语音乐律研究报告

Status Report of Phonetic and Music Research

2010

北京大学中文系语言学实验室

Linguistics Lab
Department of Chinese Language and Literature
Peking University

说　明

本论文集是北京大学中文系语言学实验室年度研究报告的第三期，收录了实验室 2010 年发表和未发表的论文 17 篇。2010 年，本实验室主要开展了如下活动：

一、 科研进展

实验室的研究主要包括以下几个方面：基础理论和方法、普通语音学、生理语音学、民族语音学、心理语音学、司法语音学等，详见本论文集内容。

二、学术交流

实验室与许多科研机构和大专院校建立了广泛的合作关系，如 Berkeley、JAIST、CNRS、台湾中央研究院、香港大学、香港中文大学、西北民族大学、广西大学等，并互派访问学生。同时，实验室开设的语音学课程，也接受了各地院校机构的学生选课或旁听。

2010 年，台湾中央研究院语言研究所特聘研究员、阳明大学神经科学研究所合聘教授曾志朗先生、伯克利加州大学荣休教授、香港中文大学伟伦研究教授王士元先生、美国麻省州立大学亚洲语言文学系的沈钟伟教授、英国伦敦大学许毅教授、中研院语言学研究所研究员、所长孙天心教授等来访，并做了相关的研究报告。

实验室的大部分人员参加了 2010 年 10 月在北京举行的“四川境内藏缅语国际研讨会”。

三、田野调查

2010 年 6 月，孔江平教授与陈保亚教授等北京大学茶马古道语言文化研究课题组和云南大学茶马古道文化研究所组织了茶马古道 20 周年纪念考察。

北京大学中文系语言学实验室
Linguistics Lab, Department of Chinese Language and Literature, Peking University
Tel: 86-10-62753016
Email: ell@pku.edu.cn
Website: http://www.phonetics.ac.cn/index.htm

实验室人员		联系方式
孔江平	教授	jpkong@pku.edu.cn
李铎	副研究员	lid@pku.edu.cn
周燕	高级工程师	ell@pku.edu.cn
汪高武	04 级博士生	wanggaowu@pku.edu.cn/wanggaowu@gmail.com
李永宏	博士后	lyhweiwei@126.com
尹基德	05 级博士生	yoonrealjdm@gmail.com
吉永郁代	05 级博士生	i_yoshinaga@hotmail.com
李英浩	07 级博士生	leeyoungho@pku.edu.cn
潘晓声	07 级博士生	itol_xs@pku.edu.cn
吴韩娜	07 级博士生	hanna.pku@gmail.com
杨峰	08 级博士生	yangzihuai@163.com
董理	09 级直博生	dongli_pku@yahoo.com.cn
曹洪林	09 级博士生	caohonglin2009@163.com
于谦	10 级博士生	mauriceyq@hotmail.com
吴君如	08 级硕士生	yuerr@163.com

（自 78 年以来，实验室共培养研究生 30 名，其中博士 20 名，硕士 10 名。）

本期执行编辑：董理

目录

A Study on the Origin of Tibetan Tones by Homonym Rate

Kong Jiangping
Center for Chinese Linguistics;
Department of Chinese Language and Literature
Peking University, P.R. China

1. Introduction

The languages in Sino-Tibetan language family most are tonal languages, so the origin of tones is very significant in the study of Sino-Tibetan language. However, the present living languages in this family are already tonal languages and only few with non-tonal and tonal dialects. For lack of written language materials, the study on origin of Sino-Tibetan tones is very difficult. In Sino-Tibetan language family, there are 3 languages with old traditional writing systems. They are Chinese, Tibetan and Burmese among which only written Tibetan reflects a non-tonal ancient speech sound system (马学良, 2003). So we say that Tibetan is a very important language in the research of the Sino-Tibetan language family, especially in the research of the origin of tone. The reasons are: 1) it has large scale of native speakers; 2) the geographical environment is very special, because most of Tibetan speakers live at high altitudes; 3) the humanity environment is that Tibetan lives relatively isolated and use the same language; 4) it has a traditional writing system with large quantity of writing materials; 5) it has many dialects which show the different phases of Tibetan evolution, especially the evolution from non-tonal to tonal.

The written Tibetan system was established in 7th century and the earliest written Tibetan materials now we can have are those between 8th century and 9th century (王尧, 1982; 李方桂等, 1987; Richaydson, 1985). At present, most of scholars think it reflects the ancient Tibetan sound system. The earliest written According to our database[1] of written Tibetan dictionary, there are more than 90,000 words, among which there are about 5649 single syllable words and more than 43000 di-syllable words. From the viewpoint of historical linguistics, the evolution of Tibetan tones has close relationship with the decrease of Tibetan initials and finals. Usually people think that there are 213 initials and 77 finals or 220 initials and 98 finals (格桑居冕等, 1991，2002，2004). According to our database, there are 220 initials, among which there are 30 single consonants initials, 118 initials with two consonant, 66 initials with three consonants and 6 initials with four consonants and 100 finals among which there are 5 single vowel finals, 15 diphthong finals, 45 finals with consonant ending and 35 finals with double consonant ending (孔江平, 2010).

From the written Tibetan and dialects, we know that decrease of initial has relationship with the increase of homonym and the homonym leads to the origin of Tibetan tones. So the Tibetan dialects can reflect the different stages and information of the origin of Tibetan tones. Tibetan dialects divided by scholars: a) 5 dialects by Roerich G.N. in 30s last century (Roerich, 1931); b) 4 dialects by Uray G. in 50s and 5 dialects in 60s (Uray, 1949); c) 4 dialects by Shafer R. (Shafer, 1955); d) 2 dialects by Miller R.A. (Miller, 1955). e) 3 dialects by Qu Aitang (瞿霭堂, 1963); 5 dialects by 西田龙雄(Nishida Tatsuo) in 70s (西田龙雄, 1970); f) 3 dialects by Li Fanggui (Li Fanggui, 1973); g) 2 dialects by Encyclopedia Britannica in 70s (Encyclopedia Britannica, 1974). In 80s, 3 dialects and many sub-dialects are divided by Qu Aitang (瞿霭堂等, 1981): 1) Weizang (central) dialects with 4 sub-dialects; 2) Kang dialect with 6 sub-dialects and 3) Amdo dialects with 4 sub-dialects. See figure 1.

[1] The Tibetan database includes 6 Tibetan-Chinese dictionaries and Tibetan many textbooks. They are: 《藏汉大字典》、《安多口语字典》、《拉萨口语字典》、《格西曲扎藏文辞典》、《新编藏文字典》、《藏文同音字典》和《藏语文课本（小学 12 册、初中 6 册、高中 6 册）》。

Since it is very difficult to identify the sound systems which is developing tone system from pitch pattern to tone pattern, there are some problems in the principle of dialect division.

Figure 1: the map of Tibetan dialects[2].

Briefly speaking, the dialects which have tones and do not have voiced stop, voiced affricate and voice fricative are regarded as Weizang dialect, the dialects which do not have tones and have voiced stop, voiced affricate and voiced fricative are regarded as Amdo dialect and the rest dialects are regarded as Kang (中国社会科学院和澳大利亚人文科学院, 1987).

2. Pitch and tone patterns

In the phonetic study, the basis of historical sound change is very important. According to the linguistic study, the evolution of tone has at least three periods. The first one is the period of pitch pattern in which the pitch pattern of single syllable mainly is “high falling pitch”, and the pitch pattern of di-syllable is usually “middle level + high falling”. Most of Amdo dialects are in this period. The second one is the period of tone development in which the pitch or tone patterns of single syllable mainly are “high falling pitch” and “high level pitch”, such as the dialects in Yushu prefecture. The pitch or tone patterns of disyllables usually “middle level + high falling” and “middle level + high level”. The third one is the period of developed tone system in which there are usually 4 or 6 tone patterns of single syllables and about 10 tone pattern of disyllables. Some words in Xiahe, Yushu and Lasa are listed in table 1. See table 1.

Table 1. Examples of single syllable words in 3 dialects.

Tibetan	transcirption	Meaning	type	xiahe	yushu	Lasa
རྟ	rta	horse	QD	hta$^{51/51}$	ta$^{51/51}$	ta^{53}
ས	sa	place	QD	sha$^{51/51}$	sha$^{51/51}$	sha^{53}
ཨཱུ	au	grand mother	QC	ʔi$^{51/51}$	ai$^{51/51}$	ʔau^{55}
གསུམ	gsum	three	QC	səm$^{51/51}$	saŋ$^{51/51}$	sum^{55}
ཚིག	tshig	sentence	QS	tshək$^{51/51}$	tshɛʔ$^{55/55}$	tshk^{53}
བཅིབ	bcib	ride	QS	tɕəp$^{51/51}$	tɕeʔ$^{55/55}$	tɕip^{53}
རི	ri	mountain	ZD	rə$^{251/51}$	rə$^{351/51}$	ri^{13}
ང	nga	I	ZD	ŋa$^{251/51}$	ŋa$^{353/353}$	ŋa^{13}
མདའ	mdav	arrow	ZC	nda$^{2251/51}$	nda$^{5351/51}$	ta^{15}
གཟན	gzan	cassock	ZC	zan$^{251/51}$	zɛ̃$^{353/353}$	sɛ^{15}
བརྡུངས	brdungs	beat	ZS	doŋ$^{2251/51}$	dɑŋ$^{3153/51}$	tuŋ^{132}
གཟིགས	gzigs	look	ZS	zək$^{2251/51}$	zɛ$^{2125/15}$	saʔ^{132}

In table 1, the first column is written Tibetan, the second is the Latin transcription of Tibetan, the third is the word meaning, the fourth is syllable type in which ‘Q’ stands for voiceless initial, ‘D’ stands for short final, ‘C’ stands for long final and ‘S’ stands for final with stop ending, the fifth column to seventh column are the words of Xiahe, Yushu and Lasa in

[2] The figure is copied from “Atlas of Chinese Languages” 《中国语言地图集》, 中国社会科学院和澳大利亚人文科学院合编, Longman, Hong Kong, 1987.

IPA. The numbers before '/' reflect the actual fundamental frequency by 5 letters tone system and the numbers after '/' is obtained by the perception of linguists.

From the table 1, we can find that the F0 contours of some voiced initials, especially those with voiced stops, can't be perceived as pitch pattern. So there is only one pitch pattern in Xiahe, though the syllable types are different. There are two pitch patterns in Yushu, though the syllable types are different. There are 5 tone patterns in Lasa, which have close relationship with the syllable types.

Table 2. Pitch pattern or tone pattern of single syllable words in 3 dialects.

	Xiahe	Yushu	Lasa
QC	51/51	51/51	55
QD	51/51	51/51	53
QS	51/51	55/55	53
ZC	2251/51, 251/51, 251/25	5351/51, 351/351, 353/353	15
ZD	2251/51, 251/51	251/51, 351/51, 353/353	13
ZS	2251/51, 251/51	3151/51, 55/55, 2125/15	132

In Table 2, the first column stands for the syllable type and the others display the pitch or tone patterns of the three dialects from more examples in our database. The pitch patterns in Xiahe dialect are displayed in the second column, the pitch or tone patterns in Yushu are displayed in the third column and some tone patterns of Lasa are displayed in the 4th column. From table 2, we can find that the pitch patterns are not very fixed which can be easily affected by initials and finals.

From what we talked above, we can find that there is great difference between F0 contour and the perceptual pitch pattern in Xiahe and Yushu which are usually regarded as non-tonal dialects and the tone developing dialects. The difference of F0 contour and perceptual pitch pattern is small in Lasa dialect which is usually regarded as tonal dialect. From this analysis, we can find that the difference between F0 contour and perceptual pitch pattern can be used to define tonal or non-tonal dialect or language, because the sound system does not depend on F0 contour but perceptual system of our brain. No matter how long and how complex a F0 contour is, a syllable only can be perceived as one single pattern. That is to say the physiological restriction and psychological perception play a very important role in the evolution of tone in Tibetan. That is the reason why it is very difficult to identify the pitch or tone systems in the tone developing dialects.

3. Homonym rate of initial and final

As is well known, there are 3 speeches which should be recorded when people investigate a Tibetan dialects. They are local speech of ordinary people used in daily life, local speech for reading written Tibetan and speech used by Lama which is more close to ancient Tibetan. Here we use local speech of ordinary people to explain the pitch and tone patterns, and local speech for reading written Tibetan to calculate the homonym rates. The reason is that different word corpuses have to be used, if the local speeches of daily life in different dialects are used.

Since the numbers of initials and finals in the dialects of Tibetan decrease gradually from 7th century, the result is that the homonyms increase gradually and finally may affect the normal communication. In this study, written Tibetan and 6 dialects are selected for explaining the relationship between the evolution of tones and the homonyms. The words used in this calculation are all single syllable words. The total number is 5649. In this paper, HR1 stands for homonym rate of initials and finals, also called homonym rate 1. TN stands for the total number of all single syllable words. NIF stands for the number of syllables with different initials and finals.

Method for calculating homonym rate 1:

HR1=TN/NIF – 1

For example: The homonym rate 1 of written Tibetan.

HR1 (written Tibetan) = 5649/5649 – 1= 0

The results of homonym rate 1 of written Tibetan and the dialects of Xiahe, Zeku, Batang, Dege, Rikaze and Lasa are listed in Table 3. See Table 3.

Table 3. Homonym rate calculated through initials and finals

Dialect	TN	NS	HR1
Tibetan	5649	5649	0
Xiahe	5649	1942	1.9089
Zeku	5649	1579	2.5776
Batang	5649	1398	3.0408
Dege	5649	1336	3.2283
Rikaze	5649	1086	4.2017
Lasa	5649	1052	4.3698

Table 3 shows the homonym rate 1 of written Tibetan and 6 Tibetan dialects, in which the first column is the dialect name, the second column is the total number of single words, the third column is the number of syllables with different initials and finals and the fourth column is the homonym rate 1. In order to see the result clearly and directly, the homonym rates are displayed in Figure 1.

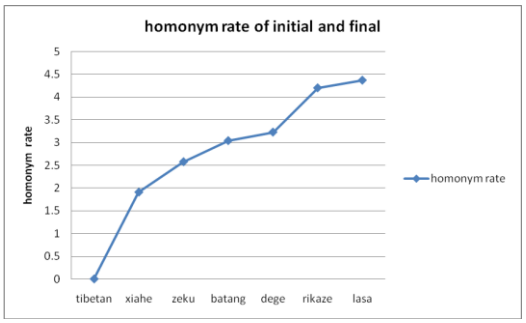


Figure 2. The homonym rate 1 of written Tibetan and 6 dialects are sorted and displayed in an increasing order.

From Figure 2, it can be clearly seen that the homonym rates increase gradually and the homonym rate 1 of written Tibetan is 0 that means there is no homonym in 5649 single syllable words, the homonym rate 1 of Xiahe is 1.9088, the homonym rate 1 of Zeku is 2.5775, the homonym rate 1 of Batang is 3.0407, the homonym rate 1 of Dege is 3.2282, the homonym rate 1 of Rikaze is 4.2014 and the homonym rate 1 of Lasa is 4.3697. As is well known, the written Tibetan, Xiahe dialect and Zeku dialect are usually regarded as non-tonal dialects in Tibetan and the others are regarded as tonal dialects according to the data now used. From the results, we can find that the homonym rate 1 '3' which is close to the homonym rate 1 of Batang is very important, because it is the watershed for tonal and on tonal.

If the procedure of evolution starts at one point and diffuses gradually from the centre and the evolution speeds of all dialects are as same as that in Lasa, the development of homonym rate in Xiahe is only in 524 of the 1200 years (from the beginning of 9th century to the end of 20th century), Zeku is in 707, Batang is in 835, Dege is in 886 and Rikaze is 1154. From these results, we may also get to know the stages of Tibetan dialects and the distance between among the written Tibetan and dialects. See Figure 4.

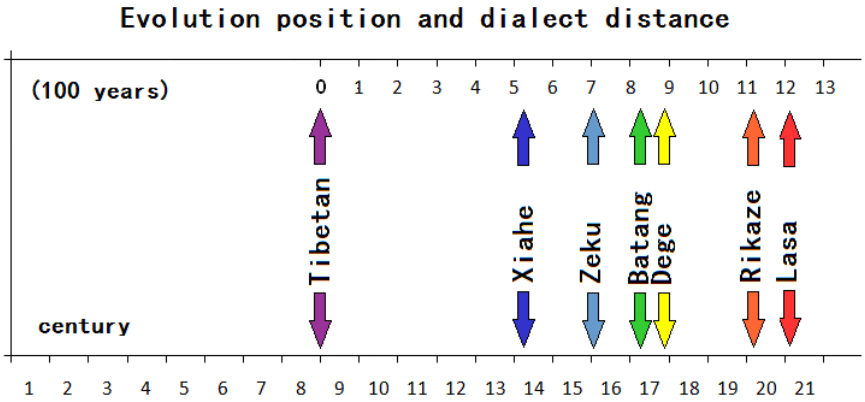


Figure 3. The evolution of dialects and the distances among dialects. The upper scale is defined as 100 years and the lower scale is the century (AD).

From figure 3, we can see that the evolution stages of different Tibetan dialects, for example, the the dialect which develops fastest is Lasa and the dialect which develops slowest is Xiahe, and the distances among written Tibetan and Tiebtan dialects, for example, there is only 46 (years) between Lasa and Rikaze and 676 (years) between Lasa and Xiahe.

4. Homonym rate of syllable

In the last section, we have discussed the homonym of initial and final, because some of the dialects are non-tonal. In this section, the homonym of syllable are calculated and discussed. HR2 stands for homonym rate of syllables, also called homonym rate 2. NS stands for the number of different syllables.

Method for calculating homonym rate 1:

HR2=TN/NS – 1

For example: The homonym rate 2 of Lasa.

HR2 (Lasa) = TN/NS – 1=1.7263

The words used in this calculation are all single syllable words. The total number is 5649. The results of homonym rate 2 of written Tibetan and the dialects of Xiahe, Zeku, Batang, Dege, Rikaze and Lasa are listed in Table 4. See Table 4.

Table 4. Homonym rate calculated through syllables

Dialect	**TN**	**syllable num.**	**homonym rate**
tibetan	5649	5649	0
xiahe	5649	1942	1.9088
zeku	5649	1579	2.5775
batang	5649	1760	2.2096
dege	5649	1891	1.9873
rikaze	5649	2130	1.6521
lasa	5649	2072	1.7263

Table 4 shows the homonym rate 2 of written Tibetan and 6 Tibetan dialects, in which the first column is the dialect name, the second column is the total number of single words, the third column is the number of syllables with different initials, finals and tones, and the fourth column is the homonym rate 2. In order to see the result clearly and directly, the homonym rates are displayed in Figure 5.

From table 4, we can see that the homonym rate of written Tibetan is 0, the homonym rate of Xiahe is 1.9088, the homonym of Zeku is 2.5775, the homonym rate of Batang is 2.2096, the homonym rate of Dege is 1.9873, the homonym rate of Rikaze is 1.6521 and the homonym rate of Lasa is 1.7263.

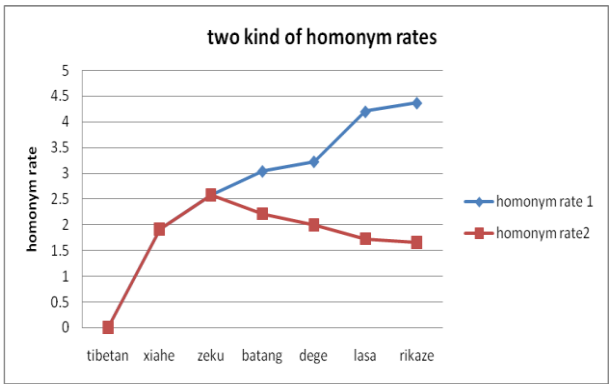


Figure 4. The homonym rate 1 calculated from initials and finals and the homonym rate 2 calculated from syllables in written Tibetan and 6 dialects.

In figure 4, the HR 1 and HR2 are displayed in an increasing order of HR1. We can see that the turning point of HR2 is at 2.5775 which is the dialect of Zeku. The homonym rate 2 of dialect which is around 2.5 may indicate that the dialect begins to develop its tone system. According to this, the stages of tone development in these dialects can be predicted by HR1 and HR2, though the systems in these dialects cannot be investigated exactly. In addition, we can see that the time of Zeku is around 14

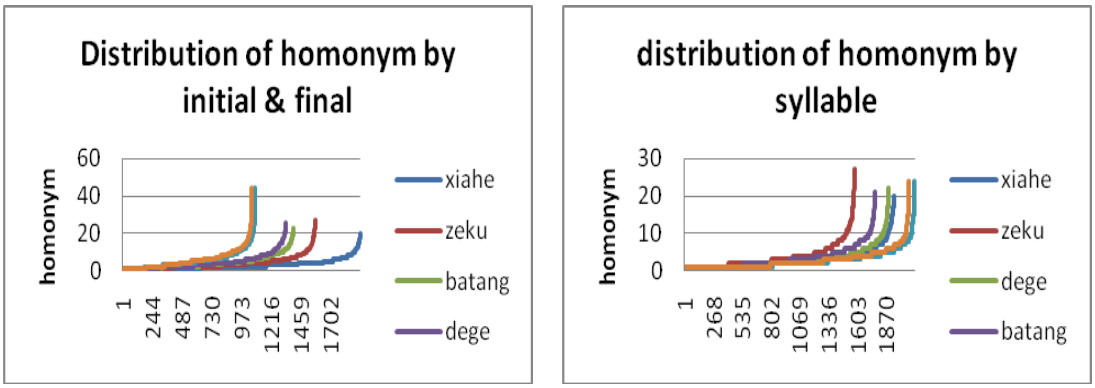
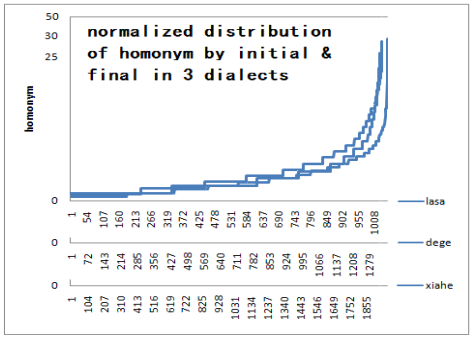


Figure 5: There are two plots in this figure. The left and right display HR1 and HR2 distributions of the 6 dialects in increasing orders respectively.

From figure 5, the distribution contours of HR1 show the procedure of HR1 development, from which we can find that the differences are in numbers of HR1 and syllable, and the same nature which look like 'exponential function' in all the 6 dialects. That means that HR1 develops and increases fast only in few syllables which may become tonal first the dialects. The contours of HR2 show the procedure of HR2 development, from which we can find that the differences are only in number, and the same nature which also look like 'exponential function' in all 6 dialects. That means that the highest numbers of HR2 are very close to each other in the 6 dialects.



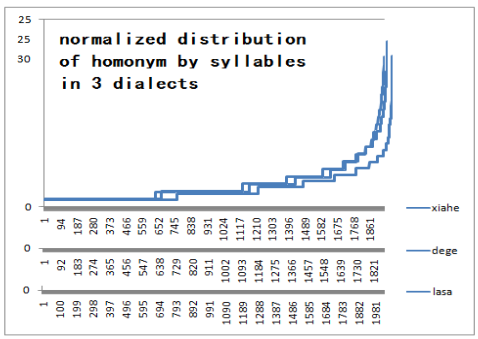


Figure 6: There are two plots in this figure. The left and right display the normalized HR1 and HR2 distributions of the 6 dialects in increasing orders respectively.

When these contours have been normalized by numbers of syllable HR1 or HR2, the nature of them can be clearly seen. They really have the same nature which tells us that the homonyms appear easily in some of the syllables and the others are not, that is to say, from the syllables which have more homonym will develop the tone first.

5. Conclusions and discussion

According to the analysis, the following conclusions can be drawn: 1) the increase of homonym rate 1 in Tibetan leads to the development of tone system; 2) Homonym rate 1 can be used to study the distances of Tibetan dialects; 3) The homonym rate 1 and 2 which is around 2.5 indicate that the dialect begins to develop its tone system; 4) the syllable with high homonym rate may become tonal; 5) The distributions of homonym rate 1 and 2 in each dialect have the same nature and the distribution contours is like an exponential function.

In historical linguistic viewpoint, the origin of Tibetan tones is usually studied through voiced and voiceless initials and the stop endings. The main problem in Tibetan dialects is that the pitch patterns and tone patterns are very difficult to investigate in the dialects which are developing their tone systems. To solve this problems needs the further studies on phonation types and the perception of pitch and tone patterns. However, the study in this paper opens up a new way to study the origin of Tibetan tones through the homonym rate. From a phonetician viewpoint, more and more attention has been paid not only to the study on evolution of tones linguistically, but also to the study on the information system, physiological mechanism and psychological event.

Reference

格桑居冕，1991，“藏语字性法与古藏语音系”, 民族语文，第六期。

马学良主编， An Introduction to Sino-Tibetan Languages, 汉藏语概论，民族出版社，2003 年 10 月。

孔江平，2010，古藏语音位系统的结构和分布,《研究之乐：庆祝王士元先生七十五寿辰学术论文集》潘悟云沈钟伟主编，上海教育出版社

《藏语方言概论》，格桑居冕，格桑央京，民族出版社，2002 年。

《实用藏文文法教程》（修订本），格桑居冕，格桑央京，四川民族出版社，2004 年。

瞿霭堂，1963,《藏语概况》,《中国语文》，第 6 期。

瞿霭堂，金效静，1981,《藏语方言的研究方法》,《西南民族学院学报》，第 3 期。

西田龙雄，《西番馆译语的研究---言语学序说》，京都松香堂，1970 年。

王尧, 1982,《吐蕃金石录》，文物出版社。

李方桂，柯蔚南，1987,《古代西藏碑文研究》，史语所专刊 91。

Rorich G.N., Modern Tibetan Phonetics, with special reference to the dialects of Central Tibet, JASB, XXVII, 1931.

Uray G., Kelet-Tibet nyelvjarhsainak o sztalyozasa (The classification of the dialects of eastern Tibet), Budqpest, 1949.

Shafer R., Classification of the Sino-Tibetan Languages, Word, 11:1, 1955.

Shafer R., Introduction to Sino-Tibetan, I-V, wiesbadan: otto Harrassowitz, 1966-1974

Li, Fangkuei, Language and Dialects of China, JCI, No.1, 1973

Miller, R.A., The Independent Status of the Lhasa Dialect within Central Tibetan, Orbis, Tome 4, No.1, 1955

Encyclopedia Britannica, Encyclopedia Britannica Inc. New york, 1974, reprint, 1984. (fifteenth edition)

中国社会科学院和澳大利亚人文科学院合编,《中国语言地图集》(Atlas of Chinese Languages), Longman, Hong Kong, 1987.

Richardson M.E., 1985, A Corpus of Early Tibetan Inscriptions, Royal Asiatic Society, Hertford.

An Exploration on the Measurement of Psycho Distance across Mandarin Phonemes

Wu Junru

Peking University

yuerr@163.com

Abstract

This study investigated the measurements of correct rate (CR) and reaction time (RT) under priming paradigm on language specified perception of similarity (termed "psycho distance") across phonemes carried by syllables. Mandarin speakers participated in two experiments and take speeded same-different determination tasks and shadowing tasks in both experiments, to see the psycho measurements' consistence with articulatory similarity (feature co-existence). Tasks and different feature co-existence conditions of stimuli pairs were manipulated to see their effect on CR and RT. It is proved that manipulating articulatory feature co-existence conditions would cause different correct rate as well as different reaction time for speeded same-different determination tasks. Sharing either an articulatory feature of place of articulation or aspiration would result in longer reaction time, and with both these features different the reaction time would be the shortest. Moreover, stimulus pairs different in place of articulation but only sharing the value of aspiration would result in significantly longer reaction time and lower correct rate than pairs in the other way..

1. INTRODUCTION

Many linguistic theories and hypothesis involve the belief of "similarity" across phonetic or phonological units. How "similar" two units sound alike? As for phones or phonemes, researchers have explored two main approaches. One is feature theory[1] [2] [3], which attribute similarity to sharing same features. The other is acoustic[4] [5] [6], which "quantifies" similarity by acoustic parameters. The two approaches do integrate, as what Stevens find [5]. However, do we have other ways to measure human concept of similarity directly? Can we measure "similarity" following this approach?

As for an objective and reliable measurement, psychological measurements of reaction time (RT) and correct rate (CR) of phonetic priming is worth trying. Perceptual identification of spoken words in noise is less accurate when the target are preceded by spoken phonetically related primes[7]. And the reaction times of shadowing tasks are also under the interference by similar priming [8]. These effects are attributed to competition between similar units. However, most the former studies were carried out based on similarity no lower than phonotactic level[9] [10],, and except Radeu's work not much concern was put on the differences caused by different degrees of similarity.

