舌形姿态的调整，限于篇幅，此处不做详述。V1/i1/后接C2发音部位为VA的条件下PST的升幅超过10%,但在其他的条件下 ΔPST 的升幅不超过 10%,前者与软腭成阻有关，后者与 C2 对舌体动作要求有关。V1/u/后过渡段的舌体动作过渡同样也受到 C2 对舌体发音动作要求的影响。C2 发音部位为 AP 的条件下 PST 增幅最小，而 C2 发音部位为 RE 的条件下 PST 增幅最大，C2 发音部位为其他的条件下 PST 的增幅在前两者之间。V1/a/后过渡段的舌体抬高的幅度整体上大于其他 V1 的情况，PST 的增幅同样也反映了 C2 对舌体动作的发音要求。

从上面的结果来看,V1 向 C2 过渡过程中舌头的动作分为 2 种情况:当 C2 发音部位不是 VA 的时候，舌尖或舌页的成阻部位位于硬腭之前的区域，舌体受 C2 发音条件的制约做主动或者耦合的运动;当 C2 发音部位为 VA 的时候，舌体作为主发音器官在硬腭后/软腭前成阻，舌前部动作受舌体动作的制约。

表 1 表示以 C2 发音部位不是 VA 条件下的 2 个腭位参数为自变量、ΔF2 为因变量的回归分析结果(使用强制选入法)。从表 1 可知：所有回归模型对 ΔF2 的解释程度均小于 20%；V1/i, a, u/的 ΔF2 只与 ΔPST 显著相关，而与 ΔANT 无关。V1/a, u/的回归结果符合预期结果[2,8]。V1/u/出现了负相关的情况，根源在于 V1/u/后接舌尖音(包括 AP 和 RE)的条件下 ΔF2 与 ΔPST 并无对应关系，具体原因将另文处理。

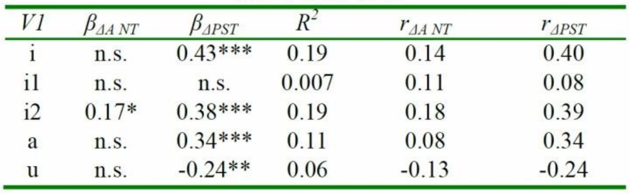


(a) C2 部位对 V1 的 ΔANT的影响 (b) C2 部位对 V1 的 ΔPST的影响 (c) C2 部位对 V1 的 ΔF2的影响

图 2 C2 发音部位对腭位参数和 F2 变化的影响

V1/i1/的 Δ*F*2 与 2 个腭位参数没有显著相关关系，从图 2c 可以看出，V1/i1/的 Δ*F*2 与 C2 对舌体动作的要求有关系，此处不展开说明。V1/i2/的 Δ*F*2 与 2 个腭位参数均显著相关，但从图 2 可知，Δ*F*2 与 ΔANT 的关系可能仅限于C2发音部位为PA的情况，此时，虽然 PST 下降，但是 ANT 的接触重心却比较靠后(与舌面前音在齿龈后成阻有关)，因此 V1/i2/的 *F*2 轨迹上升。

表 1 回归分析表



V1	$\beta_{\Delta ANT}$	$\beta_{\Delta PST}$	R^2	$r_{\Delta ANT}$	$r_{\Delta PST}$
i	n.s.	0.43***	0.19	0.14	0.40
i1	n.s.	n.s.	0.007	0.11	0.08
i2	0.17*	0.38***	0.19	0.18	0.39
a	n.s.	0.34***	0.11	0.08	0.34
u	n.s.	-0.24**	0.06	-0.13	-0.24

C2 发音部位为 VA 的条件下，V1/i, i1, i2/的 ANT 下降，V1/i/的 PST 小幅下降，而 V1/i1, i2/的 PST 上升。V1 后过渡段的 *F*2 轨迹一般下降，这与舌体后移的动作有关。V1/a, u/只有 PST 上升，前者的 Δ*F*2 接近零，而后者的 Δ*F*2 在 250 Hz 附近。C2 发音部位为 VA 的条件下的舌体运动与 *F*2 轨迹变化的关系比较复杂，一方面 C2 的舌体动作易于受两侧元音舌体动作的影响使得 C2 成阻部位发生变化，另一方面 *F*2 腔体归属会在软腭部位发生置换[9]。