The overall goal in the present investigation is to test these measurements on the perceived "similarity" across Mandarin phonemes. It is reasonable to assume that perceived "similarity" is closely related to articulatory "similarity". This makes a more specific goal to investigate if the differences of feature co-existence result in significantly difference in the variables of CR and RT. That is to test the difference of priming effect across pairs of phonemes (carried by syllables) with different conditions of feature co-existence. "Psycho distance" is used in present study to term the language specified perception of similarity across phonemes. The more similar two units in the perception by speakers in a language, the smaller the psycho distance between the two units are in that language. If the difference of priming effect exists in this experiment, the operational definition of "psycho distance" can be set to the correct rate or reaction time in phonetic priming.

However, to use the priming paradigm and the two variables, some covariance factors should be considered. RT of phonetic priming is influenced by phonotactic probabilities of segments and sequences[11] [12].. Task and design also take their place [15] [16], auditory naming task may bias processing toward the sublexical level (facilitation), while the lexical decision task and naming task appear to bias processing toward the lexical level (interference). On the other way, using non-word stimuli or testing word stimuli in non-word environment increase the processing toward the sub-lexical level. As for Chinese, Zhou et

al. found Compound words sharing segmental templates but differing in lexical tone did not prime each other significantly in visual-visual lexical decision tasks but in auditory-visual priming lexical decision task[13]. Another variance comes from inter-stimulus interval (ISI). Shadowing latencies were significantly longer for target words preceded by phonetically related primes, but only when the prime–target inter-stimulus interval was short (50 vs. 500 msec) in another experiment by Luce et.al[10]. Zhou also found the influence of SOA.

2. EXPERIMENT 1

2.1. Method

Subjects. Seventeen undergraduate or graduate students in Peking University participated in this experiment (6 Male, 11 Female). All subjects are paid.

Stimuli. Prime and targets were Mandarin syllables. Four groups of stimuli pairs were used. The prime and target shared the same rhyme (/a/ or /u/) in each pair of every group and the difference only lies in initials. (Table 1) Another 24 pairs for fillers and 12 pairs for practice were produced. Stimuli and fillers were spoken by a female speaker in a format of 16 bit, 44100 Hz, 2ch, and then manipulated with PSOLA synthesizer in Praat program to a duration of 700ms and a level pitch of 290-285Hz.

Design. Two variables were examined by within group design: (1) feature co-existence of phonemes in primes and targets (same vs. different in PLA, identical in ASP vs. different in ASP identical in PLA vs. different in); (2) task (shadowing task vs. speeded same-different decision task). Reaction times of both tasks and correct of same-different decision tasks were examined. Word frequency type was also considered as a covariance measured but not manipulated.

In shadowing task, items were assigned to 12 groups, each group with 4 pairs of stimuli and 1 filler. Across groups interaction of PLA of prime initials with condition of feature co-existence was balanced out by Latin square. Pseudorandom (n=4) was applied to balance out order effect. In speeded same-different decision tasks, similar balance was made, except that the fillers were pairs of identical syllables to balance the number of same pairs and different pairs in the experiment. 7 subjects took shadowing task before same different task, 10 took same different task before shadowing task, eliminating the influence across tasks. In same-different decision tasks, 9 subjects used right hands to response to same pairs, left hand to different pairs and 8 used left hands to response to same pairs, right hands to different pairs, eliminating the influence of dominant hand.

	Same ASP		Different ASP	
	Same PLA	Different PLA	Same PLA	Different PLA
I	11	21	12	22
	pa-pa	pa-ta	pa-pha	pa-kha
	pha-pha	pha-kha	pha-pa	pha-ta
	ta-ta	ta-ka	ta-tha	ta-pha
	tha-tha	tha-pha	tha-ta	tha-ka
	ka-ka	ka-pa	ka-kha	ka-tha
	kha-kha	kha-tha	kha-ka	kha-pa
	pu-pu	pu-tu	pu-phu	pu-khu
	phu-phu	phu-khu	phu-pu	phu-tu
	tu-tu	tu-ku	tu-thu	tu-phu
	thu-thu	thu-phu	thu-tu	thu-ku
	ku-ku	ku-pu	ku-khu	ku-thu
	khu-khu	khu-thu	khu-ku	khu-pu

Table 1

Table 1: prime-target pairs. Each prime and target shared a same rhyme.

Procedure

The subjects were tested individually in a booth equipped with a voice response key interfaced to a minicomputer that controlled stimulus presentation and response collection. The experimental procedure was controlled by DMDX program [21].

The procedure of shadowing task is as follows. The subjects were instructed to repeat back the second word of the pair as quickly and as accurately as possible into a microphone attached to the headphones. RTs were recorded from the onset of the target word to the onset of the shadowing response.

silence (500ms)-mark sound (500ms) -prime (700ms)- ISI (50ms)-RT clock on -target (700ms)

The procedure of speeded same different task was the same as the shadowing task, except that the voice key was replaced by keyboard. The subjects were instructed to decide if the pairs are the same as quickly and as accurately as possible. RTs were recorded from the onset of the target word to the onset of the button press response.

2.2. Results and discussion

T	Same ASP		Different ASP	
a	Same PLA	Different	Same PLA	Different
s		PLA		PLA

k	RT		CR		RT		CR		RT		CR		RT		CR	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
De-	805	169	95	.11	837	150	76	.15	798	139	97	.08	791	146	98	.04
Sh-	889	190	-	-	901	178	-	-	878	194	-	-	881	171	-	-

Table2: Reaction Time (RT, in Milliseconds), Correct Rate (CR, in Percentage), and Standard Deviation for Determination Task (De-); Reaction Time (RT, in Milliseconds) and Standard Deviation for Shadowing Task (Sh-) in the 700ms Stimulus-Duration Condition

(1) Correct Rate of Speeded Same-Different Task (700ms)

One-way repeated ANOVA were conducted for correct rate and reaction time for decision tasks by participants with the two factors PLA (2 levels) and ASP (2 levels) integrated as four levels of one factor of feature co-existence. Effect of feature co-existence conditions on correct rate for decision tasks was significant by participants, $F(3, 48)=16.805$, $p<0.05$, LSD test showed significant difference between con11 and con21 $M=0.181$, $p<0.05$, con12 and con21 $M=0.206$, $p<0.05$, con 21 and con22 $M=0.216$, $p<0.05$, but no significant difference is between con11 and con12 $M=-0.025$, $p>0.05$, con11 and con22 $M=-0.034$, $p>0.05$, con12 and con22 $M=-0.010$, $p>0.05$. These indicate that, with inter-participant difference controlled, pairs with different PLA but same ASP tended to have lowest correct rate, however differences between other conditions were not higher than opportunity-level. The results on correct rate were negatively consistent with results on reaction time. Effect of feature co-existence on reaction time for decision tasks was insignificant by participants, $F(3, 48) = 1.766$, $p<.05$, which indicates that in this experiment the measurements were still not strong enough to stand out from inter-participant difference.

Two-way repeated ANOVA were conducted for correct rate and reaction time for determination tasks by participants.

For correct rate, main effect of both PLA co-existence, $F(1, 16) = 10.206$, $p<0.05$, ASP co-existence, $F(1, 16) = 29.829$, $p<0.05$, and their interaction, $F(1, 16) = 14.318$, $p<0.05$, was significant. For reaction time, main effect of neither PLA co-existence, $F(1, 16) = 0.577$, $p>.05$, nor ASP co-existence, $F(1, 16) = 3.678$, $p>.05$, was significant, nor was the interaction between the two factors, $F(1,16) = 1.631$, $p>.05$. This indicates that sharing ASP value in a pair caused significantly lower correct rate; with different ASP value, sharing PLA value also caused lower correct rate. In spite of the existence of interaction in statistics, the fact that identical pairs required a different button-press reaction from the other three conditions could explain the high correct rate under this condition.

(2) Shadowing Task (700ms)

One-way repeated ANOVA were conducted for reaction time for shadowing tasks by participants with the two factors PLA and ASP integrated as four levels of feature co-existence. Effect of feature co-existence on reaction time for shadowing tasks was insignificant by participants. $F(3, 48) = 1.292$, $p<.05$, which indicates that in this experiment by this measurement feature co-existence conditions didn’t cause significant enough effect.

Two-way repeated ANOVA were conducted for reaction time for shadowing tasks by participants. Main effect of either PLA co-existence, $F(1, 16) = 1.673$, $p>.05$, and ASP co-existence, $F(1, 16) = 1.289$, $p>.05$, was insignificant, nor was the interaction between the two factors, $F(1,16) = 0.017$, $p>.05$.

The results of Experiment 1 proved the existence of effects from feature co-existence. However, more concrete effects across conditions seemed to be submerged by effects of some extra factors, and some participants suggested that the tasks were too easy that they hardly made any hesitation or mistake. This ceiling effect was proved by the lack of significance of difference across other feature co-existence conditions - only different PLA-same ASP condition was significantly different from other feature co-existence conditions. To access significant differences across all conditions, we designed Experiment 2.

3. Experiment 2

3.1. Method

Subjects: Thirty-two undergraduate or graduate students participated in this experiment (14 Male, 18 Female). All subjects are paid.

Stimuli: Stimuli were similar in Experiment1 except that the duration of each prime or target was adjusted to 400ms instead of 700ms.

Design: Like Experiment1, two variables were examined by within group design: (1) feature co-existence of phonemes,(2)task. Reaction times of both tasks and correct rates of same-different decision tasks were examined.

Procedure: The procedure was close to Experiment1, except that the duration of each stimulus was 400ms.

3.2. Results and Discussion

Task	Same ASP								Different ASP							
	Same PLA				Different PLA				Same PLA				Different PLA			
	RT		CR		RT		CR		RT		CR		RT		CR	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
De-	669	123	95	.07	716	132	73	.15	658	122	96	.07	643	107	98	.03
Sh-	709	132	-	-	720	113	-	-	696	111	-	-	708	118	-	-

Table 3: Reaction Time (RT, in Milliseconds), Correct Rate (CR, in Percentage), and Standard Deviation for Determination Task (De-); Reaction Time (RT, in Milliseconds) and Standard Deviation for Shadowing Task (Sh-) in the 400ms Stimulus-Duration Condition

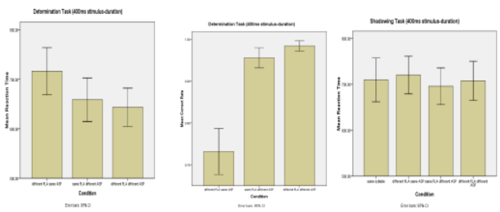


Figure 1 Reaction time (left) and correct rate (middle) under different feature co-existence conditions (identical condition excluded) for determination tasks and reaction time (right) for shadowing tasks in the 400ms stimulus-duration condition.

(1) Speeded Same-Different Task (400ms)

One-way repeated ANOVA were conducted for correct rate and reaction time by participants with the two factors PLA (2 levels) and ASP (2 levels) integrated as four levels of one factor of feature co-existence. Effect of feature co-existence conditions on correct rate for decision tasks was significant by participants, F (3, 93)=51.303, p<0.05, LSD test showed significant difference between con11 and con21 M=0.219, p<0.05, con11 and con22 M=-0.34, p<0.05, con12 and con21 M=0.224, p<0.05, con12 and con22 M=-0.29,and p<0.05, con 21 and con22 M=-0.253, p<0.05, but no significant difference is between con11 and con12 M=-0.05, p>0.05. These indicate that, with inter-participant difference controlled, pairs with different PLA but same ASP tended to have lowest correct rate, followed by pairs with different ASP but same PLA and pairs with different PLA and same ASP had the highest correct rate (Fig. 1) Effect of feature co-existence on reaction time for decision tasks was significant by participants, F (3, 93)=12.62, p<.05, LSD test showed significant difference between con11 and con21, M=-46.969, p<.05, con11 and con22, M=25.906, p<.05, con12 and con21, M=-57.656, p<.05, con 21 and con22, M=72.875, p<.05, but no significant difference is between con11 and con12 M=10.688, p=>.05, nor between con12 and con22, M=15.219, p=>.05. Two-way repeated ANOVA were conducted for correct rate and reaction time for determination tasks by participants. For correct rate, main effect of both PLA co-existence, F (1, 31) = 32.577, p<0.05, ASP co-existence, F (1, 31) = 59.827, p<0.05, and their interaction, F (1, 31) = 62.926, p<0.05, was significant. For reaction time, main effect of either PLA co-existence, F (1, 31) = 4.473, p<.05, and ASP co-existence, F (1, 31) = 20.223, p<.05, was significant, so was the interaction between the two factors, F (1,31) = 10.486, p<.05. This indicates that sharing ASP value in a pair caused significantly lower correct rate and longer reaction time; with different ASP value, sharing PLA value also caused lower correct rate and longer reaction time. In spite of the existence of interaction in statistics, the fact that identical pairs required a different button-press reaction from the other three conditions could explain the high correct rate under this condition.

(2) Shadowing Task (400ms)

Two-way repeated ANOVA were conducted for reaction time for shadowing tasks by participants with PLA and ASP as two factors. Main effect of either PLA co-existence, F (1, 31) = 0.803, p>.05, and ASP co-existence, F (1, 31) = 1.286, p>.05, was insignificant, nor was the interaction between the two factors, F (1, 31) = 0.003, p>.05.

For shadowing tasks, effect of feature co-existence on reaction time was insignificant by participants, F (3, 93) = 0.691, p>. 05, which indicated that in this experiment by this measurement feature co-existence conditions didn't cause significant enough effect.

By adding difficulty to the tasks (stimulus-duration shorten), Experiment2 successfully elicited more difference across different feature co-existence conditions. Sharing either feature in pairs would cause lower correct rates and longer reaction time. Nevertheless, effect of sharing place of articulation (PLA) seemed to have larger effect. Pairs sharing aspiration (ASP)

value but by different place of articulation (PLA) (e.g. pa-ta), with lowest correct rate and longest reaction time, were most difficult to discriminate and caused largest priming interference effect, thus had smallest psycho-distance; pairs sharing place of articulation but by different aspiration value (e.g. pa-pha) were the second difficult to discriminate, thus had the second small psycho-distance; pairs with different place of articulation and aspiration value (e.g. pa-tha) were most easy to discriminate, thus had the largest psycho-distance. Identical syllables (repetition condition), like what Radeau [8] noted, were exceptions, and showed facilitation effects in his experiment. According to our results, no difference in articulatory feature didn't mean most difficult to "discriminate" or largest priming interference effect.

However, data of shadowing task still failed to be significant in this experiment. According to former studies, shadowing tasks tended to have smaller effects than determination tasks [16]. Perhaps differences of similarity on feature level were too subtle to elicit significant difference for this task.

4. Conclusion and General Discussion

Back to our goal of present experiments, the hypothesis that priming effects across pairs of phonemes (carried by syllables) vary under different feature co-existence conditions was proved by the experiment results. The results supported the existence of measurable psycho distances across phonemes in Mandarin, and proved their correlation with feature co-existence conditions – phoneme pairs with more features shared had smaller psycho distances than pairs with less features shared, and different values of aspiration contribute more to increase psycho distances than different places of articulation in Mandarin.

However, besides the effect of feature co-existence conditions, the measured psycho distances varied by tasks and measurements too. The effects on reaction time were significantly larger and more stable for speeded same-different tasks than shadowing tasks, in consistent with Vitevitch and Luce's reports[15] [16]. Effects of correct rates were significant, and more stable. However, the problem of correct rate lies on the ceiling effect – when tasks were easy, the subjects just barely made any error, and then the correct rates tended to be near 1.00. Though word frequency factors were mentioned a lot in literatures, they were not so significant in the present two experiments and only significant in interaction effect of Experiment2. Main effect of feature co-existence tended to be smaller primed by low frequency syllables.

In addition, we found besides the ISI and pronouncing problems by longer word in literature[16], duration of stimuli also mattered. The paradigm was based on hesitation and errors. When stimulus duration was 700ms, participants didn't make lots hesitation or errors. Only the most similar pairs were significantly different and the ceiling effect on correct rate was strong, since some participants just barely made any error. With durations of stimuli adjusted to 400ms in Experiment2, more subtle effects appeared - differences between other conditions turned significant, and even those best started making some error, eliminating the ceiling effect.

5. References

[1] M. Halle, "Preliminaries to Linguistic Phonetics," *Language,* vol. 49, pp. 926-933, 1973.

[2] J. Goldsmith, "An Overview of Autosegmental Phonology," *Linguistic Analysis,* vol. 2.1, pp. 23-68, 1976.

[3] P. Ladefoged, "Phonetic prerequisites for a distinctive feature theory," *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre,* pp. 273-285, 1972.

[4] K. N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," *The Journal of the Acoustical Society of America,* vol. 69, p. S116, 1981.

[5] K. N. Stevens and S. E. Blumstein, "The search for invariant acoustic correlates of phonetic features," *Perspectives on the study of speech,* pp. 1-38, 1981.

[6] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics,* vol. 17, pp. 3–45, 1989.

[7] S. D. Goldinger, P. A. Luce, and D. B. Pisoni, "Priming lexical neighbors of spoken words: Effects of competition and inhibition," *Journal of Memory and Language,* vol. 28, pp. 501-518, 1989.

[8] M. Radeau, J. Morais, and A. Dewier, "Phonological priming in spoken word recognition: Task effects," *Memory & Cognition,* vol. 17, pp. 525-535, 1989.

[9] P. A. Luce, D. B. Pisoni, and S. D. Goldinger, "Similarity neighborhoods of spoken words," *Cognitive models of speech processing: Psycholinguistic and computational perspectives,* vol. 540, pp. 122–147, 1990.

[10] P. A. Luce, S. D. Goldinger, E. T. Auer, and M. S. Vitevitch, "Phonetic priming, neighborhood activation, and

PARSYN," *Perception and Psychophysics*, vol. 62, pp. 615-625, 2000.

[11] M. S. Vitevitch and P. A. Luce, "When words compete: Levels of processing in perception of spoken words," *Psychological Science*, pp. 325-329, 1998.

[12] M. S. Vitevitch and P. A. Luce, "Probabilistic phonotactics and neighborhood activation in spoken word recognition," *Journal of Memory and Language*, vol. 40, pp. 374-408, 1999.

[13] X. Zhou, "Phonology in lexical processing of Chinese: Priming tone neighbours," *Psychological Science*, vol. 23, pp. 133-140, 2000.

[14] K. I. Forster and J. C. Forster, "DMDX: A Windows display program with millisecond accuracy," *Behavior Research Methods Instruments and Computers*, vol. 35, pp. 116-124, 2003.

Effect of Speech Rate on Inter-segmental Coarticulation in Standard Chinese

Yinghao LI, Jiangping KONG
Phonetics Lab
Peking University, Beijing
leeyoungho@pku.edu.cn; jpkong@pku.edu.cn

Abstract—**Coarticulatory variations of consonant clusters spanning morpheme boundary and transconsonantal vocalic articulatory transition in Standard Chinese as a function of speech rates were investigated. Eight /V1C1#C2V2/ and eight /V1#pV2/ disyllable sequences were uttered by a female speaker at slow, normal, and fast speech rate. Electropalatographic, electroglottographic, and acoustic data were recorded and the kinematic properties of segments and the inter-segmental temporal coarticulation were analyzed. Results showed articulatory duration shortening and increased temporal coarticulation were the two major motor control mechanisms involved in fast speech. Consonant production manner, coupling effect of tongue body movement in consonant production, and prosodic affiliation of vowels were found to affect kinematic properties of segments and inter-segmental coarticulation when speech rate is increased. In the acoustic domain, the F2 trajectory was found to be sensitive to the increased gestural overlapping in fast speech. However, the correlation between the linguapalatal contact in mid-portion of vowels and the relative same-time F2 values was vague. It was also found that inter-segmental coarticulation was independent of the suprasegmental tier, which tends to speak against the tenets of the time structure model.**

Keywords-speech rate; inter-segmental coarticulation; Standard Chinese; electropalatography

I. INTRODUCTION

One of the principal objectives in studying speech rate effect is to find out speech motor control mechanism and its relevance to the linguistically based invariance. Previous studies have found that several articulatory mechanisms are responsible for articulatory variations in fast speech. Segment shortening and increased temporal coarticulation are claimed as the main articulatory adjustment in fast speech [1], [2]. Segment shortening in fast speech tends to result in phonetic target undershoot, or spatial reduction of articulatory target [3], which leads to reduced articulatory displacement toward the phonetic goal and slower peak velocity of participating articulators [4]. Speaker-specific articulatory strategies are also an important factor in explaining the articulatory variations [4], [5].

The variation of articulatory dynamics induced by varying speech rates has been dealt with in different speech production models. Linear models consider the duration of articulatory movements are compressed or contracted proportionally by a constant factor [6]. Nonlinear models contend that timing of successive speech movements is restructured and the inter-segmental phasing relations are shifted with varying speech rate [1], [7], and target undershooting is not a necessary result of increased speech rate. The articulatory constraint model adds that inter-segmental coarticulation is conditioned by the articulatory constraints of segment in question [8]. In recent speech recognition model, the coarticulation is modeled through bidirectional temporal filtering, assuming that the rate-induced vocalic formant deviation from target value is the result of duration-dependent articulatory target reduction [9].

The coarticulatory processes in the Standard Chinese exist in tonal and segmental tiers. According to the time structure model [10], the unit for the speech temporal organization is syllable with its boundaries being aligned with tonal target. Coarticulation can only occur between initial consonant and the following vowel in the syllable domain. Though the time structure model is the first of its kind to elucidate that the coarticulation is constrained within the tonal domain, no serious articulatory study has been done to examine the coarticulatory effect in segmental tier and articulatory adjustment as a function of speech rate in the Standard Chinese.

Electropalatographic (EPG), electroglottographic (EGG), and acoustic data are used to examine rate-induced inter-segmental coarticulation in this paper. Two experiments are addressed: first, the relative timing relation of C1#C2 (# indicates a morpheme boundary) and kinematic characteristics of individual consonants as a function of speech rate are investigated. The articulatory and acoustic properties for flanking vowels are also examined, as they might be affected with the increased consonantal gestural overlapping. Second, the transconsonantal vocalic anticipatory effect is studied through measuring the dynamic properties of vocalic articulatory transitions. The articulatory and acoustic properties of vowels, either static or dynamic, are also investigated.

The aim of the experiments is threefold: first, the articulatory mechanism controlling the speech rate is explored for comparison of the similar study for American English [2]. Second, the acoustic consequences of altered kinematic properties of segments and temporal

Supported by National Project of Scientific and Technical Supporting Programs Funded by Ministry of Science & Technology of China （NO. 2009BAH41B00）

coarticulation are examined to show the acoustic manifestation of articulatory adjustment. Third, the inter-segmental coarticulatory pattern in the Standard Chinese is tentatively put forth with small set of speech samples to challenge the tenets of [10].

II. METHOD

A. *Stimuli*

The stimuli were drawn from a large EPG speech corpus for the Standard Chinese currently under construction. The stimuli for the first experiment were eight nonsense disyllables /V1C1#C2V2/. C1 was velar nasal consonant /G/ and the C2 was either unaspirated alveolar stop /t/ or alveolar fricative /s/. Two vowels contrasting in height were selected to enrich phonetic context. The only high vowel that can appear after /s/ is /i1/(The apical vowel appears after /s/) instead of /i/. The monotonous rising or falling tones were applied because it was expected that the F0 contour movement was independent of the coarticulatory process in the segmental tier.

. In the second experiment six nonsense disyllables /V1#pV2/ were constructed with the vowels differing in tongue body specification in each disyllable. They were /pi#pa/, /pa#pi/, /si1#pa/, /Si2#pa/(S stands for retroflex fricative, and i2 for apical vowels appearing after retroflex fricative), /si1#pi/, and /Si2#pi/. The intervocalic unaspirated stop was assumed to minimally interfere with the transconsonantal articulatory transition. /ɿ/ and /ʅ/ are traditionally called apical vowels, and can only appear after alveolar and retroflex consonants. Symmetrical disyllables /pi#pi/ and /pa#pa/ were also included for comparing the vocalic coarticulatory pattern.

Disyllables are imbedded in the carrier sentence "wo214 shuo55 ___ zhe51 ge51 ci35." ("I say ____ this word"). Additional four filler utterances with the same pattern as the test utterance were placed in between the test utterances. All utterances were randomly grouped into four blocks with each block started and ended with a dummy utterance.

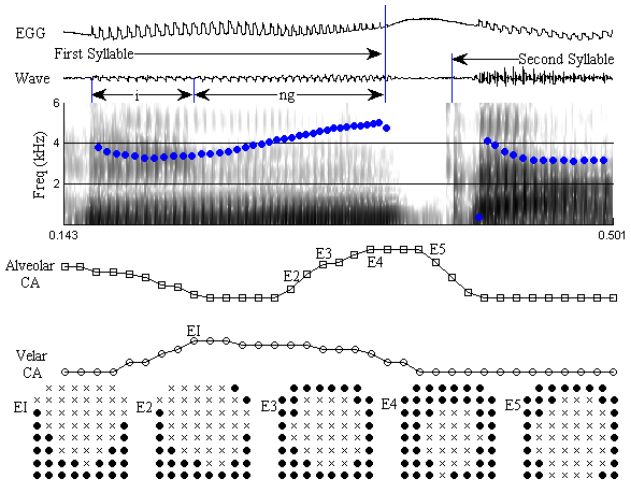


Figure 1. The marking protocols for /V1C1#C2V2/ sequences. (The five EPG frames at the bottom row respectively indicate maximum contact for C1 (E1), onset frame for C2 (E2), alveolar closure frame for C2 (E3), maximum contact for C2(E4), and last closure frame for C2 (E5). The pitch contour was shown in the spectrogram)

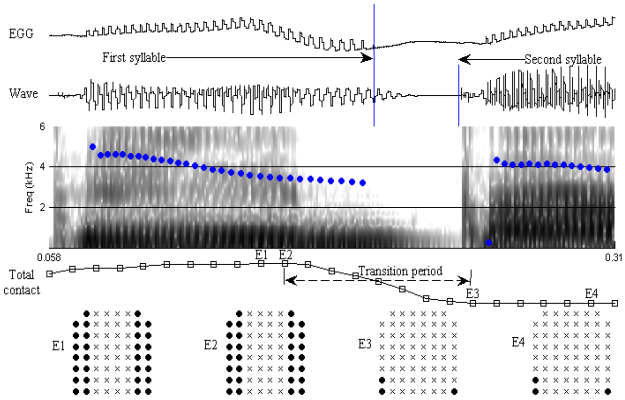


Figure 2. The marking protocols for /V1#pV2/ sequences. (The four EPG frames at the bottom row respectively indicate maximum contact for V1 (E1), onset frame for the articulatory transition (E2), offset frame for the transition (E3), and charateristic contact frame for /a/)

B. *Procedure*

One 27-year-old female speaker pronounced the utterances. A 62-electron artificial pseudo-palate was custom-made to place on the speaker's palate. The WinEPG system was used to acquire the linguapalatal contact signal every 10ms. The Kay EEG and the SONY ECM-44B microphone were connected to the WinEPG's serial port interface, and the laryngeal and acoustic signals were synchronized internally with the linguapalatal contact signal and sampled at 22050Hz. The recording was conducted in a sound-attenuated room.

The speaker was required to wear the pseudo-palate and practice for 15 minutes before recording. During recording the speaker was instructed to use self-directed speaking rate to read the sentence block for three times respectively at slow, normal, and fast rate consecutively. By doing so, it was observed that the speaker adopted a consistent rhythmic pattern after several trials.

The EPG, EGG and acoustic signal of the test disyllables were excerpted from the carrier sentences in the Matlab-based EPG analysis platform developed at the Phonetics Lab of Peking University. In total, 144 tokens were selected for further analysis.

C. *Marking and measurement*

The articulatory landmarks were determined by linguapalatal contact in specific contact area (CA) that was characteristic for consonantal place of articulation. The contact region for velar nasal was defined as the contact in the mid four columns of the last two rows, following [2]. The onset was the first frame when there were activated electrons in this region, and the offset the frame after the last frame with contacted electron on in this area. For velar nasal following tautosyllabic /a/, the acoustic landmarks served as articulatory onset and offset because no electrons were activated in the velar region

TABLE I. LINEAR REGRESSION R^2 VALUES (SLOP VALUES FOR A LINEAR FIT) FOR LATENCY AND KINEMATIC PROPERTIES IN V1C1#C2V2 SEQUENCES AS A FUNCTION OF SPEECH RATE (SIGNIFICANCE LEVEL WAS SET RESPECTIVELY AT P<0.05(*), P<0.01(**), P<0.001(***), AND MARGINAL SIGNIFICANCE P<0.1 (※). THE SHADED AREA REFERS TO DATA NOT AVAILABLE.)

Measure	*[iG#ta]*	*[iG#ti]*	*[iG#sa]*	*[iG#si1]*	*[aG#ta]*	*[aG#ti]*	*[aG#sa]*	*[aG#si1]*
ΔONSETS	0.806*** (0.56)	0.701** (0.66)	0.835*** (0.83)	0.766** (0.69)	n.a.	n.a.	n.a.	n.a.
ΔPEAKS	0.84*** (0.71)	0.3 (0.21)	0.466* (0.72)	0.525* (0.53)	n.a.	n.a.	n.a.	n.a.
C1 OLP	0.61* (-117.23)	0.849*** (-134.71)	0.24 (-150.35)	0.479※ (-104.96)	0.16 (-132.66)	0.17 (-95.74)	0.27 (-62.64)	0.24 (-194.68)
C1 DIS	0.14 (8.23)	0.27 (-4.52)	0.24 (24.74)	0.27 (18.72)	n.a.	n.a.	n.a.	n.a.
C1 DUR	0.73** (0.6)	0.513* (0.72)	0.591** (0.90)	0.651** (0.81)	n.a.	n.a.	n.a.	n.a.
C1 MaxC	0.14 (8.23)	0.27 (-4.52)	0.24 (24.74)	0.27 (18.72)	n.a.	n.a.	n.a.	n.a.
C1 VEL	0.05 (0.11)	0.545* (-0.75)	0.19 (0.87)	0.01 (0.18)	n.a.	n.a.	n.a.	n.a.
C2 DIS	0.886*** (51.77)	0.03 (2.82)	0.14 (14.54)	0.14 (14.67)	0.07 (10.15)	0.64* (24.12)	0.13 (-8.92)	0.16 (-15.93)
C2 DUR	0.746** (0.48)	0.795*** (0.34)	0.554* (0.6)	0.859*** (0.6)	0.58* (0.39)	0.77** (0.52)	0.9*** (0.58)	0.76** (1.01)
C2 MaxC	0.759** (45.84)	0.007 (1.2)	0 (0.51)	0.16 (-14.6)	0.07 (10.15)	0.64* (24.12)	0.13 (-8.92)	0.16 (-15.93)
C2 VEL	0.001 (0.06)	0.38 (0.54)	0.007 (0.21)	0.25 (1.34)	0.19 (-1.07)	0.2 (-2.06)	0.29 (-0.1)	0.26 (1.57)

The first three rows were defined as the alveolar region. The onset frame was determined in the same fashion as for velar nasal, but the offset frame was judged differently depending on following vowels. When /a/ was followed, the offset was the first frame with no activated electron in alveolar region. If /i/ or /i1/ was followed, the offset was defined as the first frame of the following vowel portion. The marking protocols for /V1C1#C2V2/ sequences were shown in Figure 1.

The marking protocols for vocalic transition in /V1#pV2/ were exemplified by Figure 2. The onset was the last frame in the plateau of the total contact profile for V1, and the offset the first frame reaching the phonetic target for V2. The transition from apical vowels to /i/ involved two articulatory processes: the loosening of the alveolar constriction and alveolo-palatal linguapalatal contact change. In this case, the onset frame was the last frame before considerable contact change, and the offset frame was the first frame that approach stabilized linguapalatal contact for /i/, normally one or two electrons less than the maximum contact for /i/.

The acoustic landmarks were marked manually for vowel onset and offset as well as velar nasal onset. The vowel onset and offset in a syllable was marked by the first and last detectable peaks in the first derivative of EGG signal. The nasal onset was determined through the observation of spectrogram and the formant trajectory. In /V1#pV2/ sequences, the F2 and F3 formant trajectories of V1were truncated toward the end of V1 due to lip closure gesture of the intervocalic consonant. Thus additional mark was made to delimit the end point for formant trajectories.

In the first experiment, the articulatory parameters included latency, overlap, duration, displacement and average velocity. The latency was defined following [2], Δ_{ONSETS} was the temporal interval between C1 and C2 onset frames, and Δ_{PEAKS} between C1 and C2 maximum contact frames. C1 overlap (OLP) was the percentage of the C1 interval during which C2 contact also appeared. Duration (DUR) was the time interval of segment onset and offset. Displacement (DIS) was the difference between maximum contact frame (MaxC) and the onset frame. Average velocity (VEL), which was an indirect measure to reflect dynamic characteristics of articulators, was defined as the ratio of the displacement against the time interval between the onset of consonant gesture and offset of maximum contact.

In the second experiment, displacement, duration, average velocity, shift of the vocalic transition, and maximum contact for vowels were measured. The displacement was the absolute margin of the contact between the onset and offset frames. Duration and average velocity were defined in similar fashion as in experiment 1. The shift was the percentage of the V1 period during which transition toward V2 was initiated. The linguapalatal contact for vowels in the middle portion of vowels (MidC) was also measure for correlating with F2 value in the same portion.

The speech rate measure was taken as the average of acoustic length of the two syllables in the test disyllable. The formant trajectories were obtained by using Burg method in Praat (20-ms window length, 5-ms time step, 5 maximum number of formants, 6000Hz maximum

formant), and fed into the Matlab-based EPG analysis platform for manually adjust-

TABLE II. LINEAR REGRESSION R^2 VALUES (SLOP VALUES FOR A LINEAR FIT) FOR TONGUE KINEMATIC AND ACOUSTIC PROPERTIES IN V1#PV2 SEQUENCES AS A FUNCTION OF SPEECH RATE

Measure	[i#pa]	[a#pi]	[i#pi]	[i1#pa]	[i1#pi]	[i2#pa]	[i2#pi]
DIS	079*** (78.1)	0.003 (-3.25)	n.a.	0.01 (1.28)	0.07 (18.42)	0.31※ (22.81)	0.62* (42.24)
DUR	0.65** (0.26)	0.65 ** (0.87)	n.a.	0.38※ (0.13)	0.58* (0.62)	054* (0.4)	0.76** (0.99)
VEL	0.03 (-0.07)	0.79*** (-0.917)	n.a.	0.43※ (-0.55)	0.20 (-0.55)	0.49* (-0.91)	0.09 (-0.04)
Shift	0.87*** (-263.47)	0.77*** (-149.62)	n.a	0.51* (-173.07)	0.13 (-87.26)	0.66** (-324.21)	0.57* (-277.1)
V1 MaxC	0.53* (60.19)	0.08 (-20.28)	0.03 (-9.91)	0.59* (34.91)	0.01 (-3.06)	0.69** (53.18)	0.37※ (30.1)
V2 MaxC	0.27 (-14.58)	0.02 (-7.9)	0.18 (-25.64)	0.2 (-5.69)	0.29 (40.22)	0 (0.28)	0.63* (45.53)
V1 MidC	0.01 (3.34)	0.6** (-17.65)	0.06 (-9.65)	0.85*** (35.39)	0.21 (24.21)	0.53* (45.1)	0.49※ (22.62)
V1 MidF2	0.34※ (803.76)	0.17 (489.39)	0.84*** (2051.12)	0.3 (915.94)	0.07 (-316.8)	0.76*** (3456.15)	0.49* (1858.74)
V2 MidC	0.52* (-16.58)	0.001 (2.0)	0.24 (22.94)	0.2 (-7.86)	0.42※ (60.0)	0.04 (-2.45)	0.67** (42.09)
V2 MidF2	0.17 (467.01)	0.54* (1424.87)	0.48※ (1576.76)	0.32 (-436.53)	0.28 (754.84)	0.10 (-253.49)	0.86*** (1007.2)

ment with reference to the cepstral spectrum. The F2 trajectory for V1 was averaged respectively for three speech rates and then normalized. The mid-time F2 value (MidF2) was the F2 value in the original F2 trajectory nearest to the mid-time point.

III. RESULTS

All articulatory and MidF2 were regressed against speech rate. Systematic changes as a function of speech rate were evidenced in both experiments. It was also found that the highest or lowest point of pitch contour aligned well with the syllable ending, thus leaving the questions only in the segmental tier.

A. *Experiment 1*

Table I shows the results of the significant regression (r^2) and goodness of a linear fit (slope value in the parenthesis) for the eight consonant sequences cross morpheme boundary.

Speaking rate has a significant effect on the absolute latency for the four sequences. Both Δ_{ONSETS} and Δ_{PEAKS} shorten proportionally as a function of rate. Although no significant correlation is achieved for Δ_{PEAKS} in /iŋ#ti/, the interval between the two peak contact frames tends to increase as speech rate is reduced.

A linear increase of temporal coarticulation is evidenced for three sequences in faster speech, which shows increased overlapping between the tongue body and tongue tip gestures. The gestural overlapping tends to be constrained by the consonant production manner and the lingual specification for flanking vowels. /s/ versus /t/ is more resistant to overlap with the preceding nasal consonant, indicative of the aerodynamic requirement for high-pressure buildup to produce an intelligible fricative. In /aG#tV2/ sequences a consistent larger gestural overlapping is observed across three speech rates, thus no significant increase of temporal coarticulation is found.

The acoustic consequences as a result of the combined articulatory constraints are reflected in the F2 trajectory of V1. The F2 trajectories in /aG#ta/ and /aG#ti/ are found to be sensitive to the increased gestural overlapping. As speech rate is increased, the direction of F2 trajectory approaches the loci value of the /t/. The anticipatory effect from V2 may also affect the V1's F2 trajectory direction, as is shown in Figure 3 (b). The F2 trajectories of V1 for other sequences tend to show a consistent pattern, which supports the resistance of gestural overlapping for /i/ before velar nasal and /s/ after velar nasal.

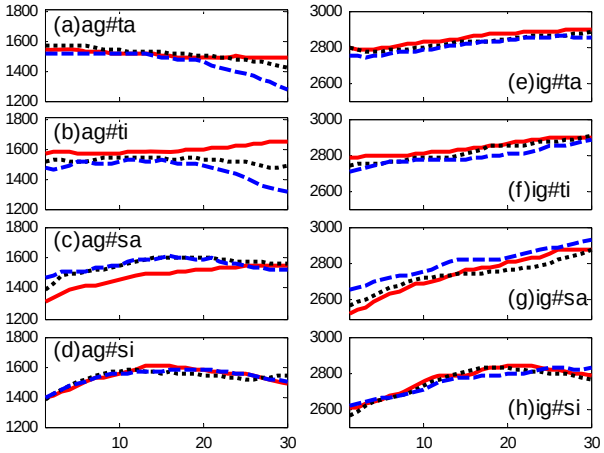


Figure 3. Normalized F2 trajectory of V1 averaged respectively for three speech rates. The solid stands for fast speed, dotted line for normal

speed,
and dashed line for slow speed.

The articulatory duration for individual consonants shortens proportionally as a function of speech rate. The displacement and maximum contact for C1 have no significant difference across speech rate, which may be attributed to the ceiling effect because the electrons in this area show relative full contact in the four sequences. Linear spatial reduction as a function of rate is only found for displacement and maximum contact for /t/ in the context of flanking vowels with contrasting tongue body height. This may be attributed to the small set of speech samples. Alveolar fricative tends to be resistant to the effect of speech rate or phonetic contexts, which is in conformity with [2].

Average velocity is an indirect measure reflecting the dynamic feature for participating articulators. Significant linear correlation is only found for velar nasal tongue body gesture in /iG#ti/. No significant linear relation is found for average velocity for C2 as a function of speech rate. The different average velocity is suggestive of the dynamical features for tongue body and tongue tip/blade. While tongue tip/blade movement normally involves precise and fast movement in consonant production, the movement for tongue body is relatively slower, which in most cases involves in vowel production [12].

The results of the first experiment show that strategies in producing consonant clusters across morpheme boundary are similar to that discovered in English [2]. Articulatory shortening and increased temporal coarticulation are the two speech motor control mechanisms in faster speech. Spatial reduction for consonantal gestures as a function of speech rate depends on consonant identity and phonetic context, and is not a necessary result when speech rate is increased. The tongue body movement velocity tends to become faster in fast speech, whereas tongue tip/blade movement velocity tends to constant across different speech rates. In the acoustic domain, the F2 trajectory for V1 is sensitive to the increased gestural overlapping, but the V1 and C2 identity constrains such sensitivity.

B. Experiment 2

Table II shows the regression results for the kinematic properties of transconsonantal articulatory transition against speech rates. The regression results for vocalic mid-portion linguapalatal contact and same-time F2 values are also included.

Significant displacement is obtained for three out of six /V1#C2V2/ sequences. While the increased maximum linguapalatal contact for V1 in /i#pa/, /i2#pa/ may be ready to explain the observed larger tongue body movement excursion in slow speech, the transconsonantal transition from /i2/ to /i/ is hard to explain because the maximum contact for both vowels is significantly increased in slower speech. A possible explanation is that the near-target realization of /i2/ in slower speech distorts the gestural phase relation of the following syllable, leading to the delayed tongue body gesture against lip closure gesture. This is supported by the larger linguapalatal contact for the first frame of V2 in faster speech compared with that in slower speech.

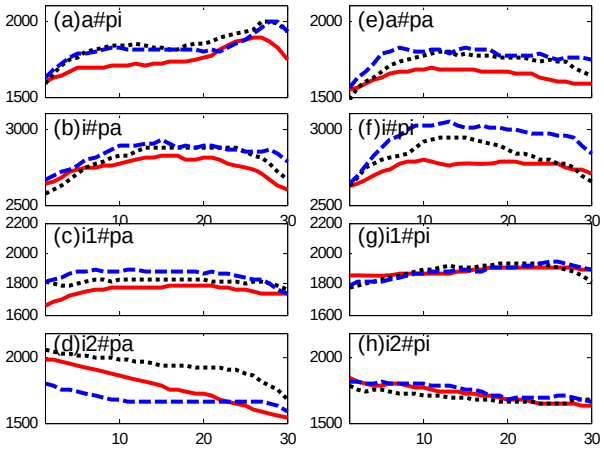


Figure 4. Normalized F2 trajectory of V1 averaged respectively for three speech rates. The solid stands for fast speed, dotted line for normal speed,
and dashed line for slow speed.

The articulatory transition duration is significantly reduced as speech rate is increased. But the shorter vocalic transition does not guarantee the faster articulatory velocity, which in part supports [11]. The significant regression values are inevitable result of transitional duration when the displacement is held relatively constant, as the shorter of the transition is, the faster the articulatory velocity becomes.

The articulatory adjustment toward V2 is initiated earlier in faster speech. The only exception occurs for /i1#pi/, for which the tongue blade gesture of the apical vowel tends to be preserved toward the end of the vowel. This may reflects the differential coarticulatory resistance for /i1/ depending on the tongue height of the transconsonantal vowels. While the channeled tongue body shape is easy to go downward, it is relatively sluggish to respond to the tongue body front rising gesture. The acoustic consequence of earlier initiation of tongue body gesture adjustment toward V2 as speech rate is increased is reflected in the F2 trajectory movement in the left column of Figure 4 (a-d). The F2 trajectory of the pair-wise comparison sequences are also provided in the right column of Figure 4 (e-h). The prime example is /a#pi/ in three speech rates. In Figure 4(a) both the starting point for F2 trajectory upward movement and the time F2 trajectory reaching the peak value are earlier in faster speech than in normal or slow speech. However, the rate-induced increased temporal coarticulation is hard to be observed in other sequences with the flanking vowels associated with differing tongue body gestures.

Articulatory undershooting is more frequent in V1 position as compared with V2 position. The linear relationship between maximum linguapalatal contact and speech rate is found in V1 for /i#pa/, /i1#pa/, /i2#pa/, and /i2#pi/. This can be attributed to the increased anticipatory vocalic effect of the transconsonantal V2. As speech rate is increased, the articulatory reconfiguration toward V2 starts earlier, resulting in undershooting of the V1 target. The maximum contact for /a/ in all sequences is not a

meaningful measure in describing the associated tongue body gesture and thus is not discussed here. Spatial reduction for V2 in fast speech is only found for /i/ in /i2#pi/, while /i/ in V2 position of other sequences tends to achieve the phonetic target in question. Since tongue body gesture is highly constrained for both /i2/ and /i/, it is expected that both targets would be reduced as speech rate is increased.

No simple correlation is found between the MidC and MidF2 for individual vowels. This may be attributed to the coarse correlation between the linguapalatal contact index and the acoustic properties of vowels on the one hand and the small capacity of speech samples on the other.

IV. SUMMARY AND CONCLUSION

The results in the first experiment confirm the finding in [2]. As speech rate is increased, the kinematic properties associated with consonant clusters are proportionally changed. First, not only does the latency between the two gestures in consonant clusters become closer, a linear increase of temporal coarticulation is evidenced. But the gestural overlapping tends to be constrained by the consonant production manner and the lingual specification for flanking vowels. Second, the articulatory duration for consonants is significantly shorter in fast speech, whereas the spatial reduction as a function of speech rate is only found for alveolar stop in the context of flanking vowels with contrasting height. The consistent linguapalatal contact for /t/ followed by /i/ may be the result of the coupling effect of tongue dorsum raising gesture toward the following vowel during the closure interval, which compensates the latent spatial reduction for the preceding /t/. The aerodynamic requirement for the production of alveolar fricative is assumed to be the reason for the stable linguapalatal contact for /s/ across speech rates, which is additionally supported by the relative consistent gestural overlap with C1. The consistent linguapalatal contact for velar nasal is the result of ceiling effect, for full or near full contact in the velar region is achieved. The displacement tends to be strongly correlated with the maximum contact of the consonant in question. The average velocity of linguapalatal contact change for tongue body gesture and tongue tip/blade gesture is different, which may be attributed to the fine articulatory gesture associated with tongue tip/blade [12].

The results in the second experiment shows the similar motor control mechanisms in fast speech: the vocalic articulatory transition shortens in duration and initiates earlier in the V1 interval as speech rate increased. The articulatory displacement is dependent on the maximum linguapalatal contact for V1 whereas the average velocity for tongue body movement tends to be correlated with either articulatory duration or displacement. Phonetic target undershooting is more frequent in V1 position than in V2 position, indicating that the target in the prosodically stronger position tends to resist spatial reduction. Coarse relationship is found between the vocalic mid-point linguapalatal contact and F2 values, which may be attributed to the insufficient vocal tract information as reflected by linguapalatal contact.

The combined results of the two experiments show that articulatory gestural shortening and linear increase of temporal coarticulation are the two motor control mechanisms involved in fast speech. Besides, it is found the transconsonantal vocalic anticipatory effect becomes stronger as speech rate is increased. Spatial reduction is not a necessary result for consonants as speech rate is increased, whereas the tongue body gestural reduction associated with vowels tends to be constrained by prosodic position.

The acoustic consequence of the increased gestural overlapping is shown in the F2 trajectory excursion in the V1 position in both experiments. However, the correlation between the altered linguapalatal contact near the mid-portion of vowel and the same-time F2 values is hard to explain. More speakers and speech samples have to be collected.

In both experiment, the increased gestural overlapping and transconsonantal anticipatory effect are shown to be independent of the suprasegmental tier, which speaks against the tenets of the time structure model. In terms of coarticulatory pattern, the anticipatory effect is shown to be dominated in both articulatory and acoustic domain. Anticipatory coarticulation occurs between neighboring segments as well as transconsonantal vowels. In fastest speech, the gesture associated with V2 shows anticipatory effect on V1 by crossing two intervocalic consonants, which is not manifested in slow speech.

ACKNOWLEDGMENT

The authors want to express gratitude to Jin Zi for her patience and hardworking in pronouncing the utterances. The authors are also grateful for the comments of reviewers.

REFERENCES

[1] T. Gay, "Mechanism in the control of speech rate," Phonetica, vol. 38, 1981, pp. 148-158.

[2] D. Byrd, and C.C. Tan, "Saying consonant clusters quickly," JOP, vol. 24, 1996, pp. 263-282.

[3] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory", in W.J.Hardcastle, and A.Marchal, Eds. Speech Production and Speech Modelling. Kluwer,1990, pp. 403-439.

[4] D.R.Kuehn, and K.L.Moll, "A cineradiogaphoic study of VC and CV articulatory velocities," JOP, vol. 3, 1976, pp. 303-320.

[5] J.E.Flege, "Effects of speaking rate on tongue position and velocity of movement in vowel production," JASA, vol. 84, 1988, pp. 901-916.

[6] S.J.Moon, and B.Lindblom, "Interaction between duration, context, and speaking style in English stressed vowels," JASA, vol. 96, 1994, pp.40-55.

[7] C.Browman, and L.Goldstein, "Tiers in articulatory phonology, with some implications for casual speech," in Papers in Laboratory Phonology I, J.Kingston and M.E.Beckman, Eds. Cambridge University Press, 1990, pp. 341-376.

[8] D.Recasens, M.D.Pallares, and J.Fontdevila, "A model of lingual coarticulation based on articulatory constraints," JASA, vol. 102, 1997, pp. 544-561.

[9] L, Deng, Dynamic Speech Models, Theory, Algorithms, and Applications, Morgan & Claypool, 2006.

[10] Yi Xu, and Fang Liu, "Tonal alignment, syllable structure and coarticulation: toward an integrated model," Italian Journal of Linguistics, vol. 18, 2006, pp. 125-159.

[11] T. Gay, T. Ushijima, H. Hirose, and F.S.Cooper, “Effect of speaking rate on labial consonant-vowel articulation,” Status Report on Speech Research, SR35/36, pp. 7-30.

[12] J.S.Perkell, Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. Cambridge, MA: The MIT Press,1969.

Effects of Prosodic Boudnaties on Segment Articulation in Standard Chinese

Yinghao Li, Jiangping Kong

Phonetics Lab, Peking University, Beijing, China
leeyoungho@pku.edu.cn; kongjp@gmail.com

ABSTRACT

This study investigates the effect of prosodic boundary on the articulation and the acoustic properties of consonant and vowel segments, and vowel-to-vowel coarticulation across five prosodic boundaries in the Standard Chinese. The articulatory measures are obtained by electro-palatography. The result shows that the domain-initial consonant /t/ is strengthened in a hierarchical fashion for prosodic boundaries above Foot; however, no difference is found between Syllable and Foot domain. For /i/, differential influences are found for the left boundary of the syllable where the vowel is in and the right boundary immediately after the vowel. The last result shows that the vocalic anticipatory effect is constrained both by prosodic factor and articulatory constraint of vocalic gesture, but the vocalic carryover effect is more likely conditioned by Foot constraint.

Keywords: electroplatographic; prosodic structure; Standard Chinese; V-to-V coarticulation

1. INTRODUCTION

Increasing articulatory studies have shown that prosodic structure modulates segment articulation in a hierarchical fashion in various languages [1,3, 6]. The two prime prosodic effects on segment articulation are domain-initial consonant strengthening, which might spread to the whole syllable [2], and domain-final lengthening. Prosodic structure also constrains the inter-segmental coarticulation [7].

In the Standard Chinese, the pitch contour and pause are the main apparatus for marking the higher prosodic boundaries, but the foot boundary rely less on the two acoustic cues [10]. In [10] it is conjectured that the articulatory and acoustic correlates for marking foot boundary may be related to the consonantality of the domain-initial consonant in that it is more consonant-like at the foot-initial position than within-foot syllable-initial position. It is also hypothesized that the normal bisyllabic foot in the Standard Chinese is phonetically characterized by the closeness of segmental and suprasegmental articulation within the foot. While the suprasegmental pattern for foot is stacked with a large literature, the articulatory evidence at the segmental tier has been rare.

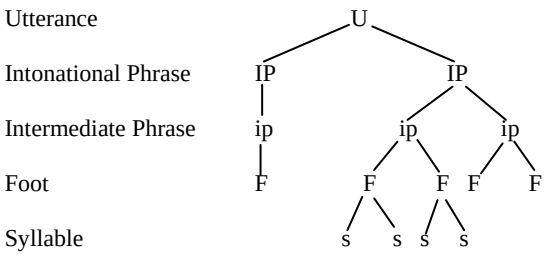
The paper aims at finding out the prosodic signatures in articulation and acoustic properties of segment and inter-segmental coarticulation in the standard Chinese. Specifically, two issues are to be addressed: firstly, does the prosodic structure modulate the segment production in the hierarchical fashion? Secondly, is the vowel-to-vowel coarticulation constrained by prosodic domains?

2. METHODOLOGY

2.1. Prosodic structure

A five-level prosodic hierarchy [5] is adopted in this study and is shown in Figure 1. While utterance and intonational phrase are unanimously agreed upon, the intermediate phrase is hard to distinguish from foot boundary phonetically. In [10] the intermediate phrase is encoded as a smaller pause whereas the foot boundary a potential pause. The foot domain, normally comprised of two syllables, is described as tightly combined phonological unit, for lexical tones is not a prominent feature in foot domain in the Standard Chinese. Besides the tone sandhi group is not a prosodic domain in the Standard Chinese [8].

Figure 1: A five-level prosodic hierarchy for Standard Chinese.



2.2. Speech material

The test segments were unaspirated /t/ and high front vowel /i/, for both have considerable linguapalatal contact. The low vowel /a/ was used to investigate vowel-to-vowel coarticulation. The four bisyllable sequences, /ta#ti/, /ta#ta/, /ti#ta/, and /ti#ti/, were constructed where # stands for five prosodic boundaries. When the sequences were separated by a Syllable boundary, higher boundaries were placed before the first syllable. A total of 32 sentences were designed. The last syllable before the test bisyllable sequences was always low vowel /a/ and the first syllable after the sequences was not strictly controlled. The tonal specification of the test bisyllable sequences was not controlled.

2.3. Recording

The recording was taken in a sound-attenuated booth at the Phonetic Laboratory of Shanghai Normal University. The sentences were repeated for three times and divided into six sentence blocks in random, with two dummy sentences appeared in the first and last position in each block. One female speaker was presented with six blocks of sentences and was instructed to read the sentences in normal speed. The electropalatographic signal was recorded with a sampling rate of 100 Hz, and the speech signal 22050 Hz.

2.4. Measurements

The WinEPG electropalatography was used in the study. The maximum linguapalatal contact of the mid-portion three frames of the alveolar closure was measured (MaxC). The alveolar seal duration was defined as the interval from the first to last frame that showed alveolar closure (ASD). The acoustic silent duration was the interval from the last pitch pulse to the energy burst for stop (AD). The maximum contact for /i/ was the maximum linguapalatal contact of the frames between the one third and one half of the interval into the vowel (MaxV). The center of gravity (CoG) of the linguapalatal contact is defined following [8]. The vocalic duration and mid-portion F1 and F2 were also measured.

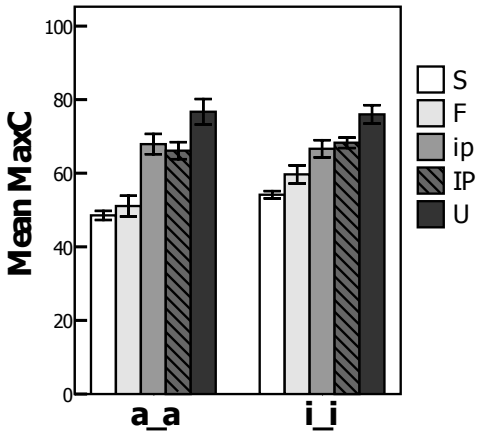
For vocalic anticipatory effect, the linguapalatal contact of the last four rows of electrons in the final frame was measured for V1 (POS_E), and the V1-end F2 (F2_E) was first derived from LPC with an order of 20, and later manually adjusted. For vocalic carryover effect, the V2 start time frame linguapalatal contact (POS_I) and the corresponding F2 for V2 was measured (F2_I).

3. RESULTS

3.1. Domain-initial consonant strengthening

Figure 2 shows the maximum linguapalatal contact for alveolar consonant /t/ at five prosodic domain-initial positions when the contextual vowel was controlled.

Figure 2: Maximum contact of /t/ at five prosodic domain-initial positions. Error bar refers to one standard error.



The overall contact data shows that the prosodic domains are generally distinguished by the speaker in both vocalic contexts ($p<0.001$), except that the ip-initial has more contact than IP-initial in /a_a/ context. In the symmetrical /a/ context the MaxC is increased in the order of S, F<ip, IP<U ($p<0.05$, Turkey HSD), and in the symmetrical /i/ context in the order of S<F<ip, IP<U ($p<0.05$, Turkey HSD). The mean seal duration, as shown in Figure 3, is increased progressively in higher prosodic boundary in the order of S, F<ip<IP<U in both vocalic contexts ($p<0.05$, Turkey HSD). The AD shows the similar hierarchical pattern.

The prosodic domains above Foot are distinguished consistently by ASD and AD, indicating the pausing is the prominent cues for marking prosodic boundaries in higher prosodic domains. However, the distinction between Syllable and Foot is not consistent. The only distinction appears in the symmetrical /i/ context, which might be attributed to the high rising tongue body gesture interfering with tongue tip closure gesture in a short interval of time, which does not appear in symmetrical /a/ context.

Figure 3: Alveolar seal duration of /t/ at five domain-initial positions.

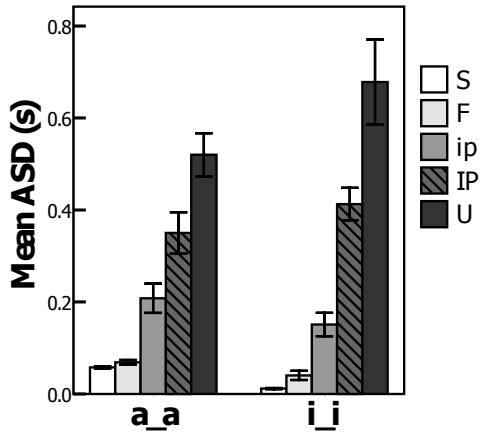
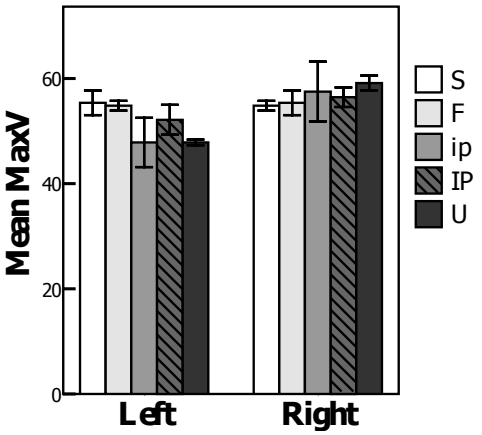


Figure 4: The maximum contact for /i/ at five left and right boundaries.