3.2 C2 发音方式对 V1 的逆向协同发音影响

一系列的单因素方差分析结果表明：C2 发音部位为 PA、AP 和 RE 的条件下，C2 发音方式对 V1 的 ΔPST 和 Δ*F*2 没有显著影响，ΔANT 的显著差异只与舌前部与齿龈、硬腭成阻有关，此处不做分析。C2 发音方式对 V1 的腭位和 *F*2 轨迹的影响出现在 C2 发音部位为 LA、AL 和 VA 的条件下。

由 C2 发音部位为 LA 的条件下的 3 个参数的单因素方差分析结果可以发现：V1 为/a, i1, i2/条件下的 ΔPST 和 ΔF2 不受 C2 发音方式的影响；而当 V1 为/i, u/的时候，至少有 1 个参量受到 C2 发音方式的影响。V1 为/i/的时候，C2 发音部位为 LA 且发音方式为 S 和 F 的条件下 PST 的降幅显著大于发音方式为 N 的条件下的降幅。C2 发音方式为 F 的条件下的 *F*2 的降幅显著大于为 S 和 N 的条件下，这与 PST 的快速下降有关。V1 为/u/的时候，C2 发音方式为 F 的条件下的 *F*2 上升，而其他条件下的 *F*2 的变化趋近于零。这可能与 V1 后过渡段下唇内收有关。

C2 发音部位为 AL 的条件下，发音方式为 L 的 C2 对 V1 腭位参数和 *F*2 轨迹的影响与其他发音方式的 C2 有显著区别。单因素方差分析结果表明：C2 发音方式为 L 条件下 V1/i, i1, i2/的 ΔPST 的降幅显著大于 C2 为其他发音方式条件下的降幅。在声学方面，除了 V1/i1/以外，C2 发音方式为 L 的条件下 V1/i, i2/的 Δ*F*2 的降幅显著大于 C2 为其他发音方式条件下的降幅，同时 V1/a, u/的 Δ*F*2 的升幅显著小于 C2 为其他发音方式条件下的升幅。由于边音 L 发音时舌体两侧下降，因此前高元音向边音过渡时，PST 会迅速降低，而低元音和后高元音的过渡过程无法被 EPG 准确捕获。

发音部位为 VA 的 C2 对 V1(特别是高元音)后过渡段 PST 的变化有显著影响，发音方式为 S 的 C2 在软腭区域形成阻塞，因此 PST 总是增加。C2 发音方式为 F 的条件下 PST 总是减少。C2 发音方式为 F 的条件下 V1 的 Δ*F*2 降幅一般较大，这与 PST 的下降有关。但是 V1/u/的情况特殊，C2 发音方式为 F 的条件下 V1/u/的 *F*2 轨迹上升，且 *F*2 的升幅大于 C2 发音方式为 S 条件下的升幅，这可能与 V2 的影响有关。

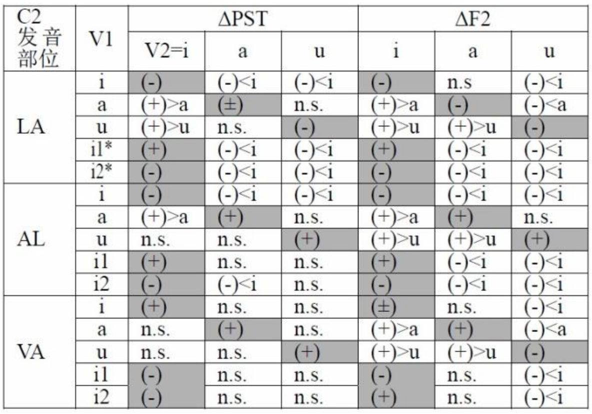
3.3 V2 对 V1 的逆向协同发音影响

按发音部位计算每个 V1 条件下不对称元音序列(V1≠V2)和对称元音序列(V1=V2)的 V1 后过渡段的 3 个参数是否具有显著差异。V1 为舌尖元音且 C2 不是舌尖辅音时，把/V1#C2i/作为比较的基准；C2 为 VA 时使用/i#C2ei/作为比较的基准。

单因素方差分析结果发现 C2 为 LA、AL 和 VA 的条件下，V2 对 V1 后过渡段的 *F*2 轨迹和/或腭位参数有显著影响，如表 2 所示。表 2 中的灰色单元格表示对称序列，括号内的符号表示参数的变化方向，(±)表示基本没有变化；其他单元格的结果表示参数的变化方向、非对称序列与对称序列的参数是否