3.2. Articulatory and acoustic properties of /i/

As is shown in Figure 4, the linguapalatal contact for /i/ varies differentially with the left boundaries immediately before the syllable /ti/ and the right prosodic boundaries immediately after the vowel. The strengthened domain-initial consonant tends to lead to reduced vocalic gesture with lesser linguapalatal contact in higher prosodic domains, though no significant difference is found for MaxV and CoG. The duration tends to be longer in higher prosodic domains, but there is no significant difference, which might be attributed to the intrinsic vowel duration. The F1 differs significantly across five prosodic domains ($p=0.05$), with the mid-point F1 for Syllable and Foot domains significantly lower than higher prosodic domains ($p<0.1$), which is indicative of larger supra-laryngeal opening.

The linguapalatal contact tends to be increased progressively with ascending right prosodic boundaries, but still no significant difference props up. Neither does the mid-frame CoG have the significant difference. The acoustic properties show significant difference in vowel duration ($p=0.001$) and mid-portion F1 ($p=0.001$). The vowel duration in Syllable and Foot domains is significantly shorter than in higher prosodic domains, indicating that final-lengthening is a prominent cue for marking higher prosodic boundaries. The mid-point F1 is increased in the order of S, F<IP, U ($p<0.05$), and S<ip<IP, U ($p<0.05$). This result indicates that the oral cavity might have a larger opening in higher prosodic boundaries.

3.3. Vowel-to-vowel coarticulation

Because the lexical tone plays no meaningful role in the Standard Chinese, it is hypothesized that the segments within the lower prosodic domains show more gestural overlap. To this ends, the vowel-to-vowel coarticulation is investigated by comparing the non-symmetrical and symmetrical /V1#tV2/ sequences, where # denotes five prosodic boundaries.

Figure 5: The mean POS_E and F2_E for V1/a/ in sequences /a#ta/ and /a#ti/ across five prosodic boundaries. (* for $p<0.05$)

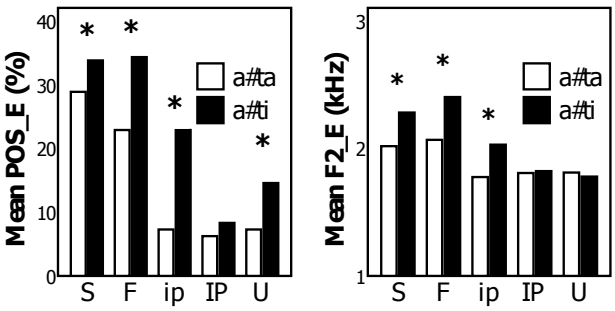
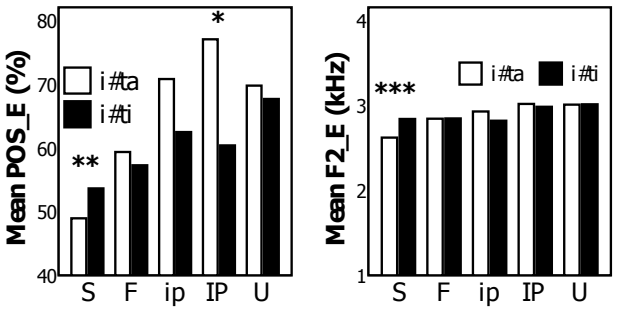


Figure 5 shows the comparison of POS_E and F2_E for V1/a/ between the /a#ti/ and /a#ta/. The one-way ANOVA for the posterior contact area of the last frame for V1 /a/ and the final F2 shows significant difference between the sequences of /a#ta/ and /a#ti/ below ip boundary, indicative of significant vocalic anticipatory effect. The unexpected significant difference at Utterance boundary is not the result of the transconsonantal high front vowel, for the large time interval between the V1 and second syllable make the transconsonantal effect impossible to happen.

Figure 6 shows the comparison of POS_E and F2_E for V1/i/ between the /i#ta/ and /i#ti/. The

results of one-way ANOVA indicate that transconsonantal V2 effect only occurs at Syllable boundary articulatorily and acoustically, which is reflective of the strong articulatory constraints for high front vowel [9].

Figure 6 The mean POS_E and F2_E for V1/i/ in sequences /i#ta/ and /i#ti/ across five prosodic boundaries. (** for p<0.01, and *** for p<0.001)



In Figure 7-8 the transconsonantal V1 carryover effect is investigated articulatorily and acoustically. A series of one-way ANOVA results indicates that the domain for carryover effect is within the Foot domain. The only exception is that the effect of V1 /i/ on the posterior contact for the first frame of V2 /a/ is hard to measure, however, the significant difference on initial F2 between /a#ta/ and /i#ta/ is reminiscent of tongue body lowering gesture initiated from the achievement of the tongue body high raising gesture.

4. SUMMARY AND CONCLUSION

The first result of the paper is that the domain-initial consonant strengthening does distinguish the higher prosodic boundaries from lower ones. However, the consonant production at the initial position of Syllable and Foot does not differ in strength, which goes against the hypothesis in [10]. The second result finds that the prosodic boundary at the left of the consonant preceding the vowel in question and the right boundary immediately after the vowel affect the vocalic gesture in differential manners. The higher degree of the left boundary, the more reduced for the vocalic gesture compared with hierarchically strengthened consonantal gesture; however, the vocalic gesture tends to progressively reinforced as the right prosodic boundary goes up. The third result is that vocalic anticipatory effect is constrained by prosodic factor and the articulatory constraints for the vocalic gesture in question. However, the vocalic carryover effect is more likely to be constrained by the prosodic factor.

Figure 7 The mean F2_I for V2/a/ in sequences /a#ta/ and /i#ta/ across five prosodic boundaries.

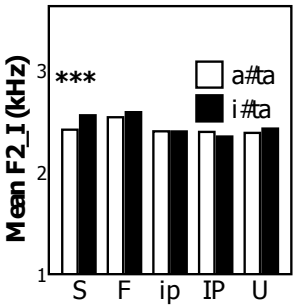
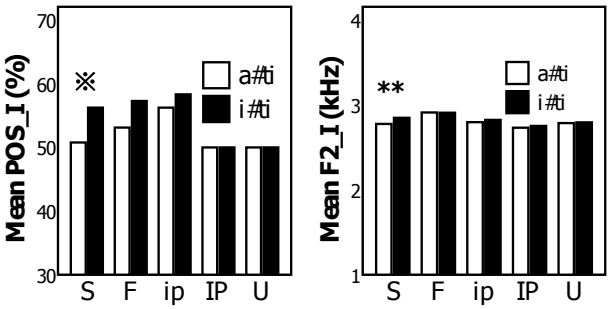


Figure 8 The mean POS_I and F2_I for V2/i/ in sequences /a#ti/ and /i#ti/ across five prosodic boundaries.（※ for p=0.09)



5. REFERENCES

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

[2] Cho, T., Keating, P. 2009. Effects of initial position versus prominence in English. *Journal of Phonetics* 37, 466-485.

[3] Fougeron, C. 2001. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics* 29, 109-135.

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

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

[6] 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.

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

[8] Kuang, J.J., Wang, H.J. 2006. An acoustic study on the third tone sandhi across various prosodic boundaries. *Proc. 7th PCC Beijing*.

[9] Recasens, D, Pallares, M.D., Fontdevila, J. 1997. A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America* 102, 544-561.

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

The Relation between Larynx Height and F0 during the Four Tones of Mandarin in X-ray Movie

Gaowu WANG, Jiangping KONG
Phonetics Lab, Peking University, Beijing
wanggaowu@pku.edu.cn; jpkong@pku.edu.cn

Abstract— **The relation between vertical larynx movement and vocal frequency (F0) change has attracted the attention of many researchers. This paper studies the vertical laryngeal position (larynx height) during the four tones of Mandarin, based on the X-ray movie data from one female speaker, with 36 monosyllables in tones. The location and movement of the larynx have been measured, and compared with the variation of the F0 during the four tones. Results show that, during the first tone, the larynx height is weakly correlative with F0, while during the other three tones the larynx height is positively correlative with F0. This suggests that the upward movement of larynx is partially (in tone 1) independent of F0, while the downward movement is accompanied by decreasing F0. And the quantitive mechanism of larynx height and F0 during the four tones can be integrated into articulatory model of vocal tract in Mandarin for better speech synthesis.**

Keywords- laryngeal position; tones; Mandarin; X-ray

I. INTRODUCTION

It is well-founded that the vertical laryngeal position (larynx height) is related to the voice fundamental frequency (F0), which roughly means that the larynx moves up and down as F0 rises and falls[1]. This suggests that vertical movement of the larynx is a critical component of F0 control mechanisms, and therefore the relation between vertical larynx movement and F0 change has attracted the attention of many researchers [1-7].

On the other side, in articulatory model during F0 fluctuation, especially for those tonal languages, should the larynx height be adjusted accordingly, which will lead to the change of the vocal tract length? In other words, when the F0 of the synthesized speech rises and falls, should we adjust the larynx height? This will change the length of vocal tract, and accordingly influence the acoustic characteristics of synthesized speech. Specifically, what kind of quantitive mechanism of larynx height and F0 should be integrated into articulatory model for synthesizing tonal speech? This quantitive mechanism during the four tones will bring more precise vocal tract length when synthesizing tonal speech in Mandarin. From engineering point of view, all these also require us to explore the relation between larynx height and F0 [8-10].

This study is sponsored by National Project of Scientific and Technical Supporting Programs Funded by Ministry of Science & Technology of China （NO. 2009BAH41B00）.

However, the empirical observation is difficult because the larynx hides in human body. Previous measurements of larynx position have employed various techniques using optical instruments[2, 4], mechanical facilities[5], X-ray devices[6, 11]), and Magnetic Resonance Imaging (MRI)[1,7,12]. Although MRI can give full volume vocal tract images, it is limited to a static, sustainable configuration. The X-ray permits real time tracking, so the full sagittal view of vocal tract including larynx during running speech provided by X-ray movie (cineradiography) remains unsurpassed, although the poor contrast of the cartilages in lateral cineradiography makes it a little technically difficult in measuring laryngeal movements

In the literature, there is scarce data of laryngeal movement in Mandarin, a tonal language with four tones: *yin ping*, *yang ping*, *shang*, *qu*. Therefore, in this study, the vertical laryngeal movement during the four tones of Mandarin has been observed and investigated, to explore the relation between the vertical movement and the F0 during Mandarin tones.

II. METHOD

A. X-ray movie

An X-ray movie database has been established from the original PAL videotape, which is the only cineradiography video of Mandarin available at present. Both sagittal cineradiography and speech sound are given simultaneously. A female speaker, who is a national broadcaster of Mandarin, pronounced 9 sets of monosyllables with different consonants and vowels in the 4 tones, totally 36 monosyllables in their corresponding characters. The speaker was required to articulate each syllable clearly and rest sufficiently between every two syllables, to avoid the interference of articulatory movements. The monosyllables and their corresponding characters in Mandarin are listed in table 1.

TABLE I. THE MONOSYLLALES AND THEIR CORRESPONDING CHARACTERS IN FOUR TONES OF MANDARIN

	Tone 1	**Tone 2**	**Tone 3**	**Tone 4**
	yin ping	***yang ping***	***shang***	***qu***
ba	巴 ba1	拔 ba2	把 ba3	爸 ba4
bi	逼 bi1	鼻 bi2	比 bi3	闭 bi4
bu	逋 bu1	醭 bu2	补 bu3	布 bu4
lao	捞 lao1	劳 lao2	老 lao3	涝 lao4

liao	撩 liao1	辽 liao2	了 liao3	料 liao4
qian	千 qian1	前 qian2	浅 qian3	欠 qian4
qiang	枪 qiang1	强 qiang2	抢 qiang3	呛 qiang4
zhai	摘 zhai1	宅 zhai2	窄 zhai3	债 zhai4
chuai	揣 chuai1	膗 chuai2	揣 chuai3	踹 chuai4

This video has been transferred to avi files at 25 frames per second (fps), which is the same as original videotape. Next, the avi files are adjusted to make sure that the PAR (Pixel Aspect Ratio) equals 1, which means each pixel in the image is square, to avoid vertical or horizontal distortion. Thereafter, the avi files are cropped with a window of 400*400 pixels, to focus on the movements of articulators and to alleviate the processing burden. Finally, the avi files are segmented to single syllable.

The original speech sound is recorded simultaneously in the X-ray device room, where there exists some kind of noise. However, this will bring little influence to the detection of F0. The audio format is as follows: 44.1 kHz sampling rate, 16bits quantization, 1 channel, PCM wav file.

B. X-ray movie processing

A platform 'VocalMarker' in Matlab is programmed to trace the mid-sagittal articulatory movements, including the preprocessing of the images, the semi-automatic tracing of the articulatory movements, and the speech sound processing. Figure 1 is the demo of the 'VocalMarker', in which a syllable sample /biao1/ is being processed. The upper window shows the marking of articulatory movement, in which there are several lines with key points to mark the edges of articulators. The shape of larynx is marked by two lines 'larynx front wall' and 'larynx back wall'. The articulators are marked in each frame of the X-ray movie, so finally the movement of the larynx can be obtained. The lower window shows the processing of synchronous speech sound, where the waveform and spectrogram of current syllable are shown. At the same time the acoustic parameters including formant frequency and F0 are calculated.

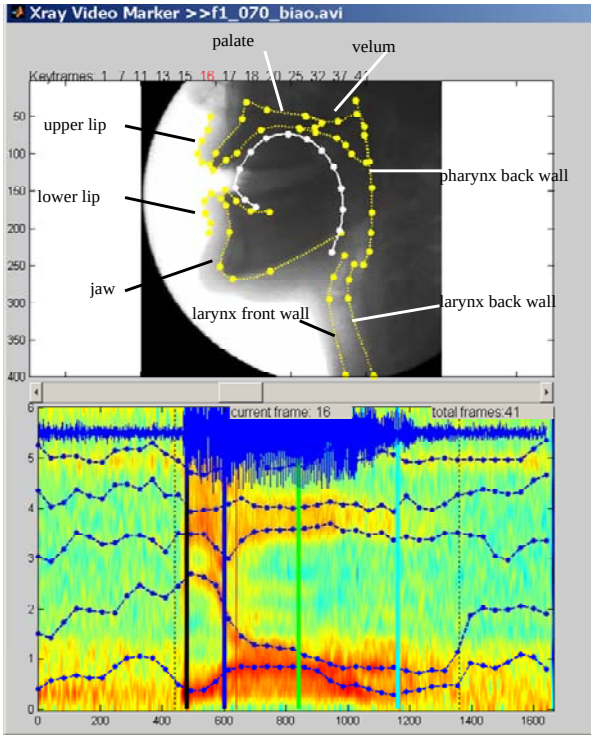


Figure 1. The demo of 'VocalMarker'

C. Larynx height measuring

Due to the poor contrast of the cartilages in lateral X-ray movie, we adjust the contrast of the movie, trying to make a precisely positioning of the larynx. Figure 2 shows the measuring of the larynx height. In the X-ray movie, the images of the cartilages are blurry, so the whole shape of the cartilages are not clearly shown. However, we can position the vocal fold as the maximum curvature along the larynx front wall, where is also the most protrudent place of the cricoid cartilage.

The original larynx height is defined as the vertical distance between the vocal fold (the third key point along the larynx front wall) and the lowest place of trachea in this video (the forth point along the larynx front wall). We set the original larynx height during speechless rest as the base larynx height. And finally the larynx height is defined as original larynx height minus the base larynx height. The higher value of larynx height indicates the higher position of vocal fold, and accordingly the shorter length of the vocal tract when other articulatory configurations remain the same. If the value of larynx height is minus, it means that the position of vocal fold is lower than its position during speechless rest.

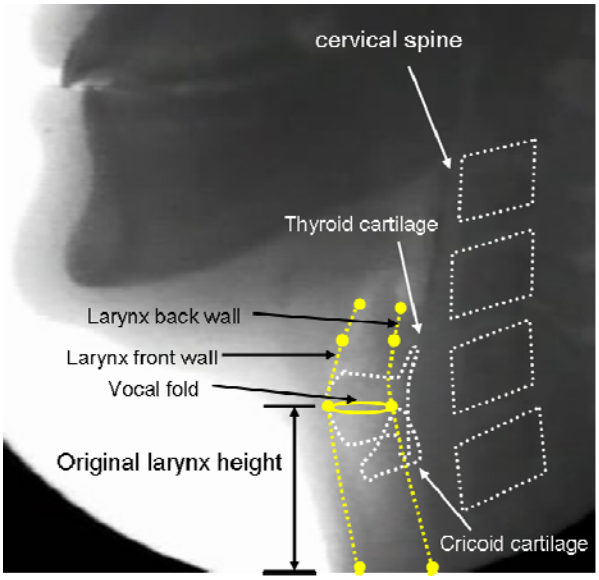


Figure 2. The measuring of the larynx height

III. RESULTS

The results are observed and discussed in two ways: the movement of the vertical laryngeal position and the relation between larynx height and F0.

A. The movement of the vertical laryngeal position

Figure 3 shows the movement of the vertical laryngeal position and fluctuation of the F0 during the four tones of the 9 sets of syllables. The x-axis is time in frames, and 1 frame is equivalent to 40 ms, since the fps is 25. The y-axis is F0 in Hz and larynx height in 0.1 mm at the same time for showing them both in the same figure. The red stars are F0 data, which show that the fluctuant F0 curves during the four tones: tone 1 (*yin ping*), tone 2 (*yang ping*), tone 3 (*shang*), tone 4 (*qu*), from left to right along the time axis. The starts and ends of every syllable are marked by eight vertical dashed lines. The blue points are larynx height data, which shows the rises and falls of the larynx during the four tones with 9 sets of syllables.

As shown in this figure, there usually is a lowering movement of larynx before each syllable with the decreasing larynx height. Moreover, some of the values of larynx height are minus, which means that the larynx goes lower than its position during speechless rest. This indicates the preparing action before the speaker pronounces each syllable. As we can notice, the 9 sets of syllables have different larynx height because of their different vowels, where the vowel /i/ has the maximum larynx height and the /u/ has the minimum one. However, for the moment, the intrinsic F0 and larynx height of different vowels are not considered.

As we can see that, during the tone 1, the larynx moves up while the F0 already starts at a high level around 250Hz and maintains flat, which means the speaker can produce a high tone without lifting up the larynx to a certain height necessarily. During the other 3 tones, basically, the larynx height moves up and down as F0 rises and falls.

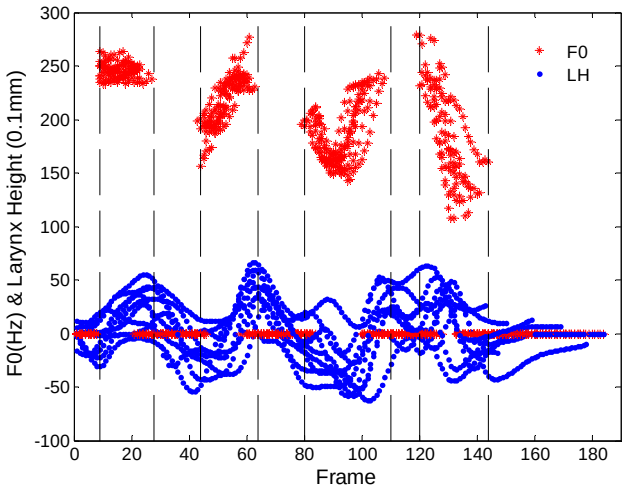


Figure 3. The fluctuant curves of larynx height and F0 during the four tones

B. The realtion between the larynx height and F0

Having observed the fluctuation of larynx height and F0 in the time domain, we may explore whether it exists a quantitive correlation between them during the tone 2, 3, and 4, and what kind of mechanism between the larynx height and F0 we can adopt into the articulatory model in Mandarin.

Figure 4 shows the result. There are four sub-windows in which the results during four tones are shown respectively. The x-axis is F0, ranging from 100 to 300Hz, as a typical female speaker's voice. The y-axis is larynx height, ranging from -6 mm to 6 mm. The different colors and symbols of the points represent the 9 sets of syllables with different consonants and vowels. Still the same, we do not consider the intrinsic F0 of vowels, so we just study the relation between the larynx height and F0.

In the first sub-window (from the left), during tone 1, when the F0 maintains at the same level around 250Hz, the larynx moves up. The correlation analysis result shows that the correlation factor between F0 and larynx height is 0.0274, which indicates they are weakly relevant, or nearly irrelevant.

During tone 2, when F0 rises, the larynx moves up, as indicated by the red line arrow. And the correlation factor between them is 0.5876. Therefore we use linear regression analysis to get the relation between larynx height (LH) and F0: LH = 0.6071*F0 - 135.9667.

During tone 4, when the F0 falls, the larynx moves down, as indicated by the red line arrow. And the correlation factor between them is 0.4639. Therefore we use linear regression analysis to get the relation between them: LH = 0.3055*F0 - 44.2634.

During tone 3, the relation is complicated. At first, the F0 falls from 200Hz to 150Hz, while larynx moves down, and we can use the equation during tone 2 to describe it. And then, the F0 remains low, but the larynx still moves down or up. Here the relation between the larynx height and F0 is still unclear since the phonation type of vocal fold may be different from modal voice. Finally, the F0

rises from around 150Hz to 250Hz, while the larynx moves up, and we can use the equation during tone 2 to describe it.

The changing F0 will bring a fluctuant range from -6 mm to +6 mm to larynx height, which would lead to the varying vocal tract length. The typical length of the whole vocal tract is about 150mm for an adult female speaker, so in articulatory model this will bring an 8% range of vocal tract adjustment during tonal speech synthesizing, thus may be beneficial to Mandarin TTS (Text to Speech).

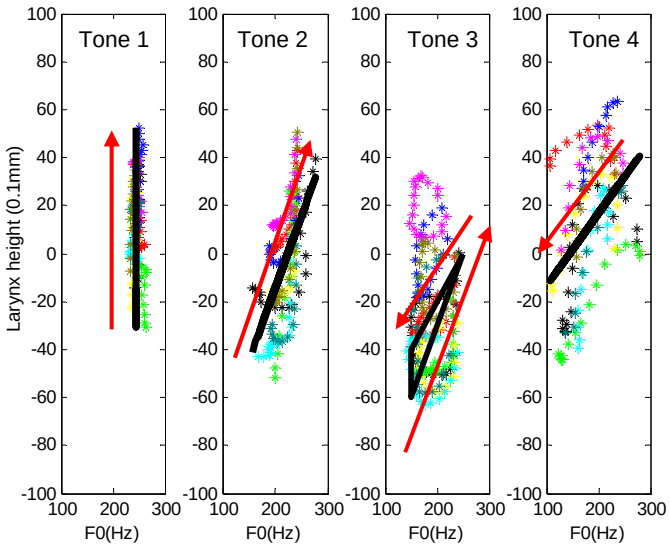


Figure 4. The relation between larynx height and F0 during the four tones

IV. CONCLUSION

As to the movement of the vertical laryngeal position and the relation between larynx and F0 during the four tones in Mandarin, there are mainly two different kinds of results: During tone 1, the vertical position of larynx is not correlative with F0. The F0 stays at a high level while the larynx moves up from its rest position, which means the speaker can produce a high tone without lifting up the larynx to a certain height necessarily. During the other 3 tones, the larynx position is positively correlative with F0, which roughly means that the larynx moves up and down as F0 rises and falls. Two equations can describe how the larynx height changes when F0 changing, and the larynx height changes at different ratio to F0 when moving up or down.

The results suggest that the upward movement of larynx is partially (in tone 1) independent of the F0, while the downward movement is accompanied by decreasing F0. This may be explained by the mechanism of the larynx structure, which still needs clarifying unfortunately. One drawback of the study is that the data size is scarce due to the radioactive harm of X-ray. Only a single speaker's database containing 36 isolated syllables is used. This may need more empirical data and further study to explore the underlying mechanism controlling the movement of larynx and the change of F0. However, from the viewpoints of engineering, the quantitive mechanism of larynx height and F0 during the four tones can be integrated into articulatory model of vocal tract in Mandarin, which will bring more precise vocal tract length when synthesizing tonal speech in Mandarin.

ACKNOWLEDGMENT

We give our devout appreciation to Prof. Huaiqiao BAO for providing the original videotape of the X-ray movie.

Moreover, we here express our gratitude to the paper reviewers for their scrupulous work.

REFERENCES

[1] K. Honda, H. Hirai, S. Masaki, and Y. Shimada, "Role of vertical larynx movement and cervical lordosis in F0 control," *LANGUAGE AND SPEECH*, vol. 42, pp. 401-411, 1999.

[2] Y. Kakita and S. Hiki, "Investigation of laryngeal control in speech by use of thyrometer," *Journal of the Acoustical Society of America*, vol. 59, pp. 669-674, 1976.

[3] T. Chiba and M. Kajiyama, *The Vowel: Its Nature and Structure*. Tokyo, 1942.

[4] W. Ewan and R. Krones, "Measuring larynx movement using the thyroumbrometer," *Journal of Phonetics*, vol. 2, pp. 327-335, 1974.

[5] J. Gandour and Maddieson, "Measuring larynx movement in Standard Thai using the cricothyrometer," *Phonetica*, vol. 33, pp. 241-267, 1976.

[6] J. Lindqvist and M. Sawashima, "An investigation of the vertical movement of the larynx in a Swedish speaker," *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, vol. 7, pp. 27-34, 1973.

[7] H. Hirai, K. Honda, I. Fujimoto, and Y. Shimada, "Analysis of magnetic resonance images on the physiological mechanisms of fundamental frequency control," *Journal of the Acoustical Society of Japan*, vol. 50, pp. 296-304, 1994.

[8] G. WANG, "An Articulatory Model of Vocal Tract in Mandarin," Doctor Dissertation, Department of Chinese Linguistics and Literature, Peking University, Beijing, 2010.

[9] P. Mermelstein, "Articulatory model for the study of speech production," *The Journal of the Acoustical Society of America*, vol. 53, pp. 1070-1082, 1973.

[10] S. Maeda, "Articulatory Modeling of the Vocal-Tract," *Journal De Physique Iv*, vol. 2, pp. 307-314, Apr 1992.

[11] T. Shipp, " Vertical laryngeal position during continuous and discrete vocal frequency change," *Journal of Speech and Hearing Research*, vol. 18, pp. 707-718, 1975.

[12] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *Journal of the Acoustical Society of America*, vol. 90, pp. 799-828, 1991.

TIME SERIES ANALYSIS OF JITTER IN SUSTAIN VOWELS

Li Dong

ABSTRACT

This paper proposes a new equation to show the time domain features of jitter: Jitter = $(T_{i+1}-T_i)/T_i$. The subjects are tonal language speakers. Jitter is extracted from EGG signals. The results show that: (1) jitter does not obey Gaussian distribution but multimodal distribution; (2) the first peak position near zero of the jitter distribution curve increases with the pitch; (3) each peak position is an integral multiple of the first peak position; (4) jitter values are around 0, and in most cases, adjacent values have opposite algebraic sign; (5) the zero value percentage and zero-crossing rate of jitter increase with the pitch; (6) the variance of pitch keeps in the allowance of just noticeable difference, which can be attributed to the effect of audible feedback, and it affects the variance of jitter further.

Keywords: jitter, time domain features, distribution

1. INTRODUCTION

Jitter involves small fluctuations of the glottal cycle lengths. Lieberman was the first to compare the small cycle-to-cycle fluctuations in the fundamental periods of healthy and diseased larynxes [3]. From then on, most studies of jitter focus on using jitter value to distinguish between healthy voice and pathological voice. Another function of jitter is to make the synthesized voice much livelier. Previous studies of jitter can be divided into two categories, mean analysis and time series analysis. Studies using time series analysis method have a small number. Because in time domain, each pitch value is around the mean value, most studies use Gaussian distribution to simulate the distribution of jitter [2]. In [3, 4], jitter value was broken down into predictable and random components.

This article focuses on the time domain features of jitter, and a new equation is proposed to show them. The subjects are native Chinese speakers, for previous studies chose non-tone languages speakers whose jitter may be different from tone languages speakers. In this article, the distribution of jitter is discussed, and jitter values are broken down into predictable and random components.

Previous studies used different equations to calculate jitter. Some couldn't reflect time domain features; some were affected by the pitch; some used the mean value on the local scale. Therefore, a new equation was proposed to show the time domain features of jitter.

$$(1)\quad \text{Jitter} = \frac{T_{i+1}-T_i}{T_i}$$

This equation is free from the pitch effect and can show the local fluctuations exactly.

2. METHODS

Two male subjects and two female subjects participated the recording. They are all college students and the average age is 25. The recordings were made in a sound-proof room. The sample rate is 44.1 kHz and the resolution is 16 bit. The left channel is speech signal and the right channel is EGG signal. Jitter is very small, so that the sample rate is very important. In [2], the average jitter extracted in 40 kHz is close to that extracted in 80 kHz. So the jitter value in this article is reliable.

Some recorded sounds were played back to the subjects and they pronounced with the same pitch. The target frequencies are 95Hz, 113Hz, 131Hz, 149Hz and 167Hz for male subjects and 180Hz, 210Hz, 240Hz, 270Hz and 300Hz for female subjects. Before the test, the subjects' lowest and highest sounds were recorded, and then each step's value was calculated. Each subject pronounced /a/, /i/, /u/ in each frequency twice.

Jitter was extracted from EGG signals. EGG signals are more pure than speech signals, for speech signals contain more information such as formant information. The differentials of the EGG

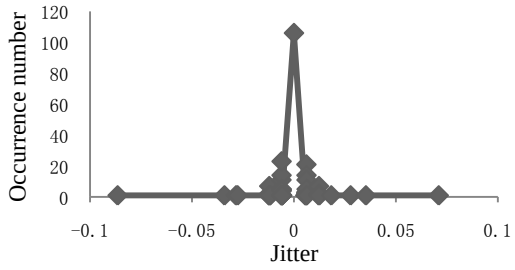
signals were calculated and each peak of the differentials was marked as the initial point of each cycle. Then jitter sequences were calculated.

3. RESULTS

3.1. The distribution of jitter

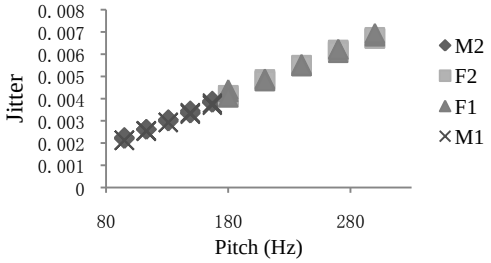
Jitter does not obey Gaussian distribution but multimodal distribution. The peaks of the multimodal distribution are the foundation of producing jitter. Figure 1 is a typical distribution of jitter. It is /a/ in 270Hz pronounced by a female subject, and it is multimodal distribution. The distribution is symmetrical and the occurrence number peaks when jitter value is equal to zero.

Figure 1: The distribution of jitter.



Jitter is related to pitch. The number of the peaks is irrelevant to pitch, but the first peak position near zero is positively correlated with pitch. That is to say, the first peak position near zero increases with the pitch. So jitter calculated through equation 1 is relative that the positive correlation is not simply caused by the increase of the minuend and subtrahend. Figure 2 shows the first peak positions near zero of /a/, /i/, /u/ in five frequencies of four subjects. Because the distributions are symmetrical, the average absolute value of the left and right first peak positions near zero was used to plot Figure 2. Figure 2 shows that the first peak position near zero is positively correlated with the pitch, and it is not affected by vowels and subjects.

Figure 2: The first peak position near zero of the distribution of jitter



Another result is that each peak position is an integral multiple of the first peak position. Table 1 shows each peak's position and the occurrence number in the jitter distribution of /a/ in 270Hz pronounced by a female subject. The second peak position is twice the first peak position and the third peak position is three times it. Though the left and right fourth peak positions are not equal, they are all at integer multiples of the first peak position.

Table 1: The jitter distribution of /a/ in 270Hz

Jitter value	The occurrence number
-0.307	1
-0.093	1
-0.077	1
-0.018	2
-0.012	14
-0.011	1
-0.006	74
0	88
0.006	73
0.012	10
0.018	1
0.019	1
0.036	1
0.103	1

3.2. The time domain features of jitter

Figure 3 shows the time domain features of jitter of /a/ in 300Hz pronounced by a female subject and Figure 4 is the features of /a/ in 95Hz pronounced by a male subject. The ordinate represents jitter

value and the abscissa represents the ordinal of the cycles. In these two Figures, all jitter values are around 0, and in most cases, adjacent values have opposite algebraic sign. In Figure 3, most jitter values are near 0.0068; in some cycles, the jitter values jumped to about 0.0137 in random. In Figure 4, the jitter values are irregular compare to Figure 3. But, adjacent values still have opposite algebraic sign in most cases. By observing all the data, the pitch instead of gender is the key factor that made the curves change.

Figure 3: The time domain features of jitter of /a/ in 300Hz

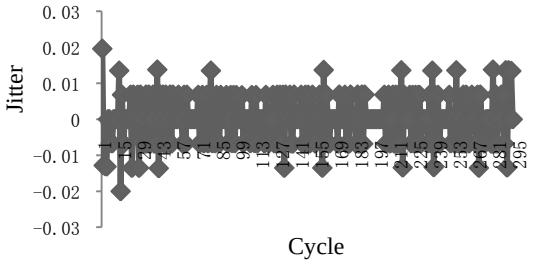
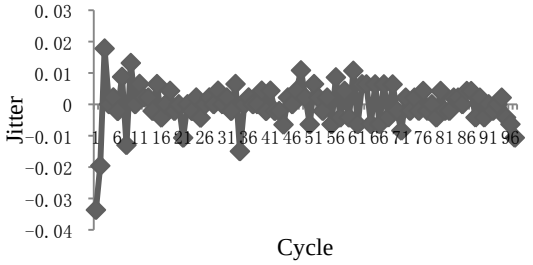


Figure 4: The time domain features of jitter of /a/ in 95Hz



Two parameters are introduced to show the trends of jitter varying with the pitch. The zero value percentage is the percentage of zero of all the jitter values. The zero-crossing rate of jitter is the proportion of the number that adjacent values have opposite algebraic sign of all the cycles. Figure 5 and 6 show that the zero value percentage and zero-crossing rate of jitter increase with the pitch (A few exceptions on the local scale). The range of zero value percentage is from 10% to 50% and the range of zero-crossing rate of jitter is from 0.6 to 1.

Figure 5: The zero value percentage

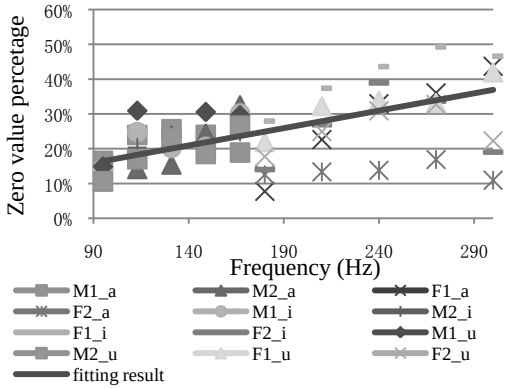
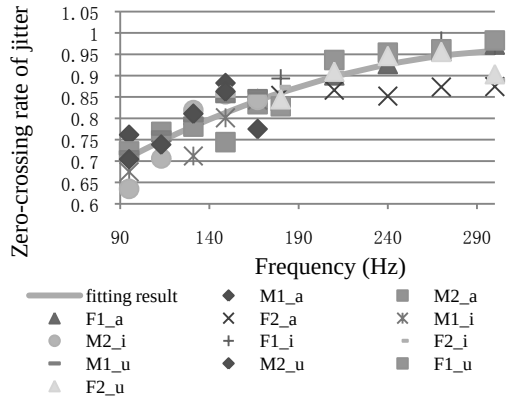


Figure 6: The zero-crossing rate of jitter



3.3. Jitter and audible feedback

The pitch changes with the jitter, and the jitter's change is limited by the pitch. Figure 7 is the F0 of /i/ pronounced by male subject and Figure 8 is the F0 of /i/ pronounced by female subject. These two F0 curves seem to be limited by two boundaries. When the F0 increases to the upper bound, the F0 starts to decrease. When it decreases near the lower bound, it starts to increase. It can be attributed to the effect of audible feedback. A simple example is that if one can't hear his own voice when he sings, he will be off key. Similarly, when one speaks, there are target tones in one's brain. So when the pitch drifts too far from the target value, the speaker will adjust the jitter value. It can explain why the synthetic voice is unnatural, even if synthetic jitter

is small, but the voice is natural, even if the natural jitter is large in some cycle. Just noticeable difference (JND) is a complex variation. Klatt [1] indicated that the subjects could detect a change of 0.3 Hz in a constant F0 contour when F0 = 120 Hz, but the JND is an order of magnitude larger (2.0 Hz) when the F0 contour is a linear descending ramp (32 Hz/sec). So JND varies with pitches. JND is also different between tone language speakers and non-tone language speakers; because as Wang [5] showed that, for the previous, the perception of tone is category perception that their JND may be larger.

Figure 7: F0 of /i/ pronounced by male subject

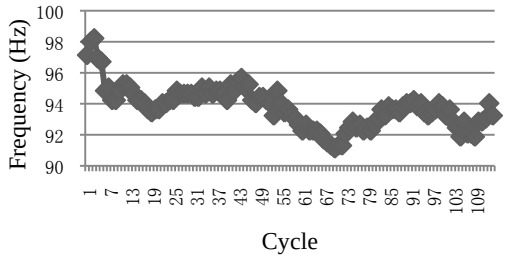
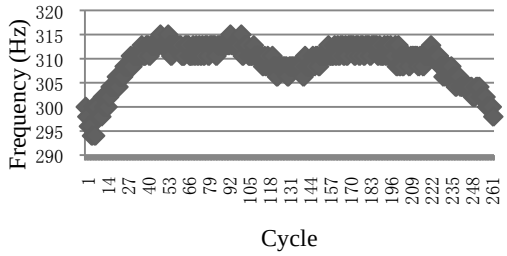


Figure 8: F0 of /i/ pronounced by female subject



4. CONCLUSION

Jitter value can be broke down into predictable and random components. The predictable components are that: all jitter values are around 0 and in most cases, adjacent values have opposite algebraic sign; the zero value percentage and zero-crossing rate of jitter increase with the pitch; the first peak position of the distribution of jitter increases with the pitch; each peak position is an integral multiple of the first peak position; the variance of pitch keeps in the allowance of just noticeable difference and it affects the variance of jitter further. The random components are that: the zero value position; the number of the peaks of the multimodal distribution; the position that jitter value jumped suddenly.

5. REFERENCES

[1] Klatt, D. H., 1973. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. In: The Journal of the Acoustical Society of America. Vol. 53, Issue 1, 8-16

[2] Heiberger, V. L., Horii, Y., 1982. Jitter and shimmer in sustained phonation. In: *Speech and Language: Advances in Basic Research and Practice*, Vol. 7, 299–332.

[3] Schoentgen, J., Guchteneere, R. D., 1994. Time series analysis of jitter. In: *Journal of Phonetics*, Vol. 23, Nos. 1/2, 189–201.

[4] Schoentgen, J., Guchteneere, R. D., 1997. Predictable and random components of jitter. In: *Speech Communication* 21, 255–272

[5] Wang, W.S.Y., 1976. Language change. In: Harnad, S.R., Steklis, H.D., Lancaster, J. (Eds.) *Origins and evolution of language and speech* Vol. 280, New York: New York Academy of Sciences.

当英语辅音遇上汉语音节

——英语音节间双辅音在普通话中的匹配[①]

When English Consonants Come Across Chinese Syllables:

Matching of English Inter-Syllable Bi-Consonants in Mandarin

吴君如

提要：本实验考察英语音节间双辅音在普通话中的匹配情况。实验令元音条件为 a，考察英语“(C)a C1 C2 a C”（C1、C2≠m/n/ng）音段配列在普通话中的匹配。当两个辅音都不为鼻音时，这个音段配列违反汉语音节结构限制，不经处理就不能直接匹配。实验假设辅音串音值和词重音两个因素对辅音串匹配音节数（syl）、辅音串匹配的从前系数（m）及辅音串匹配从后系数（n）这三个参数造成影响。被试用汉字听写英语假词，实验后转写这些汉字，并提取参数值。实验结果显示，辅音串相同时，后重单词的辅音串匹配从前系数（m）更高，而匹配成音节数（syl）更低；即后重单词中辅音串更容易影响前一音节的匹配，而前重单词中更容易发生音节增生；不同重音条件下辅音串从后系数没有显著差异；前重和后重单词的三个变量均显著相关。求每对前重和后重单词这三个变量的均值，以此均值为条件对辅音串进行分层聚类，得到了四类具有不同匹配特性的辅音串组合：中立成音节型、从后成音节型、中立非音节型、从前非音节型，并由各类的参数特点推出了一些制约辅音串匹配倾向性的音系条件。不同被试对同一个英语音段配列的匹配结果显示出明显的个体差异，音系匹配规则是通过影响某些匹配倾向性的强弱起作用的。

Abstract: In this experiment we examine how the bi-consonant string ”C1C2” (C1、C2≠m/n/ng) in English bi-syllable “(C)a C1 C2 a C” matches into Mandarin. Using pseudo words of English, three parameters are extracted: the number of Mandarin syllables each “C1C2” string matches to (syl), the reflection from “C1C2” on the former matched syllable (m), and the reflection from “C1C2” on the latter matched syllable (n). Two variants are considered: specific consonants in “C1C2” and the stress of the English pseudo words (“initial stress” and “final stress”). Subjects are asked to write down English pseudo-words in Chinese characters. Then the characters are transcribed into parameters. Paired T-test indicates that (1) The reflection of “C1C2” on the former Mandarin syllable (m) is significantly greater for final stress bi-syllables than initial stress bi-syllables; (2) initial stress bi-syllables tend to match the same “C1C2” to syllables more than final stress bi-syllables, causing epithesis; (3) the reflection of “C1C2” on the latter Mandarin syllable is not significantly influenced by stress condition. Hierarchical clustering yields five clusters of “C1C2” with different matching properties: type 1 which tends to cause epithesis and hardly reflects itself on the former or latter matched syllables, type 2 which also tends to cause peithesis but shows some more reflection on the latter matched syllable, type 3 which has little tendency of epithesis and has little reflection on the former or latter matched syllable, and type 4 which also has little tendency of epithesis but often reflects itself on the former syllable. Linguistic interpretations are offered for the clustering result. Considering the obvious personal variabilities in our data, we propose that phonological matching rules function by influencing the strength of specific matching tendencies instead of as clear-cut criteria.

关键字：音节、匹配、英语、汉语

1. 导言

1.1. 实验问题

有接触，必有匹配。有匹配，必有变形。当我们把“McDonald`s”听成了“麦当劳”，“sardine”听成“沙丁鱼”，把“Beetles”听成“披头士”时，我们在想什么？我们据何作此匹配？这个问题过于宽泛，搬出“信、达、雅”也不能生产出一个具有操作性的可以证伪的答案，本文只讨论它名下一个具体的问题：英语音节间双辅音在普通话中是如何匹配的？

1.2. 研究背景

以下从英语音节间双辅音的划分，音段配列限制及音节以上音系条件对匹配结果的影响两方面讨论本研究的背景。

1.2.1. 英语音节间双辅音划分问题

英语单词的音节划分本身就是一个饱受争议的问题。音节间的双辅音（VCCV）是其中一种情况，主要的争议在于两个辅音是都属于前面的音节，都属于后面的音节，还是应该分别属于前后两个音节。例如，“whisky”这个词的是应该划分为“whi#sky”，还是“whis#ky”，抑或“whisk#y”。关于英语单词的音节划分，研究者们提出了一系列原则和方案，下面我们来看一下这些方案在音节间双辅音归属问题上的应用。

首先是“首音原则”（Law of Initials）和“尾音原则”（Law of Finals），即一个英语单词词内部的音节，它的首音从必须是能够出现在词首的（LOI），它尾音从必须是能够出现在词末的（LOF）(Hoard 1971; Kahn 1976)。据此“whis#ky”和“whisk#y”不违反这两条原则，而“whi#sky”违反了尾音原则，因为英语没有以短/i/结尾的单词。

其次是“最大首音节方案”（Maximize Onset），即尽量把辅音从划为首音(Kahn 1976; Blevins and Goldsmith 1995)。单独应用此方案应该选择“whi#sky”这种划分，但是这样就会违反上述尾音原则，如果以首音原则和尾音原则作为该方案的使用前提，则应该选择“whis#ky”。

再次是“最大重音音节方案”（Maximize Stressed syllable），即尽量把辅音归给带重音（stress）的音节 (Hoard 1971; Bailey 1978; Wells 1990; Hammond 1999)。“whisky”是前重的，据此应该选择“whisk#y”。

最后优选论的“重重原则”（Weight-Stress Principle）也被用确定英语音节的划分(Prince and Smolensky 1993; Duanmu 2008)，即带重音的音节（stressed syllable）应该是重的（heavy），即韵母为 VC 或为 VV，不带重音的音节，应该是轻的（light），即韵母为 V，这个原则引入了“CVX 模板”，使归给重音节的辅音不至于过多，据此方案选择“whis#ky”。

研究者还通过各种行为实验测试母语者对音节划分的直觉。一个经典的测试是音节重复测试(Giegerich 1992)，即让母语者重复单词中的音节，如“after”-“af-af-ter-ter”。据此，英语母语者的直觉支持“whis#ky”，而“whi#sky”也是可以接受的，不太能接受的是“whisk#y”。总而言之，不论是母语者的直觉还是音系的处理，英语词中双辅音的音节划分都是有多种可能性的，而词重音是一个重要的影响因素。

相比之下，普通话的音节划分通常是清晰的，普通话的音节在首音、尾音位置上都只能容下一个辅音，不仅如此，哪些辅音能出现在首音位置上，哪些辅音能出现在尾音位置上是不受音节结构以外条件的约束的。

那么，当音节间双辅音违反汉语音节结构限制时，双辅音 C1C2 的音值对匹配的结果会造成什么影响呢？英语的词重音是音节以上的条件，它会对音节间双辅音在普通话中的匹配结果造成影响吗？

1.2.2. 音段配列限制及音节以上音系条件对匹配结果的影响

语言接触受到语言内部和外部因素的共同影响。本研究关注的是内部因素起作用的方式，下面就语言内部因素在音系匹配中的作用方式作简要回顾。

早在 60 年代，Weinreich 就提出了“语言干扰的机制和结构动因”（Mechanisms and Structural Causes of Interference），认为接触双方的语言结构确定了发

生干扰[②]的可能性(Weinreich 1961)。90 年代，陈保亚将语言接触的过程分为“匹配”和“回归”两个阶段；在“匹配”阶段，以音质相似为条件，母语者将受配语言（非母语）的音匹配为施配语言（母语）的音，在“回归”阶段母语者的非母语发音逐渐趋近受配语言(陈保亚 1996)。同时，Best 等人利用祖鲁语和英语的匹配实验，对特定语言对非母语的音段和对立的感知作了考察，提出了感知同化模型（PAM, Perceptual Assimilation Model）(Best 1995; Best, McRoberts et al. 2001)；该模型认为，存在一个以生理为基础的先天音系空间（native phonological space）(Liberman, Cooper et al. 1967)，非母语的音段会按照它和母语音段发音特征的异同来感知，非母语音段和母语音段共享特征的多少及共享特征在先天音系空间中的典型性决定着非母语音段会被感知同化到哪个母语音类里，也决定了非母语对立对习得的难易。音段相似性作为一个重要内部因素得到了广泛的支持。

问题在于，现在对匹配因素的讨论集中于音段的相似，而影响匹配的内部因素不限于此，我们认为至少还有两个因素对匹配结果造成影响。

其一，音段配列规则。研究中发现汉语舌面音声母以韵母为条件分别受两组傣语声母匹配，这受到傣语声韵搭配规则的限制(陈保亚 1996)。对维族人说汉语的研究也有类似发现。但是，这类讨论尚处于描述阶段，当母语的音段配列规则遭到违反时，选用母语的哪个音段配列和非母语音段配列匹配受到什么因素的控制，这有待进一步讨论。这里我们讨论英语音节间双辅音违反普通话音段配列限制的情况，假设音节间双辅音的音值会对匹配的结果造成影响。

其二，音节以上音系条件。这种条件如何影响匹配我们所知甚少。英语进入汉语的方式是一个英语单词匹配为几个汉字，因此我们认为，英语到汉语的匹配是英语的音系词到汉语韵律字的匹配，也就是说，匹配的单位是两种语言的句法韵律枢纽(王洪君 2008)。音系词可以包含多个音节，韵律字只含一个音节。英语音节划分和音段配列都受到词边界的制约，不受词组、句子层面的制约；韵律字的音节划分音段配列不受韵律词层面音系规则制约(徐通锵 1994) (王洪君 2008)。那么，当英语多音节的音系词和汉语单音节的韵律字进行匹配的时候，在英语音系词中起作用的词重音会起到怎样的作用呢？本研究讨论这里我们讨论英语词重音对英语音节间双辅音在普通话中匹配的影响。

1.3. 实验预设和操作定义

普通话不允许音节末尾有鼻音之外的辅音，也不允许音节首有超过一个辅音（介音另当别论）。因此，以下音段配列在汉语中是不能直接匹配的：

V1 C1 C2 V2 (C1≠n/ng)

当遇到这样一种音段配列的时候，根据音节通常以元音为核心的原则，不做其他预设，这个配列的音节划分有三种可能：

（1）V1 # C1 C2 V2 (C1≠m/n/ng)

（2）V1 C1 # C2 V2 (C1≠m/n/ng)

（3）V1 C1 C2 # V2 (C1≠m/n/ng)

排除音段音质本身的差异，由于普通话的声母只能是单个辅音[③]，韵尾只能是元音 i/u 或鼻音 n/ng[④]，（1）和（3）普通话中直接匹配不可能，而（2）也因为 C1 不符合普通话对韵尾的要求而不能成立。因此，在这种情况下，说普通话的人必须对 C1C2 进行处理，以下是一些处理的手段（表 1，表 2，表 3）：

（1）将 C1 或 C1C2 匹配为一个独立音节，例如 Kaplan-卡普兰 ka#pu#lan（C1C2=pl）。

表 1：

Ka	p	lan
卡	普	兰
ka	# pu#	lan

（2）将 C1 相似匹配为前一音节合法韵尾。

例如 Albert-艾伯特 ai#bo#te（C1C2=lb）将 l 相似匹配前一音节的合法韵尾 i。

表 2

Al	-	ber	t
艾	-	伯	特
ai	#	bo	te

又如，orlon-奥纶 ao#lun (C1C2=rl) 将 r 相似匹配前一音节的合法韵尾 o ([u])。

表 3

or	-	lun
奥	-	纶
ao	#	lun

（3）将 C1C2 当做一个音段匹配为后一音节的合法声母

例如 Adrian-雅芝 ia#zhi (C1C2=dr) 将 dr 一起匹配为后一音节的合法声母 zh。

表 4.

a	-	drian

雅	-	芝
ia	#	zhi

通常 C2 都会匹配为后一音节的声母。可能的情况还包括将 C1C2 分别匹配为一个音节，将 C1C2 都删除等。

a 是普通话中搭配声母能力最强的韵母，令 V1 V2 为英语字母 a 的变体，包括重音条件下的/æ/(有时为/a/)、非重音条件下的/ə/和/l/前面的/ɒ/（变体/ei/在实验中不出现），以控制匹配受到的普通话音段配列限制，这样简化了实验条件以后，我们的实验问题具体化为：

英语(C)a C1 C2 a C 双音节中的 C1C2 在普通话中的匹配（C1、C2≠m/n/ng）。

本研究预设说话人采用哪一种处理受到 C1C2 和英语词重音这两个因素影响。

为了对问题进行量化研究，我们必须赋予“匹配结果”操作定义。“匹配结果”指英语 C1C2 在汉语匹配结果中的对应成分。即找出 C1C2 之外的英语音段匹配到普通话中的成分，将其除去后剩下的部分。从“匹配结果”中我们可以提取三个统计变量。

（1）匹配音节数（syl）：C1C2 匹配形成音节的数量。

（3）匹配从前系数（m）：C1C2 匹配后进入前一个音节匹配结果的音段数。

（3）匹配从后系数（n）：C1C2 匹配后进入后一音节匹配结果的音段数。

2. 实验

2.1. 方法

参与者：两组各十名非语言学专业的被试自愿参加了实验，所有被试都收到了被试费。

材料：英语(C)a C1 C2 a C 双音节假词（pseudowords）录音（C= b/d/g/p/t/k），153×2 个。假词录音由一位受过语言学训练的母语为美式英语的发音人在隔音室录制，格式为 22050 Hz，16 bit。假词的编造依照以下程序：

（1）构造 C1C2 交叉表

（2）筛选出英语中可能的 C1C2 配列（153 个），以美式读音为准。(牛津大学出版社 2005)（例见表 5）

（3）成对构造前重和后重的英语假词，尽量保证 b/d/g/p/t/k 在 C 位置上均匀分布，（例见表 6）

表 5.“-”表示英语中没有的 C1C2 配列，“/”表示不在双音节之间出现的配列。

c2 / c1	/b/	/p/	/f/	/v/	/d/	/t/	/l/	/g/	…
/b/	–	–	–					–	…
/p/		–		–					…
/f/			–	–					…
/v/	–	–	–	–	/			–	…
/d/					–			–	…
…	…	…	…	…	…	…	…	…	…

表 6. 实验词表截录（C2=d），“//”前为前重音单词，“//”后为后重音单词

	…	/d/	…
/b/	…	gabdap//abdap	…
/p/	…	bapdab//apdab	…
/f/	…	dafdat//afdat	…
/v/	…	/	…
/d/	…	-	…
…	…	…	…

任务：用汉字听写下列英语单词的读音，要求尽　　量准确，书写清晰。

设计：实验为混合设计，组间变量为两个水平的重音条件，组内变量为 C1C2 的类型。将含有两种重音条件的测试对平衡分为两组，保证各辅音在每组的 C1 和 C2 位置都有所体现。被试随机分为两个组，分配给 2*2 个方格。这样，一方面每组被试不仅接触了所有 C1C2 配列，而且在两种重音条件下分别接触了所有的 C1 和 C2；另一方面每组被试都会听到两种重音条件的词，却又只会听到每一对的其中一个重音条件的假词，通过这种设计避免被试猜测实验意图，并避免一对假词相互之间的影响。

表 6. 实验设计

	前重	后重
假词 1 组	被试 1 组	被试 2 组
假词 2 组	被试 2 组	被试 1 组

过程：普通话数字序号-静音（0.3 秒）-假词录音两遍-作答时间（15 秒 ）两组随机播放。

自变量： C1C2、英语单词重音。

因变量：

（1）C1C2 匹配音节数（syl）。C1C2 匹配音节数为 C1C2 匹配结果中音节分隔符（“#”）的个数减去 1，syl = N(“#”) -1。

具体计算过程如下：将英语两个音节主元音和普通话匹配元音对齐，元音在匹配中忽略的情况以音节符号与英语元音对齐。提取普通话两个匹配元音之间的部分，作为 C1C2 对应的音段。（本文音段符号除非使用“/ /”标记，均为英语拼写或汉语拼音）。

必须说明几点：英语字母 a 对应汉语韵母中的 a、e、o。与英语第一个音节匹配的普通话音节的韵母有 ai/a/an/ao/e/er 等，本研究中预设第一音节的韵尾都来自 C1C2 的匹配；/æ/本身就有可能匹配为/ai/，所以严格说来 ai 里面的 i 不一定完全来自 C1C2 的匹配；er 为/ɚ/卷舌特征贯穿始终，拼音中的 r 也不能说是一个音位，而是一个特征，但用 er 匹配 a，可以认为卷舌特征来自 C1C2 的匹配，处理为一个音段在计算上相对方便。

（2）匹配的从前系数 m。C1C2 匹配结果中最前的“#”之前音段数为 m，m= N(segment before $\#_1$)。

（3）匹配的从后系数 n。C1C2 匹配结果中最后的#之后的音段数为 n，n= N(segment before $\#_{max}$)

例如：英语假词“asdad”匹配汉语“ai#zi#dan”，则英语“sd”对应汉语“#zi#d”，有两个“#”，C1C2 匹配音节数为 1。（表 7）

表 7.

a	sd	a	d
a	i#zi#d	a	n

由“asdad”-“ai#zi#dan”得到“sd”-“i#zi#d”，syl=1，m=1，n=1；

由“balshag”- “bao#shan”得到“lsh”-“o#sh”，syl=0，m=1，n=1；

由“tadshab”-“ta#sa”得到“dsh”-“#s”，syl=0，m=0,n=1。

2.2. 数据处理与统计结果

将被试在问卷中填写的汉字转写为汉语拼音，用“#”标出音节界线。求出每个 C1C2 对应的音段，并计算辅音串匹配音节数（syl）和辅音串从从前系数（m）及从后系数（n）。分别对这三个因变量求出每个假词十个被试的取值之和。需要说明一点：被试的匹配策略是有明显的个体差异的，有些被试倾向于用较少的音节匹配英语单词，而有些被试则倾向于为每个音段匹配一个音节。根据被试平均每个单词匹配的音节数对两组被试进行排名，名次相同的被试的数据相对应（表 8），可以得到十组覆盖整个词表的数据，虽然第二组比第一组在最低平均音节数上要低约 0.3 个音节，但由于设计中词对在两组中是随机分布的，因此还是可以控制被试特质在分组上不可预料的差异对统计的影响。