存在显著差异以及差异的方向。如 V1=/i/、V2=/u/以及 C2=LA 条件下的 ΔF2 的结果(-)<i 表示 i#C2u 中 V1/i/的 F2 轨迹下降，且降幅显著大于 i#C2i 中 F2 的降幅。C2=LA、V1=/i1,i2/条件下的*表示 ΔPST 没有显著差异，但是 ΔANT 有显著差异。n.s.表示没有显著差异。

表 2 C2 为 LA、AL 和 VA 条件下 V2 对 V1 的影响



C2 发音部位	V1	ΔPST			ΔF2		
		V2=i	a	u	i	a	u
LA	i	(-)	(-)<i	(-)<i	(-)	n.s	(-)<i
	a	(+)>a	(±)	n.s.	(+)>a	(-)	(-)<a
	u	(+)>u	n.s.	(-)	(+)>u	(+)>u	(-)
	i1*	(+)	(-)<i	(-)<i	(+)	(-)<i	(-)<i
	i2*	(-)	(-)<i	(-)<i	(-)	(-)<i	(-)<i
AL	i	(-)	(-)<i	(-)<i	(-)	(-)<i	(-)<i
	a	(+)>a	(+)	n.s.	(+)>a	(+)	n.s.
	u	n.s.	n.s.	(+)	(+)>u	(+)>u	(+)
	i1	(+)	n.s.	n.s.	(+)	(-)<i	(-)<i
	i2	(-)	(-)<i	n.s.	(-)	(-)<i	(-)<i
VA	i	(+)	n.s.	n.s.	(±)	n.s.	(-)<i
	a	n.s.	(+)	n.s.	(+)>a	(+)	(-)<a
	u	n.s.	n.s.	(+)	(+)>u	(+)>u	(-)
	i1	(-)	n.s.	n.s.	(-)	n.s.	(-)<i
	i2	(-)	n.s.	n.s.	(+)	n.s.	(-)<i

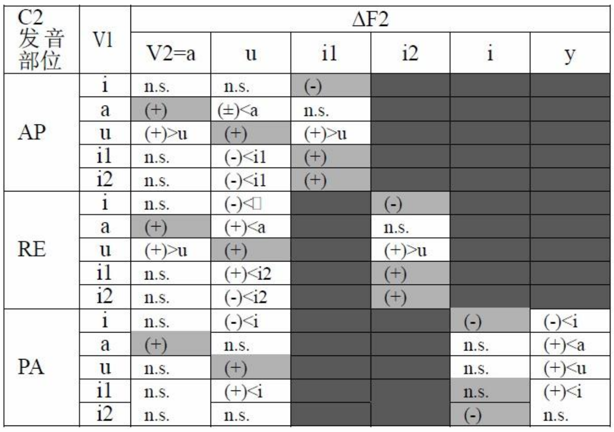
C2 为 LA 条件下 V2 对 V1 后过渡段的腭位参数有显著影响。V1/i, a, u/的 ΔPST 的方向和/或幅度受到 V2 的影响,V1/i1, i2/的 ΔPST 没有显著差异，但是 ΔANT 存在显著差异。ΔF2 受 V2 影响的方向和幅度与 ΔPST 或 ΔANT 相同。

C2 为 AL 条件下,V2 对 V1/i, a, u, i2/的 ΔPST 和 ΔF2 的影响表现为幅度而非方向的差异；但是 V1/i1/的影响却只表现为 ΔF2 的方向和幅度上的差异，ΔPST 不受 V2 的影响。

C2 为 VA 条件下，V2/ei, u/影响 V1/a, u/的 F2 轨迹的方向或变化幅度，而 V2/a/只影响 V1/u/的 F2 轨迹的方向。V1 后过渡段的腭位参数不受 V2 的影响，考虑到 EPG 只能部分地记录软腭成阻的过程，因此还需要结合其他手段来分析 C2 为 VA 条件下舌体的运动情况。

C2 为 AP、RE 和 PA 的条件下，V2 对 V1 的腭位参数没有显著影响，但是 V2 为圆唇元音的条件下 V1 后过渡段的 F2 轨迹的变化幅度有显著差异，如表 3 所示。表 3 中黑色单元格表示没有该种类型的双音节。