表 8. 被试匹配

	组 1		组 2	
十组数据	被试编号	单词平均匹配音节数	被试编号	单词平均匹配音节数
1	2	2.917722	3	2.592357
2	1	2.936709	6	2.719745
3	5	2.936709	10	2.923567
4	3	3.037975	4	3.031847
5	9	3.101266	2	3.038217

6	7	3.126582	8	3.140127
7	6	3.348101	1	3.146497
8	4	3.35443	5	3.375796
9	10	3.392405	9	3.414013
10	8	3.683544	7	3.547771

表 9. 重音条件对匹配结果的影响。

	辅音串平均匹配音节数（Msyl）	平均从前系数（Mm）	平均从后系数（Mn）
前重	0.7667	0.4706	1.0634
后重	0.6993	0.5176	1.0654

以单词前重和后重为配对变量，以辅音串的匹配音节数（syl）、从前系数（m）和从后系数（n）分别为因变量，进行配对 T 检验（不包括 9 个没有对应后重单词的前重单词）。观察重音条件对匹配结果的影响（表 9）。

对于从前系数（m），t (152)= -2.192， sig<0.05，说明音节前重与后重对应的从前系数（m）显著不同，后重单词的从前系数显著大于前重单词。这可以理解为，同样的辅音串在后重单词中比在前重单词中更容易将其特征反映在前面的音节上。对于辅音串的匹配音节数（syl），t (152)=2.567, sig<0.05，说明音节前重与后重对应的辅音串匹配音节数（syl）显著不同，前重单词的辅音串匹配音节数显著大于后重单词。这可以理解为，同样的辅音串在前重单词中比在后重单词中更容易匹配为汉语的音节。

例如，大部分被试会把“lp”中的 l 匹配为前一音节的韵尾，得到 u#p，这是总的趋势，但对于前重的 dalpad，有一小部分被试会把 l 匹配为单独的一个音节而非前一音节的韵尾，而对于后重的 alpad，就没有被试会这样做。同理，对于前重的 batlag 较少被试把前一个音节匹配为 bai，大多数被试会把 t 匹配为一个音节；而对于后重的 atlag，将前一音节匹配为 ai 的被试数量明显增多，把 t 匹配为一个音节的被试相对较少。

对于从后系数（n），t(152)= -0.210, sig>0.05，说明音节前重与后重对应的从后系数（n）没有显著不同。

同重音相同音段配列的两组单词在这三个变量上都显著相关，n 的相关系数 Cn (152) =0.483, sig <0.001，m 的相关系数 Cm (152)=0.755，sig<0.001，syl 的相关系数 Csyl (152)=0.581, sig<0.001。可见处于不同重音环境中时从后系数 m 是最稳定的，成音节数 syl 次之，从前系数 n 对重音环境最为敏感。

接下来，我们来看辅音串音段配列对匹配结果的影响。

用 Ward 方法以各对轻重音单词 m、n、syl 的平均值对辅音串进行分层聚类（聚类数 3-9），结果如下。当取 4 类时，可以得到以下 4 个类（图 1）。

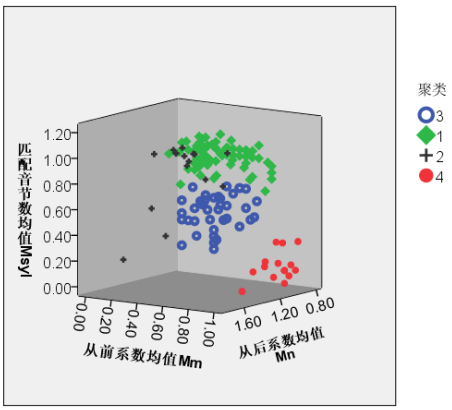
第 1 类：bd 、bt 、bl 、bch、bw 、pd 、pt 、pl 、pg 、pk 、ps 、psh、pch、pdg、pw 、fb 、fp 、fd 、ft 、fl 、fg 、fk 、fh 、fs 、fsh、vt 、vl 、vk 、vw 、dl 、dw 、tf 、tv 、tl 、tw 、gb 、gp 、gf 、gt 、gl 、gz 、gw 、kb 、kp 、kf 、kd 、kt 、kl 、ks 、ksh、sb 、sp 、sf 、sv 、zd 、st 、sl 、sg 、sk 、sth、sch、sw 、zb 、zf 、zd 、zt 、zl 、zg 、zk 、zh 、zdg、zw 、tht、thl、thk、thw、shf、shl、shr、shw、rz（Mm=0.4574，Mn=1.0259，Msyl=0.9407，N=81）

第 2 类：bdg、br 、pr 、fr 、fw 、vr 、dr 、tr 、gr 、kch、kr 、kw 、sdg、sr 、zr 、thr（Mm=0.4531，Mn=1.4344，Msyl=0.8719，N=16）。

第 3 类：bv 、bh 、pb 、pf 、fth、db 、dp 、df 、dv 、dt 、dk 、dh 、ds 、dth、dsh、tp 、tp 、td 、tg 、tk 、tth、lh 、gd 、kg 、kh 、sh 、ssh、thf、rb 、rp 、rf 、rv 、rd 、rt 、rl 、rg 、rk 、rh 、rs 、rw（Mm=0.4125，Mn=1.0050，Msyl=0.4963，N=40）。

第 4 类：lb 、lp 、lf 、lv 、ld 、lt 、lg 、lk 、ls 、lz 、lth、lsh、lch、ldg、lr 、lw，（Mm=0.9250，Mn=1.0375，Msyl=0.1344，N=16）。

图 1：分层聚类三维散点图



方差分析（ANOVA）检验显示，从前系数（m），F(3)=39.45，sig<0；从后系数（n）F(3)=105.385，sig<0；匹配音节数 F(3)=242.171，sig<0。这说明分出的类在三个因变量上的差异都是显著的，从而验证了分组的合理性。

观察分类结果的参数值， syl 值高说明容易刺激匹配音节的增生，称为“成音节型”； m 值高说明容易影响前一个音节，称为“从前型”；n 值高说明容易影响后一个音节，称其为“从后型”。如此，根据参数值将这四类分别命名为：中立成音节型、从后成音节型、中立非音节型、和从前非音节型。根据聚类结果，中立成音节型和从后成音节型（1、2）较为接近，可以合并为一类，这些辅音串容易导致两个音节中间增生音节；而剩下的两类都不容易增生音节，中立非音节型辅音串对其前后的音节都不怎么产生影响，而从前非音节型是非常独立的一类，这类的 C1 都是“l”，“l”总是匹配为前一音节的韵尾，这样 C2 也不受阻碍地匹配为后一音节的声母，只偶尔匹配为声母和介音（表 10）。

表 10：辅音串匹配特性类型。从前系数组内均值 Mm，从后系数组内均值 Mn，匹配音节数组内均值 Msyl。参数的最大值最小值分别用上标 max 和 min 表示

	类型	Mm	Mn	Msyl	例
1	中立成音节型	0.4574，	1.0259	0.9407 max	bashlag → ba#shi#lai taglap→tai#ge#la
2	从后成音节型	0.4531	1.4344 max	0.8719	apjap→a#bu#jia
3	中立非音节型	0.4125 min	1.0050 min	0.4963	cabhad→ka#pa
4	从前非音节型	0.9250 max	1.0375	0.1344 min	balshag→bao#shai
	*				

若从辅音串对这三个指数的影响观察，可以发现 C1 对是否成音节（即产生增音 epithesis）的作用比较大，C2 对从后系数有决定性的影响。首先，从 C1 来看，大多数 C1 倾向于独立成音节，如 p、f、v、g、k、s、sh、z；另一些则倾向于和 C2 合并作为后一音节的声母，如 d、t、r；l 作为惟一的边音，在 C1 位置上的匹配上也独树一帜，总是匹配为前一音节的韵尾；一些舌根和舌尖辅音在 C1 位置上可能会使元音增生韵尾，如 g、k、s、t、z。其次，从 C2 来看，C2 为 r、dg、w 时，C2 总是匹配为后一音节的介音，而 C2 为 h 时，h 倾向于匹配为后一音节的送气成分，v 在浊音后会和 C1 合并，在清音后独立匹配为后一音节声母，其余辅音在 C2 位置上倾向于匹配为后一音节的声母。

3. 讨论与结论

3.1. 词重音和音节结构

研究发现，被试用普通话匹配英语音节间双辅音时，同样的辅音串在后重单词中比在前重音节中更容易影响前面音节，在前重单词中比在后重单词中更容易匹配为汉语的音节，而重音条件对辅音串是否将其特征反映在后一个音节中没有明显影响。

导言中我们提到了音节结构和重音的关系。“重重原则”（Weight-Stress Principle）和 CVX 模板建立了重音和音节结构的联系，带重音的音节（stressed syllable）音节结构为(C)VX（重，heavy）；不带重音的音节（unstressed syllable），音节结构为(C)V（轻，light）(Prince and Smolensky 1993)。这个联系也能够为本研究的结果提供解释。端木（Duanmu）认为普通话的音节除轻声外都是重的（heavy syllable），

而英语带重音的音节才是重的(Duanmu 2008)。当英语的首音节不带重音时，首音节填入汉语首音节 CVX 之后还留下一个空位，为 C1 音段填入提供了可能性，当音位配列不允许的时候，C1 就匹配为相似的音段；当英语的首音节带重音时，首音节填满汉语首音节的 CVX 结构，C1 音段被迫独立出来，增生一个音节，或者和 C2 结合匹配为后一音节的声母，无论怎样，C2 总是很难进入汉语的第一个音节，所以重音条件对从后系数 n 的影响都不太大（除了 j 等匹配为后音节的声母和介音）。这就解释了为什么同样的辅音串在后重单词中比在前重音节中更容易影响前面音节，在前重单词中比在后重单词中更容易匹配为汉语的音节。

实验结果为重重原则的跨语言普遍性提供了一个证据。同时，本研究证明了音段相似意外，音节以上规则也是影响匹配结果的重要因素，实验结果验证了句法韵律枢纽作为匹配单位的重要性。

3.2. 个体差异匹配初期的多样性和规律

本研究证实了词重音与辅音串音段配列，在英汉匹配中的作用，这并不意味着这些因素并可以作为“非此即彼”的判断准则来使用。实验中还发现了显著的个体差异，面对匹配中的矛盾有些被试倾向于用增生音节的方式来解决，而有些被试则倾向于删除不匹配的音段，面对同样的双音节假词表，两种倾向的被试平均匹配音节数的差异能够达到 1 个音节。本研究发现的规律是隐藏在纷繁复杂的个体差异中的统计性规律，换句话说，词重音和辅音串实际上是通过控制某些匹配倾向性的强弱起作用的。

这提醒我们重新思考内部因素在语言接触中起作用的方式。

自“新语法学派”（Neogrammar）以降到生成学派的普遍语法（General Grammar），语言学家们大都认为语言规律是没有例外的，少数例外也一定是能够被解释的，中国的语言学家说“例不十，法不立；例外不十，法不破”（黎锦熙、王力等）。

不过，语言规则是一种社会规约，即使形成的过程是“不约而同”的，这个社会规约也有一个形成的过程。那么在语言接触发生时，同一个语言社团的母语者接触到同一个非母语语音片段，在形成某种规约之前，他们对这个片段的匹配会一模一样吗？实验证明并非如此，说普通话的一群人对同一个陌生音位的匹配结果是不同的，虽说不同，却呈现出总体一致，个别变异，变异有序的分布趋势(吴君如 2009)。可见所谓“结构动因”(Weinreich 1961)本身并不能预测单一的结果，而只能对匹配结果的倾向加以限制。换句话说，以语音音系为条件的匹配规则是一种统计性质的规则，它只能预测“A”在这个条件下匹配为“A1”的可能性最大，匹配为“A2”的可能性次之，匹配为“An”的可能性非常小，却不能预测“A”在这个条件下一定只会匹配为“A”。

那么，为什么一个外来词往往只有一个通用形式呢？具有一定可类推性的借词读音规则（如中国周边国家的“汉字音”）又是怎么产生的呢？这跟本研究的结果是否矛盾呢？不矛盾，因为语言接触中还有外部因素在起作用。首先，人群的大规模自然接触为多种可能性的相互竞争提供了条件，社团权威的形式会压抑其他形式，逐渐成为惟一的形式；其次，长期高强度的语言接触(Thomason and Kaufman 1991)提供了大量语言材料，为对应规则的建立打下基础，一旦规则建立，代代相传，即使后来按照规则生成的匹配结果不再符合相似原则，规则还是能够长期继续发挥作用，这时当下的语音音系条件要让位于历时的匹配规则；再次，如果接触主要是通过文献进行的，选用哪种匹配形式受到译者个人倾向和文学趣味的强烈影响；最后，现代学术和政治领域的匹配形式则受到官方规范的严格控制。在这些外部因素的作用下，其他符合语言内部条件限制的变体“阵亡”了。举最新的例子来说：“Viagra”有两个汉语音译词“万艾可”和“伟哥”，它们用来匹配“Viagra”的概率比两个或三个任意韵律字的组合高，这是内部因素在起作用；而“万艾可”得到官方的支持，“伟哥”在民间获得拥趸，这就要看外部因素了。本研究要求被试将听到的英语假词用汉字写下来，排除了以上外部因素和语义的制约，使得内部音系因素的作用得到凸显。

3.3. 结论

总之，本研究控制了外部因素的作用，从纷繁复杂的匹配变体中挖掘出潜在的音系条件及其作用对象，证明了音节结构和音节以上的规则也是影响匹配结果的重要因素，实验结果重音和符合音节结构的联系。研究探讨了这些因素的作用方式，认为音系匹配规则是通过影响某些匹配倾向性的强弱起

作用的，内部因素作用下的匹配符合总体一致，个别变异，变异有序的分布趋势。

[①]本实验得到语言学与应用语言学各位同学诸多帮助，在此特别对发音人安义宣同学，预实验参与人杨锋同学和李子鹤同学，以及协助进行实验组织的吴韩娜同学表示衷心的感谢。

[②]干扰（interference）双语者言语中出现异于任何一种语言规范的用例，就是干扰现象。（p.1）“干扰”意味着由于外来成分进入语言系统而引起的结构重组。——魏茵莱希对干扰的定义（李子鹤译）

[③]介音属于声母还是韵母有争议，性质也介于辅音和元音之间，在这里“声母”指的是不含介音的情况

[④]另外 n 在唇音前面发生语流音变，可以变成 m，单元音韵母是不是占据韵腹和韵尾也有争议，和本题无关，亦存而不论。

参考文献

牛津大学出版社 (2005). 牛津大学英语词典. 牛津大学出版社, 上海译文出版社**:** 1682.

王洪君 (2008). 汉语非线性音系学.

徐通锵 (1994). "“字” 和汉语的句法结构." 世界汉语教学 **2**.

吴君如 (2009). 双唇辅音到普通话声母的匹配分布——相似匹配的特征解释.

陈保亚 (1996). 论语言接触与语言联盟, 语文出版社.

Bailey, C.-J. N. (1978). Gradience in English syllabification and a revised concept of unmarked syllabization. Bloomington, Indiana University Linguistics Club.

Best, C. T. (1995). "A direct realist view of cross-language speech perception." Speech perception and linguistic experience: Issues in cross-language research: 171-204.

Best, C. T., G. W. McRoberts, et al. (2001). "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener’s native phonological system." The Journal of the Acoustical Society of America **109**: 775.

Blevins, J. and J. Goldsmith (1995). "The syllable in phonological theory." 1995: 206-244.

Duanmu, S. (2008). syllable Structure, Oxford University Press.

Giegerich, H. J. (1992). English Phonology: An Introduction. Cambridge, U.K, Cambridge University Press.

Hammond, M. (1999). The phonology of English: a prosodic optimality-theoretic approach, Oxford University Press, USA.

Hoard, J. E. (1971). "Aspiration, Tenseness, and syllabication in English." Language: 133-140.

Kahn, D. (1976). syllable-based generalizations in English phonology, Indiana University Linguistics Club.

Liberman, A. M., F. S. Cooper, et al. (1967). "Perception of the Speech Code1." Psychological review **74**(6): 431-461.

Prince, A. and P. Smolensky (1993). Optimality Theory: Constraint Interaction in Generative Grammar. New Burnswick, Rutgers University.

Thomason, S. G. and T. Kaufman (1991). Language Contact, Creolization, and Genetic Linguistics

语言接触、克里奥尔化和语言发生学, University of California Press.

Weinreich, U. (1961). "LANGUAGES IN CONTACT." Psycholinguistics: A Book of Readings.

Wells, J. C. (1990). "syllabification and allophony." Studies in the pronunciation of English: a commemorative volume in honour of AC Gimson. London: Routledge: 76-86.

（100871，北京，北京大学中文系，yuerr@163.com）

本文收到自然科学基金支持（项目批准号：61073085）

法庭语音学[i]

〔德〕Michael Jessen　曹洪林 王英利 译

摘要：本文对法庭语音学进行了研究综述，主要介绍了该学科的核心内容：说话人鉴定。在实际办案中，当未找到嫌疑人，只有犯罪分子的检材语音时，可以使用说话人画像/说话人分类技术。若没有犯罪分子的录音证据时，可以让受害人和证人进行说话人的听觉辨认。具体的辨认形式有两种：对熟人辨认和对陌生人辨认，在对陌生人辨认时可以采用语音辨认的方法进行。当检材语音和样本语音都齐备的时候，法庭语音分析专家就可以对二者进行比对检验了。目前语音比对分析涉及到的问题和领域有：基于贝叶斯方法的法庭推理和似然比计算、共振峰频率的测量应用、非解析感知与样例理论、法庭说话人自动识别以及不同方法的综合应用等。

【Abstract】An overview of forensic phonetics is presented, focusing on speaker identification as its core task. Speaker profiling/speaker classification is applied when the offender has been recorded, but no suspect has been found. Auditory speaker identification by victims and witnesses becomes relevant when no speech recording of the offender is available. It can take the form of familiar-speaker identification or unfamiliar-speaker identification, and in the latter case a voice line-up/voice parade can be carried out. When recordings of both the offender and a suspect are available, a voice comparison is done by an expert in forensic speech analysis. Current issues and domains in voice comparison analysis include the Bayesian approach to forensic reasoning and the Likelihood Ratio, the use of formant frequency measurements, non-analytic perception and Exemplar Theory, forensic automatic speaker identification, and the interaction between different methods.

关键词：语音（说话人）比对；语音（说话人）画像；说话人听觉鉴定；说话人自动识别；样例理论；似然比；共振峰频率；基频。

【Key words】forensic voice comparison (speaker comparison); voice profiling (speaker profiling); auditory speaker identification; automatic speaker recognition; exemplar theory; likelihood ratio; formant frequency; fundamental frequency.

Michael Jessen，德国联邦刑事侦查局（BKA）。曹洪林，证据科学教育部重点实验室（中国政法大学）；北京大学中文系。王英利，广东省公安厅刑事科学技术中心。原文刊载在Language and Linguistics Compass，2008 (2/4): 671–711。本文的翻译已经得到作者及Wiley-Blackwell出版社的授权。感谢北京大学中文系汪高武博士和中国科学院自动化研究所朱磊博士对本文部分内容提出的宝贵意见。

1. 引言

法庭语音学（forensic phonetics）是一门运用普通语音学的知识、理论与方法，为警方的侦查破案提供帮助或为法庭提供证据的应用学科，同时也研究最新的特别是与法庭语音相关的知识、理论与方法等方面的发展情况。法庭语音学的核心内容是说话人鉴定（speaker identification），也有人称之为“说话人识别（speaker recognition）”（本文中假设二者同义，尽管它们在特定的法律体系和科学领域中可能各有其特定的含义）。另一块内容是分析语音段落的语言内容，这些语音往往由于技术或人为因素致使其可懂度大大降低，有人也常称之为有争议话语的分析/检验（disputed utterance analysis/examination）。法庭语音学中的许多内容都与其他学科有着很强的联系，例如语音技术、言语工程、普通声学等。比如说音频增强（audio enhancement）就要应用各种形式的滤波技术，才能提高低质量录音的可懂度。又如音频的真实性检验（audio authentication），则要求指出被检录音的可疑之处是否对原始语音进行过增、删、改等处理。最近，对说话人自动识别技术（automatic speaker identification）而言，语音技术也变得非常重要。由于在法庭科学实验室中，往往只有语音学家拥有设备和经验可以对音频信号进行分析，因此，有时候他们还被邀请去分析一些非语音的声学现象，像枪声、飞机驾驶舱内的声音或鸟叫声。与音频增强和真实性检验相比，除了那些凭借对常见声音的经验判断可以解决的情况以外，这时需要鉴定专家具备更多跨学科的知识。本文未展开讨论法庭语音学中说话人鉴定以外的其他内容，读者若想了解音频增

强、真实性检验和非语音现象分析等技术的更多内容，请看 Hollien （1990）、Broeders （2001, 2004）和 Bijhold （2007）等学者的著作和文章。French （1990）以及 French 与 Harrison （2006）的文章中还介绍了一些对有争议话语分析的案例。

“法庭语音学”这个术语得到官方认可的时间至少可以追溯至 1991 年“国际法庭语音学协会”（IAFP，协会网址：http://www.iafpa.net）创立之时。该协会每年都举办一次讨论会，许多相关的会议报道都会在《国际语音、语言与法律杂志》（http://www.equinoxjournals.com/ojs/index.php/IJSLL）上发表，所有这些活动都对法庭语音学的发展做出了一定的贡献。正如该杂志的旧刊名（《法庭语言学》*Forensic Linguistics*）所指出的那样，该杂志不仅是法庭语音学家的研究阵地，同时也是那些参加“国际法庭语言学家协会”的法庭语言学家的研究平台（比较 Shuy 2007）。2004 年，IAFP 在它的名字中加入了“声学”一词，变成了现有的名字“国际法庭语音与声学协会”（IAFPA）。此次更名显然是为了尊重那些在该领域内做出了重要贡献的声学信号分析和语音技术专家。经常与“法庭语音与声学”一块出现的还有“法庭语音与音频分析（Forensic Speech and Audio Analysis）”这个术语。尽管实践中这两个术语的外延非常接近，但由于后者未包含“语音学”这个字眼儿，因此对其从业者而言更为中性一些，同时也更为许多在该领域内提供服务的政府组织所青睐。尽管笔者认为在法庭语音与音频分析部门中，至少应该有部分语音学家或语言学家，然而并非所有国家都认同笔者的观点，他们可能认为说话人鉴定及其相关的工作应该是技术领域的事情，工程师和警察在接受了某些软硬件的使用培训之后就可以胜任此项工作了。虽然笔者对一些个人或公司从事法庭语音分析的情况不是很了解，但是至少在英国和德国，大部分从事此业务的私人专家都是语音学家。2007 年秋季开始，英国的约克大学创办了第一个与此相关的硕士学位——“法庭语音科学理科硕士”（http://www.york.ac.uk/depts/lang/postgrad/forensic.htm）。

本文将集中讨论说话人鉴定方面的内容。相关内容读者还可以参考 Nolan（1983）、Hollien（2002）、Rose（2002）、Künzel（1987，德语，针对会说德语的读者）的专著。Baldwin 与 French（1990）以及 Hollien（1990）在其著作中介绍的内容则要更广泛一些，包含了一些法庭语音学其他方面的内容。之前关于法庭语音学（用英语撰写）的几篇综述性的文章也很好，像 Nolan（1994、1997、2001）、French（1994）、Künzel（1995、2004）、Meuwly（2000）、Broeders（2001、2004、2006）、Gfroerer（2003）以及 French 与 Harrison（2006）都值得一读。

2. 法庭说话人鉴定的类型

说话人鉴定是法庭语音学（与声学）最主要的内容，它还可以再细化为三个部分：语音比对（voice comparison）、语音画像（voice profiling）和受害人与证人进行的说话人鉴定（speaker identification by victims and witnesses）。

2.1. 语音比对

在语音比对[1]中，首先要有未知说话人的检材语音。检材语音的来源有很多种，像绑匪为索要赎金打的敲诈电话、贩毒分子通过电话（被监听）进行的非法交易以及一些蓄意骚扰或胁迫的案件等。其次，还要有（或可以录制）嫌疑人的样本语音，方可与检材语音进行比对。根据不同的法律体系，在制作样本语音时，如果嫌疑人不配合或另外还需要更多形式的语音样本时，有时监听取得的秘密电话录音或与警方会谈的录音记录也可以作为证据使用。当检材语音和样本语音（有时会出现多个检材或样本语音的情况）都具备了，鉴定专家便可以进行语音比对了。在比对检验中，要使用一种或多种不同的方法对各种各样的语音特征进行分析。分析结束后，鉴定专家会得出检材语音与样本语音是否是同一人或不同人所说的结论。然而不同国家之间，由于习惯和观念的差异，鉴定结论的表述形式也有

[1]之所以在“语音比对”、“语音画像”等相关术语中采用“语音（voice）”的说法，是想采用以点代面的方式表示言语产生（speech production）过程中涉及到的所有领域。这里说的“语音”不仅包括其字面意义上的嗓音——声带的振动类型，还包括广义上的语音（上呼吸道的调音），像鼻音，以及更广义范围上的语言学特征，像说话人的方言特点等。当非专业人员（layperson）说“我听过那个声音”的时候，常常都没有意识到那些能够反映说话人个性特点的众多语音学和语言学特征。对于语音比对、语音画像等术语来说，还有一些其他类似的叫法，像说话人比对（speaker comparison）、说话人画像（speaker profiling）等。

所不同（参见 French 与 Harrison 2007 年发表的关于英国现有程序表述的立场声明的文章）。[2]

语音比对可以作为警察侦查破案的一种手段，一般不涉及出庭问题，但最为常见的情况还是要出具书面的科学鉴定报告，作为证据在法庭上使用。至于是否出庭的问题，根据不同的法律体系、国家或地方立法情况的不同会有不同的要求，有的国家或地区会要求鉴定专家出庭，就其出具的报告进行口头解释和辩论。本文第三部分将对此问题进行详细阐述。

2.2. 语音画像和说话人分类

在警方的初期侦查阶段，常常发生这样的情况，即只有犯罪分子的检材语音而找不到嫌疑人，录不到样本语音，故而无从比对。此时，警方常常邀请法庭语音学专家就检材语音对犯罪分子作分析画像。警察或普通人往往都不是语言/语音分析的行家里手，而法庭语音学专家所作的语音画像就可以帮助他们缩小侦查范围甚至直接找到嫌疑人。根据检材语音的长度、质量及所含信息量的差别，语音画像或多或少可以反映出说话人的一些籍贯、年龄、性别、文化程度、社会背景、母语（如果带有外国口音的话）以及是否有影响发音的疾病等方面的重要信息。从一个外行人的角度看，检材语音中还可能包含了某些更深层的“不寻常”的信息。比如说，说话人可能带有高音调嗓音或说话速度特别快等特点。其实并不是说，分析时非要使用那些一般人掌握不了的语言学或语音学的专业术语，非要进行观察报告（或者甚至是声学测量！）。有些情况下，为了抓获犯罪分子，还可以直接在电视、广播或网络上播放检材语音。Jessen（2007b）曾对语音画像进行过研究综述，作者详细介绍了语音画像可以提供的具体信息。

将*语音画像*与*说话人分类*分开来讲是有意义的。正如刚才所讲，语音画像是从实践角度定义的术语，它提供的关于说话人的信息主要用来查找嫌疑人。而说话人分类则是从理论角度进行定义的，主要是由检材语音的语言学和语音学模式中推断说话人的群体属性，包括年龄、性别、社会群体和地域等信息（详细内容参见 Müller 2007 的著作）。语音画像与说话人分类的大部分内容是交叉的。语音画像主要也是对说话人进行分类，但它同时也包含了对说话人那些“不寻常”的说话特性进行分析的内容，比如说，如果发现犯罪分子带有高音调特征，可能会有助于找到嫌疑人，但这对说话人分类来说就没有多大意义。说话人分类的目的并不只是为了语音画像，同时还可以为语音比对或者语音辨认的设计服务（内容详见本文 2.3.2）。

说话人分类非常依赖鉴定专家的知识面。幻想一个专家对目标语言的所有方言都非常了解，对社会语言学、言语病理学、第二语言语音学等都非常精通，几乎是不可能的。重要的是要有责任心。如果在实际办案中，某个专家没有掌握说话人分类中某方面的知识，而且通过阅读文献、查询数据库或有针对性地进行专门研究等方式依然解决不了问题的话，那就要另请专家（这是最迟的时间）进行分析了。如果进行说话人分类的目的是为了进行下一步的语音比对和法庭展示，那么对分类动机科学性的要求就会更强。然而对于警方侦查中采用的语音画像而言，最有价值的却是其提供信息的有效性及其具体的分析速度。

语音画像与心理画像（psychological profiling）并不相同，事实上IAFPA在其网站的业务守则（www.iafpa.net/code.htm）中就推荐其会员不要进行心理画像。[3]心理画像技术源于美国联邦调查局行为科学室，并已有部分著作出版，如Douglas与Olshaker（1995）。基于对已有（主要是在犯罪现场发现的）证据的分析，心理画像专家便可以推测犯罪分子的心理类型或症状、性别、年龄、种族、体型、社会经济地位，甚至还可以进一步推测犯罪分子的衣服款式、汽车品牌、工作性质、职业行为、住所类型及地址、潜在的言语缺陷等方面的信息（“*心理画像*”这个说法只有部分的合理性，因为在这些分类中，部分分类是基于生物学和社会学而非心理学/行为科学进行的）。一般人会认为上述分析都是基于画像人员的技巧和经验——包括画像人员对案件特殊性质的整体把握以及与犯罪分子产生的共鸣——而得出的，但是这些分析结果中只有一部分有科学的统计依据，可以将具体的现场证据与上文提到的对犯罪分子群体属性的分类联系起来。

[2]关于此立场声明及其他相关信息可以参见该网站 www.forensic-speech-science.info。

[3] IAFPA 还推荐大家不要评估说话人的诚实问题，主要是因为到目前为止测谎技术还不是很可靠（比较 Hollien 1990 第 13 章；Eriksson 与 Lacerda 2007 文献中的内容）。

笔者绝非有意批评这些画像方法，其实它们可以提供许多有用的信息，对抓获罪犯也很有帮助，而且，各国进行画像分析的方法也不尽相同。比如，在美国和德国，与其他方面相比，并不怎么强调心理分析。尽管语音画像与心理画像在对犯罪分子进行分类（如性别和年龄）以及分析的紧迫程度上都有重合之处，但二者并不相同。比如，语音画像的对象仅限于录音证据（法庭语音专家很少亲临犯罪现场），主体是受过训练的语言学家或语音学家，他们能够对录音证据进行分析进而得出较为科学的意见。当要推断犯罪分子的某些心理痕迹时，比如推测犯罪分子患有某种精神错乱症，应该在查阅相关文献资料（参见Darby1981著作中的做法）的同时，还要寻求一些有能力的心理学家、精神病学专家或神经病学专家的合作和帮助。

在约克郡杀人案中，语音画像起到了很重要的作用（Ellis1994、French 等 2006）。1975 至 1980 年间，在英国的利兹、布拉德福、哈德斯菲尔德和曼彻斯特市先后共有 13 位女性惨遭杀害。1978 年至 1979 年，当地的报社和警察局都曾收到数封信件，最后还收到一盘录音带，写信人和磁带中的说话人（caller）都声称对这些谋杀案负责，还扬言要继续作案。警方遂向语音学和方言学家 Stanley Ellis 寻求帮助，请求他对涉案磁带进行分析并确定说话人的具体位置。Ellis 凭借他多年的实地研究经验，查阅了大量的方言地图和书籍资料，最终将范围缩小到了森德兰市（纽卡斯尔市附近的一个城镇），而且周围环境很可能是磁带中的说话人生活过的地方。后来 Ellis 对森德兰市进行了实地访问，并把磁带内容放给当地群众听，同时还在该地区的不同地方作了当地方言的录音采样，经过分析，Ellis 最后得出结论并报告警方说磁带中的说话人是在靠近森德兰市的索斯威克和卡斯尔敦地区长大的。事后证明，那个说话人确实出生于森德兰市西部的一个地方，距离 Ellis 所说的地方只有大约 1 英里远。鉴于此案语音画像非常准确，同时磁带内容（磁带中的说话人除了具有口音特征外，其他方面也非常有个性）也在当地媒体上反复播出过，按理说应该抓到他了，然而情况并不如意，警方在搜查嫌疑人的时候，犯了一个非常严重的错误，即只将搜寻目标锁定在与磁带中的说话人具有相同口音且没有不在现场的证据的人上了。过了一段时间，不但没有找到嫌疑人，犯罪分子又重新出来作案了。Ellis 和他的同事 Jack Windsor Lewis（对涉案信件进行过分析）开始怀疑本案中出现的磁带和信件可能是场恶作剧，便向警方表达了他们的疑虑，指出本案磁带中的声音非常特别，警方很可能已经在之前的排查工作中接触过磁带中的那个说话人了，但却由于他当时有不在场的证据而逃脱了。许多线索表明，涉案磁带中的说话人（同时也是涉案信件的书写人）只是整个恶作剧的制造者，并不是本案中的杀人犯，然而警方却忽视了这一点。又过了 2 年，直到 1981 年，真正的凶手才最终落网，并被判以终生监禁。然而，很多年后，那个制造恶作剧的捣乱分子依旧逍遥法外，直到 2005 年，警方通过对涉案信件进行 DNA 分析才抓到了他。由此，该事件还引发了法庭语音学和语言学一连串的深层分析，详情可以参见 French 等（2006）的文章。那个捣乱分子最终因扰乱司法而被判有罪，并处以 8 年监禁。关于本案中那个捣乱分子被抓之前的详细内容，可以参见 Bilton（2003）的文章。

2.3. 受害人和证人进行的说话人鉴定

以上提到的所有情况都是在已经获得了检材语音的情况下进行的。然而，还有一些情况，即没有检材语音只有证人（同时也可能是受害人）提供的对犯罪分子声音的感知情况，比如像强奸或其他侵害受害人的犯罪情况。其中，有两种情况比较典型，一种是证人对犯罪分子比较熟悉的情况，另一种则是不熟悉的情况。一般说来，进行说话人鉴定时，前一种情况要比后一种情况更加准确。

2.3.1. 对熟人进行的说话人鉴定

在第一种情况下，由于证人在案发前就认识犯罪分子，因此，证人能够辨认出犯罪分子，常常还能说出名字。在法庭上，这种类型的语音鉴定（voice identification）结果可以被视为常规的证人证言，同样也是优缺点并存（Sporer 等 1996）。然而，在一些案件中，类似语音鉴定陈述的可靠性却受到大家的质疑。影响亲听证人陈述（earwitness statements）可靠性的因素包括信道、说话人（犯罪分子）或听话人（证人）等方面。

信道限制包括多种情况，比如说话人与听话人之间的距离很大，相对于周围环境噪音和其他气压变化来说，目标语音的振幅会相对较低。还有很多其他情况，比如目标语音被较强的噪音声源遮盖了，说话人与听话人之间存在障碍物，致使那些对于言语和语音识别很重要的部分共振信息被过滤掉了

（其中也包括像犯罪分子戴面具和安全帽等类似的情况）。研究显示，与说话人和听话人直接接触的情况相比，电话录音的语音信号都经过了带通滤波，因此对电话录音进行的语音鉴定在某程度上可靠性会低一些（可以登录网站 http://www.mml.cam.ac.uk/ling/research/voice_similarity.html 参考有关该领域正在进行的一项研究课题）。对语音鉴定来说，会不时产生很多困难和问题，特别是当出现信道*错配*（channel *mismatch*）时。比如说，证人是在很好的信道条件下听到已知嫌疑人的声音，但却是在很差的信道条件下听到犯罪分子的声音（这一点与对之前不熟悉的语音进行鉴定的情况更为相关）。类似的错配问题也包括言语风格的差异（如喊声与非喊声），现在我们就来讨论这个问题。

说话人因素包括犯罪分子使用特殊的说话方式说话或者说话时间很短的情况。为了使听话人不能辨别自己的声音（也并非一定如此），有些犯罪分子故意使用很极端的说话方式进行伪装，比如说使用假声（falsetto voice）说话。但同时还有更多非伪装状态下产生的语音也会给语音鉴定带来困难，比如叫喊状态下说话（还有像带有明显的感情色彩的语音或强调的语音）就是这种情况。Blatchford 与 Foulkes（2006）曾经报道过一个类似的案例：在一起团伙涉枪案件中，两个妇女惨遭枪杀，还有一个人虽然受了枪伤，但最后还是成功逃脱了。犯罪分子在追赶他的时候，叫喊道“抓住他（get him）”，就是根据这句话，这位受害人成功认定了一名犯罪分子，而且还说出了他的名字，因为他们曾经一起在少管所呆过，所以还记得他。Blatchford 与 Foulkes 就此开展了一项实验研究，研究对象是不同时长的喊声话语，其中之一就是该案件中出现的那句只有两个音节的话——“抓住他（get him）”，另一句是包含 12 个音节的句子。在实验中，当让那些在生活中与发音人很熟悉的普通被试进行辨别的时候，发现随着语音材料的时长由长变短，正确识别率也从 81%降到了 52%，这说明喊声与时长是相互作用的。原则上来说，喊声鉴定是可行的，但这对说话人和听话人都非常依赖。这就给我们引入了说话人的另外一个因素：并非每个人的声音都具有同样的特殊性；部分说话人比另外一些人更好鉴定。Rose 与 Duncan （1995）以及 Foulkes 与 Barron （2000）都曾指出在对熟人进行的说话人鉴定中存在说话人的这种效果，但给出的可能性解释却不尽相同。Rose 与 Duncan（1995）认为，当说话人的语音展现出大量的内部变化性时，就可能不易鉴定了（比较本文 3.1.3），因为其中的部分变化与其他说话人的听觉空间存在一定的交叉。而 Foulkes 与 Barron （2000）则认为识别率低的原因与语音特征过少有关系，特别是存在方言和基频变化的情况下。

我们再来看一下听话人的因素。Blatchford 与 Foulkes（2006）、Foulkes 与 Barron（2000）以及其他学者进行的相关研究都曾指出一点，即并非每个听话人都具备能很好辨别语音的能力。因此，他们得出结论说有必要对亲听证人辨别声音的能力进行测试。感知能力也有可能与大脑皮层的周围调控水平有关，而与认知水平的关系较小。当听话人表现出存在与语音鉴定有关的听力方面的问题时，就应该进行听力学测试（audiological testing）。听话人其他方面的一些因素，像语音记忆力（vocal memory），与下文将要讨论的对非熟人进行的说话人鉴定（unfamiliar-speaker identification）之间的联系性更强（当然，上文中提到的与对熟人进行的说话人鉴定有关系的其他因素同样与对非熟人进行的说话人鉴定也有关系）。关于影响普通听话人进行说话人听觉鉴定表现因素的具体研究，可以参考 Hammersley 与 Read （1996）、Hollien （2002） 和 Rose （2002）的文章和著作。

2.3.2. 对非熟人进行的说话人鉴定和语音辨认

如果证人在案发前不认识犯罪分子，那么就会出现受害人和证人对说话人进行鉴定的第二种情况。倘若抓到了嫌疑人，就可以采用一定的程序对他/她的声音进行辨认。该辨认过程与视觉辨认（visual line-up/eyewitness parade）的程序类似，就是将嫌疑人和一组与案件无关的人（陪衬人员）排在一起进行辨认。这里的辨认只能听不能看，而且对嫌疑人和陪衬人员的辨认不是同时进行而是依顺序进行的。语音辨认（voice line-ups/voice parade），即对不同语音进行的听觉上的辨别。事实上语音辨认比起视觉辨认来说更难计划和实施。Ormerod （2001）曾经在一篇面向律师行业的论文中指出，视觉辨认的陪衬人员可以由有经验的警察来挑选，但对语音辨认而言，无论是对陪衬人员的选择还是整个过程的实施都应该由专家（语音学家或语言学家）进行。很多专家学者都曾针对语音辨认的计划和实施步骤提出过一些指导方针（Hammersley 与 Read 1996；Ormerod 2001），但最具体的要数

Broeders 与 van Amelsvoort（1999）的文章中提到的观点了，他们提出的指导方针得到了欧洲很多法庭科学实验室的认可，而且这些实验室都是欧洲法庭科学研究网工作组（ENFSI）的成员。本文仅对该方针以及其他的一些指导方针做一简要介绍。

首先需要谨记的是证人对目标语音的记忆会迅速衰减。这就要求警方要在案发后的第一时间内对听到犯罪分子声音的证人进行采访，而且越快越好，从而尽可能多地从受害人那里获取有关犯罪分子语音特点的信息。这些信息既可以帮助警方寻找嫌疑人，同时对挑选具有类似语音特点的陪衬人员也有益处。再强调一下，由于语音记忆会衰减，因此语音辨认必须及早进行。

在法庭科学领域内引起对语音记忆开展系统研究的是历史上一起著名的刑事案件：1932 年飞行员 Charles Lindbergh 的婴儿被绑架并谋杀的案子（参见 Fisher 2000）。Condon 博士是 Lindbergh 的好友，在本案中，他组织安排了一次与绑匪的会面并交付了赎金，地点在布朗克斯区的一个墓地。Lindbergh 也一块儿去的现场，并且听到了绑匪的声音。绑匪说：“在这博士，这边，这边”（Fisher 2000 第 81 页）。两年半以后，终于抓到一个叫 Bruno Richard Hauptmann 的嫌疑人，当 Lindbergh 听到这个人的声音后说：“那就是当天晚上我听到的声音”（Fisher 2000 第 249 页）。Hauptmann 最终被判有罪并被执行死刑，但是直到现在相关文献中关于 Hauptmann 到底是否有罪的争论还在继续（相关争论参见 Fisher 1999），尽管如此，除了该语音鉴定的陈述之外，很明显还有很多重要的犯罪证据证明 Hauptmann 确实有罪。[4]

心理学家 McGehee（1937）曾报道过一个相关的研究实验。实验中研究人员先让一个发音人在房间的一侧读一段文字，同时让在房间另一侧的几个被试进行听音，发音人和听音被试之间用一个屏幕隔开。相隔了不同时间之后，又让最初的发音人和另外四个发音人（在进一步的实验中，发音人的数量有所变化）给被试读了相同的文字内容，最后要求这些被试在答案纸上判断这五个声音中哪个是他/她们最初听到的那个人的声音。通过实验，McGehee 发现，一天之后，83%的被试能够进行准确辨认，在第一、二周的时间内正确识别率基本维持在这个水平上，最低到 69%，但五个月之后的正确识别率却仅有 13%。最近也有相似的研究结果（Hammersley 与 Read1996 的研究综述，以及上文提到的其他文章）。由此可见，Lindbergh 作为亲听证人的证言必须要谨慎对待。随后的研究发现语音记忆的衰退很明显不是依据一条简单的原则，相关的影响因素有很多。其中一个重要发现就是未知语音时长的影响，时长越短记忆衰退越快。由于 Lindbergh 听到的绑匪的声音只有几个单词，根据时长的这种影响效应，人们可能会对 Lindbergh 产生这样的怀疑，即他对绑匪声音的记忆能不能大于机会水平。还有一个不利因素是当时 Lindbergh 是在一个比较远的地方（相距约 200 英尺）听到绑匪的声音的（Fisher 2000 第 250 页）。值得一提的是，即使 Lindbergh 的亲听证词并不比机会水平高，那也不是他的错，相反而是那些不懂得科学，设想在本案这样不利条件下的声音仍然可以鉴定，同时还允许这样的鉴定结论可以在法庭上作为证据使用的人的错。随后的研究还发现，在非专业人员进行的语音鉴定（naïve voice identification）中，高自信度并不一定就意味着高准确度（Perfect 等 2002）。最后，在 Lindbergh 的案子中，警方的失误并非一处，除了太相信语音记忆之外，还有重要的一点就是单独呈现嫌疑人的声音，而不是通过语音辨认的方法，相比之下，在测试证人的记忆力时，前者更容易促使证人说“是”。

我们知道了语音辨认工作要在案发后尽快进行，那整个过程具体是如何计划和实施的呢？实际上，很多精力都用在挑选合适的陪衬人员身上了。通常情况下，除了嫌疑人之外，还需要五个或更多的陪衬人员。陪衬人员一般要与嫌疑人具有相同的说话人分类特性（比较本文 2.2）。举个例子，如果嫌疑人的说话语调很低，而所选陪衬人员的说话语调都很高的话就不符合要求。与陪衬人员相比，嫌疑人也不能有其他特殊的言语特点。在找到合适的陪衬人员之后，下一步就可以对他们进行录音了。

[4] 新泽西州复审和上诉法院认定 Hauptmann 犯有谋杀罪行，在诸多证据中有三个独立的有罪证据最为重要：其一，在 Hauptmann 家的车库里发现了藏匿的赎金，并且他还使用过；其二，多个笔迹专家认定勒索赎金纸条上的笔迹就是 Hauptmann 写的；其三，为了爬到 Lindbergh 家的婴儿室去绑架孩子，绑匪曾制作了一个梯子，而该梯子所用的木料与 Hauptmann 家的阁楼有关系（Fisher 2000，特别参见第 385 页）。具体说，从木料类型和工具痕迹来看，现场遗留梯子上的一块横杆与 Hauptmann 家中阁楼地板上丢失的一块木板非常吻合（详见 Fisher 1999 第 125 页）。

倘若嫌疑人不配合录音，也可以使用警方对其进行讯问时的录音材料，但同时也要对陪衬人员的录音在技术和风格上做出相应的调整，使之与警方的讯问录音相似，或者也可以直接使用他们接受警方询问时的录音。人员的选择是否合适，比对语音的录制是否成功都需要进行检验，且检验工作都要由不了解案情的人进行，检验形式也多种多样。如果发现某个说话人在某方面的表现突出，就要重新选人、录音，直到整个过程没有偏向性，每个人的声音被选中的几率都是相等的时候才能结束（因此，建议在最初选人的时候要多选一些，以便在排除不合格的人选后，人员数量仍能满足要求）。整个准备工作做完之后，就可以开始进行真正的辨认了。在具体的操作过程中，要求选择不认识嫌疑人的人做放音人，每个人的声音要按顺序进行播放，每放完一个声音就要询问证人这个声音是不是犯罪分子的声音。放音时可以多次重复播放某个声音，但不能回放之前已经放过的声音。这样做的目的不是为了让证人从所有被辨认者的声音中挑选出一个与犯罪分子的声音最匹配的声音（闭集鉴定 closed-set identification），而是让证人能够对每个声音都做判断，判断它们是不是他/她在犯罪现场听到的声音（开集确认 open-set verification）。为了更严格地执行这些原则，还可以进行一次不含嫌疑人的辨认活动。整个辨认过程都应该进行录像或者能够通过单向透明玻璃镜进行观察。这样做的原因之一是证人在听到嫌疑人的声音（而不是陪衬人员的声音）后可能做出一些非言语的反应（如可能出现强烈的情绪反应），因此有必要进行记录或监视。另外一个原因是要确认放音助手在放音过程中没有对证人做出任何动作，暗示是或不是嫌疑人的声音。在法庭上，如果此辨认方法受到质疑且批评有效的话，整个辨认活动也无法重复第二次，因为那时证人回忆起的声音很可能是来自第一次辨认活动，而不是案发当天的声音。[5]

即使对语音辨认的计划和实施做了大量的工作——同时鉴于诸多法律因素——还要再仔细审查进行语音辨认是否现实可行。比如说，如果案件发生已经很长时间了，如果案发时证人听到犯罪分子说话的时间极短，如果出现上文中提到的与对熟人进行声音鉴定有关的任何一种不利因素的话，无论从控方的角度还是从辩方的角度来说，使用语音辨认措施都是不明智的。

3. 语音比对

3.1. 引言

3.1.1. 专家

当检材语音和样本语音都具备的时候，就可以进行语音比对了。具体比对工作由法庭语音分析专家负责进行。这些专家往往都是经过语言学、心理学、言语与听力科学等相关学科（当然也包括语音学，因为在有些国家语音学是一门独立的学科）学术训练的语音学家，另外还要再接受一些法庭科学和技术领域的特殊训练，比如说语音增强。他们常常还要具备一定的实践经历，能够处理实际的案件语料、撰写专家证言报告、在法庭上质证等（可以比较参照Hollien 2002的著作中“专家”一章中的相关内容）。有时候那些使用说话人自动识别技术的工程师或电脑科学家也可以成为专家。相对于由受害人或证人进行的说话人鉴定或者“非专业人员进行的说话人鉴定”（见本文2.3）而言，由于专家的参与，语音比对活动经常被称为“专家进行的说话人鉴定”或者“技术性的说话人鉴定”。当然，不仅语音比对需要合适的专家参与，语音画像以及针对受害人与证人进行说话人鉴定的计划、实施和解释过程也同样需要专家的参与。

3.1.2. 贝叶斯方法

在语音比对过程中，有两点必须谨记。

第一，两个语音在我们比较的语音维度上相似或不同的程度有多大，也就是说话人鉴定中*相似性（similarity）*的问题。两个语音越相似（任何事物都一样），它们源自同一说话人的可能性就越大。在我们对说话人鉴定以及法庭科学其他分支学科的认识中，普遍存在一种误解，即认为只要确定两种比对取样（语音材料）具有较强的相似性，就可以进行认定匹配，得出二者同源的结论。正如Rose（2006a）指出的，我们在美国电视剧《犯罪现场调查》（CSI）和其他电视节目中经常听到检验人员使用“对上了”这个说法来表示特征检验结果。这种

[5]本段总结概括的语音辨认的相关原则主要引自上文已经提到的 Broeders 与 van Amelsvoort （1999）的文章，然而这些指导原则并非都能被大家所普遍接受，另外其他学者也提出了一些颇有见地的观点，为语音辨认的构建与实施提供了一些不错的另选方案。比如说，对要不要让证人在听到每个声音后都做出判断的程序上就有一些不同的看法（可以比较 Nolan2003 所做的进一步的讨论和例证）。

“对上”或“没对上”的说法实际上暗含了一种绝对肯定或绝对否定的判断。然而，在现有的法庭科学实践中，有学者提倡使用概率的形式（口头或定量）来表述结论，而非绝对的表述形式（少数国家除外）。要证明这种绝对比对的合理性，必须先证明任何不同取样之间的匹配类型都是惟一不可重复的。一直以来，指纹分析就以此假设为前提。然而，最近这种假设却遇到一些问题，以至于有人认为就连指纹比对都应该（也经常这样做）使用概率的形式来表述。这种概率表述的思想已经在DNA分析领域占据了支配地位，尽管如此，它仍然是一门非常年轻的学科。

第二，即*普遍性（typicality）*的问题。鉴定专家在分析检材语音与样本语音时会使用一些语音特征，这些特征在所有人的说话特点中可能比较常见，也可能比较罕见。这种普遍性越低，语音证据的证明力就越强（任何事物都一样）。本文 3.2.2 在分析基频时还将针对高、低两种普遍性的情况进行图解分析。

比较相似性与普遍性的概念和数学方法就是大家常说的*贝叶斯方法*[6]。此方法的核心概念是似然比（Likelihood Ratio，简写为LR），具体公式是（参见Rose 2002 第58页）：

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)}$$

该公式中，当起诉假设（prosecution hypothesis）正确时（即两种取样同源），分子表示获取已知语音证据 E 的概率；当辩护假设（defense hypothesis）正确时（即两种取样不同源），分母表示获取同样证据的概率。如果 LR 的数值大于 1，说明有相对更多的证据支持两种取样同源，反之，如果 LR 的数值小于 1，说明有相对更多的证据支持二者不同源。[7]分子代表比对的相似性层面，分母则代表普遍性层面，对于两种取样来说，如果二者的相似程度高，那表明二者同源的可能性也相对较大，反之亦然；当普遍性较高时，其他实体（这里指说话人）与未知取样同源的可能性也就会相对较大，反之，当普遍性较低时，除了嫌疑实体以外其他实体涉案的可能性也相对较小。

自从 20 世纪 90 年代以来，贝叶斯方法已经在法庭科学领域得到广泛认可，并在 DNA 分析技术的发展以及提交法庭的分析结果表述上起到了非常重要的作用（参见 Robertson 与 Vignaux 1995 的文章做简单了解）。在法庭语音分析领域，贝叶斯方法最早同时也是主要应用在说话人自动识别上（Meuwly2000），之后不久，又有学者（Nolan2001 和 Rose2002）曾探讨过贝叶斯方法与声学和听觉-语音学分析方法应用的关系问题。有关分析方法的具体内容将在 3.2 部分进行解释。

语音比对中，创建或获准使用有关言语特性的*人口统计数据（population statistics）*是对LR中普遍性进行量化的必要前提。说话人自动识别中确实存在一些统计数据（详见本文3.2.4），但真正与语音特性相关的却很少见（详见本文3.2.2）。甚至在一些地区没有人口统计数据可查，那量化问题也就无从谈起了，此时贝叶斯方法只能作为一种概念框架中的方法为语音比对分析提供逻辑上的支持了。Rose（2002，2005）在其著作和文章中比较全面地介绍了贝叶斯方法在法庭语音学中的具体应用。

3.1.3. 质量/数量限制和错配问题

相似性、普遍性以及人口统计数据的重要性对任何形式的语音比对活动来说都是十分必要的组成部分，当然也包括由商业原因或纯学术研究为目的的情况，这些情况下录制的语音材料，其语音质量和材料数量常常都是无可挑剔的。然而在实际案件中，情况却大不一样，语音材料的质量和数量往往存在诸多限制，这便会给比对方法的使用和结论的得出产生一定的影响。而且实践中常常遇到这样的

[6] 该方法是以 18 世纪英国汤布利韦尔斯镇的长老会牧师、数学家 Thomas Bayes 的名字命名的。Kaplan 与 Kaplan（2006）的著作是有关贝叶斯方法在司法系统中应用的一本科普读物，值得一读。

[7] 这里之所以使用修饰语“相对”一词是因为最后做出支不支持认定结论所依据的概率还取决于“先验概率（prior odds）”的大小，而此“先验概率”又与实际案情有关（比如说证人的陈述，有无不在犯罪现场的情况等），与语音分析无关。请大家注意：（在考虑先验概率的情况下）最后结果的表述形式是基于现有证据所做出的认定或否定结论的概率，而 LR 表示的却是基于认定（分子）或否定（分母）的证据的概率问题。因为这里主要讲 LR 的问题，所以解释得比较简单，详细内容可以参考 Rose（2002）的著作。

问题，即与在可控的录音室或实验室条件下录制的语音相比，案件语音中同一说话人语音内部的变化要大得多。所谓“同一说话人语音内部的变化（intra-speaker variation/within-speaker variation）”是指同一说话人在不同场合不同环境下说话时语音存在的差异（主要指语音比对分析的各种参数）。这种变化客观存在，比如当一个人大声说话时，他声音的音高就比他正常说话时高。用贝叶斯原理中的术语来讲，如果这种内部变化很大的话，即使是在确定包含了同一说话人在内的情况下，相似性也会很低，由此 LR 公式中的分子也会变小，这就可能导致误拒绝（false rejection）的产生。这种说话人语音内部的性质变化问题，在 DNA 分析和指纹鉴定中并不存在。正是这种复杂性的存在使得说话人鉴定（或任何形式的行为分析）变得非常困难。

同一说话人语音内部的变化可以有很多表现形式，其中部分变化对语言系统来说还是不可或缺的。比如说，一个人发音的音高及其主要声学相关物基频就不是一成不变的，而是随着语调和声调类型（对声调语言来说）的变化而变化的。再如，由于协同发音的作用，元音的共振峰频率也会受到相邻语音的影响而发生变化。这就意味着在语音比对过程中，不仅要选取那些在语言学意义上相同的语音片段进行比对，而且语音材料还要有足够的数量，这样才能保证检材与样本语音所受的语言学影响差不多。案件语料中同一说话人的语音常常会表现出大量的内部变化性，令人难以把握。总的说来，引起这种变化的因素可以归为两类：副语言上的与风格上的。比如像刚才已经提到的说话效果（大声和轻声说话）的影响，其他还有不少因素也会导致这种变化的产生，像紧张与情感因素的变化、朗读和正常自然说话方式的不同、醉酒和清醒状态的不同、感冒和健康状态下的不同等等。实践中，由于这些副语言或风格因素的影响，常常会导致错配的产生。有时候，检材语音中未知说话人的情绪激动，说话声音很大，而嫌疑人说话时的声音很小，好像很厌烦的样子。处理类似的错配问题有两种方法可供选择：或者是选择那些受影响最小变化不大的语音片段进行比对，或者是运用一定的语音学知识，分析其影响规律，从技术或概念上进行一定的补偿。

在同一说话人语音内部的变化中最极端的形式要数说话人故意进行*语音伪装*（*voice disguise*）的情况了。当未知说话人为了隐藏身份伪装发音而嫌疑人却没有伪装正常说话时，语音错配现象就会大量增加。有时候甚至还可能发生相反的情况（检材语音没有伪装，嫌疑人却不正常说话），此时嫌疑人的伪装手段常常隐藏的更加巧妙（这并不意味着会使分析变得更容易）。如果撇开司法鉴定来看，单说语音伪装的问题，其实很有意思，因为它恰恰凸显了发音器官的灵活性以及语言使用者的创造性。许多伪装方式都是由改变与声带（vocal folds）有关的音高或嗓音音质的方式实现的。20 世纪 80 年代末 90 年代初，德国发生过一起很有名的案子，犯罪分子使用化名 *Onkel Dagobert*（在德语中相当于迪斯尼动画中的“唐老鸭叔叔”）向一些大型百货公司勒索钱财，扬言不给钱的话就实施爆炸，所幸没有造成人员伤害。犯罪分子在实施勒索谈判时就是使用高音调假声（high-pitched falsetto voice）来伪装自己的。其他伪装方式还有改变说话速度、通过捏鼻或咬物等方法改变上呼吸道的发音特点等等。有时伪装也会只体现在语言学层面上，比方说模仿外国人的口音或者同时还使用外语说话。尽管有伪装存在，通常情况下还是可以进行语音比对的，因为很少有说话人能够采用多于一或两种方式进行伪装的。而且，在整个对话过程中，由于身体、情感或认知疲劳等原因，说话人常常不能一直保持伪装状态。由于伪装还会导致语音的可懂度降低，当犯罪分子发现其说话的可懂度降低到一定程度，以致无法实现其交流目的的时候，也会减小伪装程度或停止伪装。