表 3 C2 为 AP、RE 或 PA 条件下 V2 对 V1 的影响



C2 发音部位	V1	ΔF2					
		V2=a	u	i1	i2	i	y
AP	i	n.s.	n.s.	(-)	■	■	■
	a	(+)	(±)<a	n.s.	■	■	■
	u	(+)>u	(+)	(+)>u	■	■	■
	i1	n.s.	(-)<i1	(+)	■	■	■
	i2	n.s.	(-)<i1	(+)	■	■	■
RE	i	n.s.	(-)<□	■	(-)	■	■
	a	(+)	(+)<a	■	n.s.	■	■
	u	(+)>u	(+)	■	(+)>u	■	■
	i1	n.s.	(+)<i2	■	(+)	■	■
	i2	n.s.	(-)<i2	■	(+)	■	■
PA	i	n.s.	(-)<i	■	■	(-)	(-)<i
	a	(+)	n.s.	■	■	n.s.	(+)<a
	u	n.s.	(+)	■	■	n.s.	(+)<u
	i1	n.s.	(+)<i	■	■	n.s.	(+)<i
	i2	n.s.	n.s.	■	■	(-)	n.s.

4 结论

通过分析普通话 V1#C2V2 双音节 V1 后过渡段的腭位参数、F2 轨迹变化的方向和幅度以及两者的关系，本文发现后字音节对 V1 的逆向协同发音与 C2 发音部位、C2 发音方式以及 V2 这 3 个因素有关。V1 后过渡段的舌前部（包括舌尖和舌页）和舌体运动受到 C2 发音部位的影响，同时 F2 轨迹与后腭接触面积的变化显著相关。C2 发音部位为 LA、AL 或 VA 的条件下，C2 发音方式影响 V1 后过渡的腭位和/或 F2 轨迹，这一方面与 C2 发音动作的成阻方式有关，另一方面与 V2 类型有关。V2 对 V1 的逆向协同发音表现为 2 种形式:1)C2 发音部位为 LA 或者 AL 的条件下，V2 不仅影响 V1 后过渡段的舌位，也影响其 F2 轨迹的方向、幅度；C2 发音部位为 VA 的条件下，V2 只影响 V1 后过渡段的 F2 轨迹的方向、幅度。由于软腭辅音的舌形姿态较易受到临近元音动作的影响，还需要通过其他的实验手段做进一步观察。2)如果 C2 对舌体动作限制较强，V2 就不会影响 V1 后过渡段的舌位，但是 V2 的圆唇特征不受 C2 舌体发音动作限制条件的制约。

参考文献

[1]. Lindblom B. Explaining phonetic variation: a sketch of the H and H theory[C]// Speech Production and Speech Modeling. Dordrecht: Kluwer, 1990: 403-439.

[2]. Recasens D, Pallares M D, Fontdevila J. A model of lingual coarticulation based on articulatory constraints[J]. *JASA*, 1997, **102**: 544-561.

[3]. WU Zongji, SUN Guohua. An experimental study of coarticulation of unaspirated stops in CVCV contexts in Standard Chinese[C]//Annual Report of Phonetic Research. Beijing: Institute of Linguistics of CASS, 1989: 1-25.

[4]. 陈肖霞. 普通话音段协同发音研究[J]. 中国语文. 1997,

5, 345-350.

[5]. CHEN Xiaoxia. Segmental coarticulation in standard Chinese[J]. *Zhong Guo Yu Wen*, 1997, **5**: 345-350.(in Chinese)

[6]. MA Liang, Perrier P, DANG Jianwu. A study of anticipatory coarticulation for French speakers and for Mandarin Chinese speakers[J]. *Chinese Journal of Phonetics*, 2009, **2**: 82-89.

[7]. 潘晓声. 汉语普通话唇形协同发音及可视语音感知研究[D].北京：北京大学，2011.
PAN Xiaosheng. Labial Coarticulation and Audio-Visual Speech Perception of the Standard Chinese[D]. Beijing: Peking University, 2011.(in Chinese)

[8]. 汪高武. 汉语普通话声道调音模型研究[D].北京：北京大学，2010.
WANG Gaowu. An Articulatory Model of Vocal Tract in Mandarin[D]. Beijing: Peking University, 2011. (in Chinese)

[9]. Iskarous K, Fowler C A, Whalen D H. Locus equations are an acoustic expression of articulator synergy[J]. *JASA*, 2010, **128**(4): 2021-2032.

[10]. Tabain M. Coarticulation in CV syllables: a comparison of Locus equation and EPG data[J]. *Journal of Phonetics*, 2000, **28**: 137-159.