刚才探讨了案件材料中发音动作（言语产生）方面存在的局限和问题，实践中还有很多技术层面的局限和问题。比如，在几乎所有的语音比对案件中，无论是检材语音还是样本语音至少有一个会涉及电话录音的问题。对于电话通讯而言，无论是使用固定电话还是手机，语音信号的频带范围都被限制在 200 Hz 到 3500 Hz 左右之间，这就意味着，在实际办案中，任何超出这个范围的语音特征都是不能用的。正常情况下，男性说话人的基频范围在 80 至 170Hz 之间（低于 200Hz），不过这倒也不是问题，因为在电话的频带范围内仍会出现大量的谐波特征，而基频的数值可以通过测量谐波间的距离计算出来。真正的问题是元音第三共振峰以上的共振峰以及像[s]之类的摩擦辅音，它们的频谱会受到影响。从研究的角度讲，分析这些反映说话人特性的信息很有意思，但在实际办案中却并不适用（当然如果电话公司决定扩大电话通讯的频带范围时，这

些局限就不成问题了）。对于电话录音而言，刚才提到的理想频带范围在实际办案中可能会进一步减小：由于传输效果不好或录音设备的质量欠佳等原因，有时该范围的频率上限会进一步降低，频率下限也会进一步提高。除了频率方面的限制之外，鉴定中还可能会遇到一些语音畸变方面的难点，引起畸变产生的原因有很多，像录音级别设置过大、手机信号不好、回声较大、背景噪音很强或语音叠加等都会给语音比对带来一定的困难。另外，还有一些问题也不容忽视，像模拟录音在录制和回放转换时的速度问题，对录音采样进行模/数和数/模转换时的采样频率问题都有可能产生错配问题。

另外一个技术问题就是时长限制。一般人常常会误以为，人们可以从一段时间非常短的话语中辨别出一个人的声音，那怕只有一个元音。然而，稍微有点语言学或语音学知识的人都知道时间太短对语音辨别来说有很大影响。时间越短，录音材料中反应出的语音特征在说话人的所有语音特征中的代表性就越小。如果时间太短，可能就反映不出多少说话人的个性特征了。同时，如果录音中没有出现具有“分析价值”的词项，比如说音节末尾带/r/音的单词，那么方言分析将受到极大限制。鉴定实践中，尽管在录音材料时间长度的要求上没有固定限制，但一般推荐以检材语音不少于 8 秒，样本语音不少于 16 秒为最低标准。但这也并不排除有些特征反映非常明显的录音材料，只有 2 秒钟也可以进行鉴定的可能。

通过上文对案件语音中存在的诸多限制和问题的总结分析，我们可以得出两个结论。第一，由于这些局限和问题是客观存在的，那么我们就应该把精力集中在对最具鲁棒性的语音参数的研究与开发上，或者寻求克服这些不利影响因素的新途径（Nolan 1983 第11至14页；Rose 2002, 51f）。比如说，在实际办案中会不可避免地遇到手机通讯的问题，我们就有必要着重研究一下通讯技术是如果影响语音比对方法的，像对共振峰频率测量的影响等。有学者（Byrne与Foulkes 2004）做过这方面的研究，结果显示与高保真的语音相比，手机录音的第二、第三共振峰的频率基本没有受到什么影响，而第一共振峰的频率却有所提高。第二，要对每个案件的检材语音中存在的局限和问题进行仔细的分析评估，做到具体问题具体分析。在此基础上，首先看检材语音是否具备鉴定条件，如果具备条件，再分析有哪些语音参数可供检验，最后再看能够得出何种程度的鉴定意见。

3.2. 语音比对中使用的不同方法

在语音比对中，不同鉴定专家、实验室之间使用的分析方法可能会有所不同。关于不同的分析方法，表1中列出了一种分类方式。

表1 法庭语音比对中不同方法的分类

	解析法	整体分析法
听觉-感知	范畴语音学标音与描写	整体语音感知
声学	声学语音学	说话人自动识别

解析法（analytical approach）是语音学中最常用的方法。该方法首先将整个语音分成若干组成部分，像语音、语音特征、韵律特性等，然后再分别对部分成分进行分析或研究它们之间的相互关系。使用听觉-感知分析法（auditory-perceptual approach）时可以使用国际音标（见国际语音学会 1999 版）对语音成分进行标音。而声学-语音学法（acoustic-phonetic approach）（见 Stevens1998）首先要对语音进行听觉-感知范畴的分类，在此基础上再分析这些感知范畴的声学表现。通过声学-语音学分析，我们常常发现，与感知层面相比，在实际声学层面上，语音的渐变性更容易量化描述，语音分隔的范畴也更小。对实际鉴定而言，声学-语音学分析法的优势在于它可以提供非常准确的数值，而在这一点上，听觉-感知分析法是绝对做不到的。然而，这并不是说数字越精确，鉴定就会越准确。

感知或声学分析应该着重研究那些能够反映不同说话人之间较大个性差异的语音特征，即那些能够反映“不同说话人语音之间差异（inter-speaker variation/between-speaker variation）”的特征。而且，还应该包括那些具备一些解剖-生理学基础（能够反映不同说话人语音之间的差异）的特征。就像音高（听觉-感知范畴）或基频（声学范畴）都是由说话人的声带长度决定的，由于声带长度因人而异，因此每个人的音高或基频参数也都不一样。然而，即使部分特征有一些解剖-生理学基础，但也不完全受

它们的支配，还是会受一些语言学和副语言学的影响而发生变化。一方面，这些语言学和副语言学因素可以提供更多反映说话人个性的特征信息，但另一方面也会带来一些不必要的麻烦，额外出现一些同一说话人语音内部的变化现象（比较Braun 1995关于同一说话人语音内部变化影响基频的文章）。

进一步说，在运用表中的整体分析法（holistic approach）时，并非要在心理或技术上拆分语音，而是将其看成是一个整体。应用到听觉-感知领域中，就要集中精力着重分析两个录音材料是不是一个人说的，而不是努力去做成分分析。一方面，此方法有点类似于非专业人员进行的说话人鉴定，另一方面，在对语音进行整体听觉分析时，语音专家具备一些非专业人员所不及优势，下文将对此进行详细分析。在表1列举的方法中，还有一种方法没有说到，即运用声学处理方法的同时对语音作整体分析，在说话人自动识别当中，就是综合运用的这两种方法[8]。这里说的“整体”分析法，也可以使用“综合”、“非解析”、“格式塔式”、“全面”分析法等叫法。

相对来说，使用整体分析法进行说话人鉴定的成功率更易计算，因为比起解析法来说，前者更加快捷，因此适用范围也更广。事实上，有学者指出，整体分析法能够很好地应用于说话人鉴定实践（Alexander 等 2005）。然而该方法并不能满足语音专家和法庭的科学兴趣与要求，因为该方法对于一些问题的解释并不能令人完全信服，比如两个特定语音相似或不同的确切依据是什么，或者说一个特定语音在一定人群中常见或少见的根据是什么等问题。

在表 1 列举的 4 种方法中，哪种分析方法在鉴定实践中能够最准确地区分说话人呢？迄今为止，对于这个问题，学界还没有进行过系统研究。鉴于此，单纯使用一种方法进行鉴定是很不明智的，很可能使用两种都不够。而且，从说话人鉴定科学基础的角度上看，完全依靠整体分析法进行鉴定也是行不通的。在目前的鉴定实践中，常常是综合使用除了说话人自动识别方法之外的其他 3 种方法。德国联邦刑事侦查局（BKA）和德国其他地区的鉴定机构综合使用 3 种方法进行鉴定已经大约有 20 年的历史了。英国、瑞典、芬兰、荷兰和奥地利等国家的大部分实验室和私人从业者也都这样做。虽然说话人自动识别在法国和西班牙得到了广泛的认可，但对此我们还是应该抱以非常谨慎的态度，若要将此方法应用于实际鉴定还要进行更加严格的测试，目前我局正进行这方面的研究工作。下文几节将对表 1 中列举的 4 种方法作具体介绍。

3.2.1. 范畴语音学标音与描写

范畴语音学标音与描写（categorical-phonetic transcription and description）在鉴定实践中应用的一个重要领域可以用*面向语言学的语音学（linguistic phonetics）*这个术语来概括（比较 Ladefoged1971）[9]。这样说的原因在于，面向语言学的语音学不仅涉

[8]另外还有一种方法，即“声纹”分析法（“voice print” method），可以用“整体视图”的特点来概括。该方法基本上主要是对检材语音和样本语音中部分相同语音片断的声谱图进行比对，比对依据也主要是基于对两份声谱图相似点和差异点所进行的“目测”。因为该方法主要是在一些培训班里集中讲授的（课时较短），对象主要是一些没有语音学和语言学背景的人，所以很明显，单纯的图谱比对并不是由那些受过适当训练的人做出的声学语音学的系统分析。“声纹”分析法已经有很长的历史了，很多学者像 Nolan（1983）, Hollien（2002）与 Rose（2002）都在其著作中讲到过该方法。基于多种原因，很多业内人士都不再使用该方法了，不少法庭也不再采信该方法。法庭语音学协会的所有成员也都一致认为不能使用该方法。 然而有一点必须注意，虽然我们不用“声纹”，但不代表就要从整体上摒弃声谱分析（spectrographic analysis）甚至是声学语音学分析法了。然而遗憾的是，这种情况确实在美国的一部分州里出现了，这种“把洗澡水连同小孩一起倒掉”的做法实不可取（与 Foulkes 进行的个人交流）。（译者注：原文所说的“voice print” method（“声纹”分析法）的概念与国内常说的声纹鉴定方法完全不同。随着各方面技术的发展和进步，国内声纹鉴定领域也早已摒弃了单纯的图谱比对的方法，业内一些学者总结的“看”、“听”、“测”以及近年来兴起的说话人自动识别方法集中概括了目前国内声纹鉴定从业人员所采用的主要鉴定方法（与本文表 1 中列举的几种方法类似）。尽管“声纹鉴定”的叫法无论在业内还是在社会中使用都非常普遍，但近年来在该技术的称谓上也确实存在一些不同的观点。）

[9]这里将“标音”与“描写”结合使用的原因在于，在鉴定报告中并不要求必须要进行语音学标音。对于关键性的发音来说，还可以使用国际音标字母表的分类矩阵进行描述（例如讨论后齿龈轻擦音时就可使用描写的方式而不使用它的音标符号），也可以

及到语音的音段特征，还涉及超音段特征，它们既可以区分不同的语言，也可以对一种语言中的不同方言进行区分。面向语言学的语音学标音与描写对于说话人分类有重要作用，特别是在判断不同的地域方言、社会方言（此处可以这样理解，任何本土语言的变异都与一定的社会结构相关）和外国口音的时候。如果在案件的初期侦查阶段，其他部门的侦查人员未能区分犯罪分子说的是何种语言（或方言）的话，那么此时还要进行语言（方言）识别分析，这也是该领域的研究内容之一。然而，世界各地许多地区的语言与方言，甚至不同语言之间可能都很难区分，因此如果真遇到难题的时候，最终还是有必要寻求法庭语音学家（或语言学家）的帮助。正如本文 2.2 中提到的，对说话人进行分类的目的不仅是为了语音画像，同时对语音比对也很有意义。在语音比对过程中对说话人进行面向语言学的语音学分类具有双重作用。首先，如果两个说话人的方言、社会方言或外国口音特征不同的话，它能够在排除说话人方面提供强有力的证据。其次，即便常常不能单凭说话人分类的结果就做出鉴定结论（因为从定义上讲，它提供的是说话人整体分类的信息），但却可以大大缩小侦查范围，特别是对于那些同时含有多种分类信息的情况。笔者在办案中就遇到过一起贩毒案件与此类似，犯罪分子和嫌疑人都说德语，但话语间又都带有一些俄语和斯瓦比亚的地方口音，类似的组合分类信息可以大大缩小侦查范围。

在对说话人进行面向语言学的语音学分类时，尽管主要使用听觉-感知的方法，但并不仅限于这一种方法。Ladefoged 与 Maddieson（1996）曾介绍过很多在面向语言学的语音学中成功使用声学语音学方法进行研究的例子。这样的例子还有很多，许多文献中都介绍过不少在方言学、社会语言学、第二语言语音学中应用声学语音学方法进行研究的例子（见 Foulkes 与 Docherty 2006，比如声学方法在社会语音学中的应用）。值得注意的是，各个语言变体甚至各种语言之间的差异并非一定局限于范畴方面，而且还会涉及一些渐变特点和纯粹的语音细节（比较 Keating 1990 的文章以及他后来在“实验音系学”方面做的研究）。

法庭语音学家在语音比对（在说话人分类中没有涉及）过程中使用范畴语音学标音与描写方法时还会涉及一个领域，即*嗓音音质*（*voice quality*）的研究。每个人在发声时的嗓音音质都不一样，比方说，有些人可能是中性（也就是“正常的”）嗓音（neutral/modal voice），有些人的音质也可能发生明显偏离，产生其他类型的嗓音，像气嗓音（breathy voice）、挤喉音（creaky voice）、粗糙嗓音（rough/harsh voice）和紧音（pressed/tense voice）（或者几种嗓音类型的不同组合）等。而且，对于同一种嗓音音质，每个人发声时的强弱程度也可能不同，像气嗓音就有强弱之分。到目前为止，我们仅讨论了由喉部产生的嗓音音质问题，但实际上，说到嗓音音质，从广义上讲它还应该包括由上呼吸道产生的音质，比方说鼻音（对比注脚 1）。Laver（1980, 1994）曾对由喉部和由上呼吸道产生的嗓音音质的描述和分类进行过相关的理论和实践研究。另外，还有一个仅对喉部嗓音音质有效的听觉-感知方法也很重要，即 GRBAS 分级评估法（其中，G 代表对总嘶哑度的整体感知分级；R 代表粗糙度；B 代表气息度；A 代表无力（弱）度；S 代表紧张度，即类似于挤喉音）（Hirano 1981）。虽然 GRBAS 分级评估法是在嗓音医学（研究嗓音障碍和言语障碍的医学学科）中发展起来的，但是现已证明该方法在鉴定领域也很有用。目前，Laver 的方法和 GRBAS 分析法（Nawka 与 Anders 在德国的嗓音医学方面做了修改（1996））BKA 都在使用（Köster 与 Köster 2004）。近期的研究显示，如果评估人受过相关训练并具备一定的办案经验，就可以有效提高感知嗓音音质的准确率和不同评估人之间评分的一致性（inter-rater agreement，即不同听话人之间的感知差异相对较小）（Köste 等 2007）。

Nolan（2005）曾对在法庭语音学中运用嗓音音质进行鉴定的局限性做出过重要评价。比如说，由于电话通讯中的频带滤波限制（频率范围限定在大约 200 至 3500Hz 之间），很多信息在通话中都被滤掉了，而这些信息无论在声学上，还是最终对某些嗓音音质的感知上都有重要作用。像对气嗓音的感知，在某种程度上主要依靠第一谐波中携带的信息，但对于男性说话人来说，其语音的第一谐波的频率常常不在电话频带范围之内。另外，当录音中有噪音时，也会对气嗓音的感知产生干扰。正是由于类似问题的存在，在鉴定中分析嗓音音质时，鉴定专家更常使用听觉-感知的方法，而不是声学-语音学的

使用能够让法庭相关人员更易直接理解的多种方法进行标音或描写。而且，对于嗓音音质和一些其他的韵律现象来说，语音描写比标音要更容易一些。

方法。但是，正如 Nolan 指出的，如果关键性的声学成分丢失了（如因为电话传输所致）或有其他非言语声学信息的干扰，那么对嗓音音质的感知评估过程就不得不依赖“一项很详细复杂的感知重建过程，即要重建语音材料在没有经过电话传输时的情况。”（Nolan 2005 第 396 页）。需要说明的是，这种重建过程有时比较难以把握，如果出现偏差，则会使嗓音音质的感知效果变得更糟，甚至对于高质量的语音信号而言，也可能使不同评分人之间的意见分歧变得更大（Kreiman 等 1992）。Nolan 进一步指出，嗓音音质部分是由方言差异和说话风格变化所决定的。在语音比对中着重研究说话人之间的差异时，应该在感知或思维上将这些影响因素提取出来。在鉴定中分析嗓音音质时应该确认是否存在这些问题并想办法加以解决。

语音学标音与描写的方法还可以应用在语音比对的其他领域当中，像对*填声停顿（filled pauses）*方式的分析，所谓填声停顿，即英语对话中常出现的像“呃（uh）”和“嗯（um）”之类的停顿（Clark 与 Fox Tree 2002）。人们在说话时常常意识不到自己类似的说话方式，即便是意识到了，很大程度上也控制不了。这对伪装语音的检验来说就有作用了，犯罪分子可能会将精力集中在某种伪装方式（如假声）的实现上，但却不能控制自己某些说话方面的习惯，像方言或填声停顿等特点。填声停顿的特点因人而异，人们说话停顿时带不带鼻辅音，有没有出现声门化、鼻化和精确意义上的元音音质等都是不一样的，这些现象都可以进行语音学描写。同时，填声停顿在其他方面也会表现出一定的差异性，如在一定时间间隔内的时长和频率等参数，这些参数在声学信号中是比较容易测量的。

3.2.2. 声学语音学

对部分语音学领域来说，声学分析比听觉分析要更准一些。在众多说话人鉴定用到的参数中，最典型的就是声音的平均基频（f_0）了。基频是嗓音产生过程中声带振动频率的声学相关物。声音在平均基频上的不同，从感知角度可以由“高音调（high-pitched）”和“低音调（low-pitched）”之间的范围进行描述。然而，与对基频做声学分析的范围相比，听觉-感知范围的精确度就差多了。而且，它也不能排除声音的“高-低”与“轻-暗”之间会出现混淆的可能性，其实，声音的“轻-暗”反映的是说话人声道长短的差异，而不是声带振动快慢的不同（比较 Jessen 2007b 文中对此问题的表述）。做声学分析就不会出现这样的混淆现象，不仅是因为声学分析更加准确，还有一个原因就是与对具体音高的感知评估相比，分析基频还更加客观。目前平均基频已经存在一些人口统计数据，除此之外其他方面的统计数据还不多见。在法庭语音学的研究领域中，Künzel（1987，1995）是最先从事这方面研究的[10]，他曾对说德语的 100 位男性说话人和 50 位女性说话人在朗读方式下的语音进行过统计。最近，Jessen 等（2005）也对说德语的男性说话人做过类似的研究，结果与 Künzel（1987）的研究结果类似。不仅如此，Jessen 等（2005）还对不同的说话方式进行了区分和统计，特别是对朗读发音与自然发音、使用正常音量说话与大声说话时语音的差异进行了研究统计。大声说话时的语音是在噪声听觉反馈实验（Lombard experiment）中得到的（实验中要通过耳机给被试播放音量较大的白噪声）。该实验还对录音过程中基频的标准差进行了数据统计[11]，该参数可以反应出不同人说话旋律感的强弱。

图 1 平均基频的部分实验结果

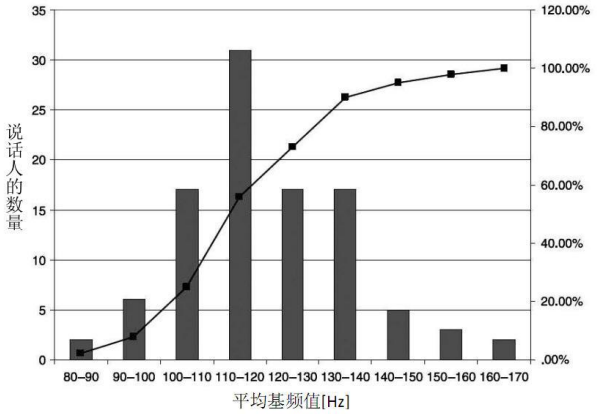


图1 100位男性说话人自然发音，使用正常音量说德语时的平均基频（X轴）分布图，其中灰色柱形图（左侧Y轴）表

[10]参见 Künzel（2000）的文章，作者让被试使用了多种不同的伪装发音方式说话，并从人口方面进行了研究。

[11] Jessen 等（2005）的研究显示平均基频与基频的标准差呈正相关关系。如果使用标准差来表示变异系数（标准差除以平均数）的话，这种相关性就会基本消失。Rose（2002 第 52 页）曾指出，为了得到更高的 LR，不同语音参数之间的相关性最好是越小越好，因此，使用基频变率（f_0 variability）来表示变异系数更可取。

示人数，黑色斜线（右侧Y轴）表示积累百分比。

图1（译者注：原文中的图示均为彩图（下同），由于印刷局限，译文中的图示及文中描述均作了部分调整）显示的是说话人在自然发音状态下使用正常音量说话时平均基频的分布情况。该图可以用来解释贝叶斯方法（见本文3.1.2）中普遍性的基本原理（比较Rose2002第306–310页）。在此，我们假设有两个案子A与B，在案件A中未知说话人的平均基频是114Hz，嫌疑人的平均基频是117Hz；案件B中未知说话人的平均基频是160Hz，嫌疑人的平均基频是163Hz。两个案子中的“相似性”是一样的（都有3Hz的差别，在可预期的同一说话人语音的内部变化中已算是比较好的了），因此，此时LR公式中的分子也是相同的，而且是比较高的；我们假设两个案子中分子都是1。但是，两个案子中的“普遍性”却并不一样，案件A比案件B的更高。在案件A中，除嫌疑人以外的其他人与未知说话人同一的可能性有0.31（见图1所示，在所有说话人中，有31%的人平均基频在110–120Hz之间，正是案件A中涉及的频率范围），用1除以0.31，得到案件A的LR是3.2，此结果仍然支持认定未知说话人与嫌疑人是同一人的说法（因为LR > 1），但是LR并不是很高，因此，该证据的证明力相对弱一些。在案件B中，除嫌疑人以外的其他人与未知说话人同一的可能性只有0.02，因为根据图1显示的结果，100个说话人中只有2个人的平均基频分布在160～170Hz之间。我们用1除以0.02，得出的LR是50。很明显，与案件A相比，要认定嫌疑人的话，此证据的证明力要强多了。

图2显示的是不同音量级条件下平均基频分布的显著差异情况。比如说，如果未知说话人是大声说话，而嫌疑人却使用正常音量说话时，就不能对二者进行直接比对（正常情况下，尽管不常见，至少有部分语音的音量是相同的）。这说明同一说话人语音内部的变化也会导致基频发生很大的变化（见Braun1995的文章做概括了解）。这确实也是基频参数在鉴定实践应用中面临的主要问题，因此就要求鉴定专家在使用基频特征时，必须要与许多其他语音参数结合使用才能下结论。最近，Hudson等（2007）又对100个讲英国南部英语的男性说话人平均基频的分布情况进行了人口数据的统计。

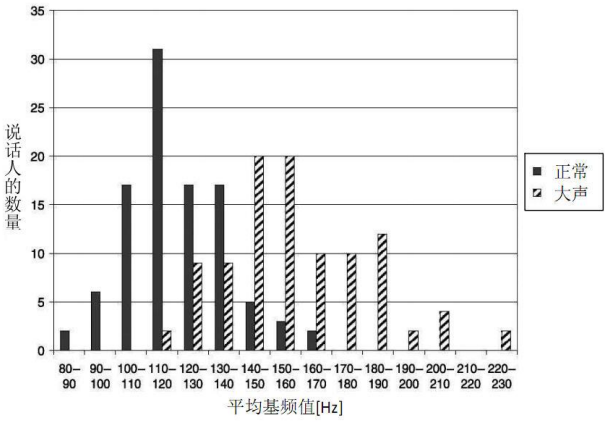


图2　100位男性说话人自然发音，分别在正常说话和大声说话两种状态下说德语的平均基频的分布情况。其中，实芯柱形图表示的是正常说话，带阴影的柱形图表示的是大声说话（在噪声听觉反馈实验中得到）。

Jessen（2007a）还曾对德语的*发音速率（articulation rate）*问题进行过人口数据的统计。发音速率（用每秒钟发出的音节数量表示）是说话人不同言语音速（speech tempo）的量化，严格说来，它还是听觉分析和声学处理的混合物：音节数量的计算是在听觉-感知的基础上进行的，而对一段话语持续时间的测量却需要声学手段。在之前对言语音速参数的研究中，“音节速率（syllable rate）”（也叫做“言语速率（speech rate）”）都要考虑停顿问题，而“发音速率”却忽略所有的停顿，只研究在连续语流中的情况。研究发现，与“音节速率”相比，在“发音速率”这个参数上，不同说话人之间的差异更大，而同一说话人内部的变化更小（见Goldman Eisler 1968以及Künzel 1997）。因此，“发音速率”参数在鉴定实践中会更有用处。

另外，*共振峰频率（formant frequency）*在使用声学语音学方法进行说话人鉴定时也是非常重要的参数之一。尽管迄今为止学界对共振峰相关人口数据的统计研究还不像对基频和发音速率的研究那样多，但目前也有不少学者正对此进行立项研究，其中剑桥大学开展的研究项目最著名（见www.ling.cam.ac.uk/dyvis）。共振峰频率不仅是区分不同辅音和元音的重要相关物，同时还携带了很多说话人的信息（比较Stevens 1998）。共振峰频率要受声道长度（即喉部到嘴唇的距离）的影响，这是其解剖学基础之一。声道变长会导致元音共振峰降低，这正是女性说话人的共振峰频率一般要比男性说话人高的原因所在，当然同性之间也存在着个

体差异（比较 Rose 2002）。决定共振峰频率的因素不仅只有声道长度，同时与声道各部分之间的比例大小也有关系。研究发现，与成年女性相比，成年男性的咽部与其声道其他部分相比会更长一些（Fitch 与 Giedd 1999），这也可能会在同性人群中产生一些个体差异。

测量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共振峰的中心频率值的方法。Rose（2002, 2006a,b 以及 Rose 等 2003）等对该测量方法如何用在对语音证据进行的贝叶斯评估中做过介绍。正如 Rose 指出的，此方法的难点，同样也是其潜在优势，在于不同元音的共振峰频率之间以及不同共振峰之间（特别是第二和第三共振峰，由于电话带宽的限制，其他共振峰要么不在其频带范围之内，要么会受到影响而发生变化）根本没有充分的相关性。为什么会缺乏高度的相关性呢？其中一个可能的原因是由于条件问题，如前文所述，说话人语音的个体模式不仅与其整体声道长度的差异有关，同时还与声道各个部分之间的差异有关。研究这种有限的相关性需要用到复杂的数学统计模型，尽管情况复杂，但同时也为通过对不同共振峰和元音的综合分析，得到较高的似然比提供了机会。

研究共振峰的另一个方法是针对其动态特性进行分析。McDougall 与 Nolan 对此做了大量的工作，最近也取得了一定的研究成果（McDougall 2004, 2006 以及 McDougall 与 Nolan 2007）。正如 McDougall（2006）指出的，我们似乎可以做出如下假设：尽管共振峰的目标值（测量共振峰的中心频率所得到的那些区域）很大程度上取决于一种语言音系上的要求，但人们说话时发音动作在不同目标值间过渡的时候都有一定的自由度。因此，人们发音时应该会留下其特定运动方式的部分痕迹，这些痕迹也能反映说话人的个性特征（一般来说，声道形状在静止状态下的差异决定说话人语音的个性特征）。这些预言已经得到证实，共振峰的动态特性在区别个体方面确实具有很高的价值。由于共振峰的动态性同时还受到音段和韵律语境的影响，特别是在自然说话时表现得很明显，有必要对如何处理这些产生同一说话人语音内部变化的来源问题进行进一步研究。

Nolan 与 Grigoras（2005）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）。该方法不是选择分析具体的目标元音，而是提取一整段语音的全部信息，当然其前提是这些录音资料可以利用线性预测编码（LPC）技术计算共振峰，并能得到可见且可靠的共振峰结构。在语音信号质量足够好的情况下，该方法常可以将所有或大部分元音和滑音的共振峰信息提取出来，而不是仅从一部分语音中提取信息。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，而如果对不同元音分别进行分析，然后再结合贝叶斯方法进行分析的话却可能得到这些特征信息（正如上文中提到的 Rose 的提议）。最近，Grigoras 设计了一种新的 LTF 算法，可以实现对一种语言的元音种类进行自动识别，同时还可以对每个元音的识别结果进行分别列表。然而，对于实际案件而言，因为检材语音的质量较差，所以至少对于一部分语音材料来说，该算法的可靠性值得怀疑。

除了刚才提到的在说话人鉴定中使用的一系列声学语音学参数之外，凭借一些实践经验，大家还能想到很多能反映说话人个体特性的声学语音学参数，这些参数在鉴定实践中也会很有用。很明显，该领域还需要进行更加深入地研究和探索。尽管每项声学语音学研究都能揭示一些个体差异，但研究成果能否应用于鉴定实践，关键要看具体的个性特征是否具有足够的鲁棒性，足以抵御来自技术条件局限、时间短促、说话风格差异以及像紧张和情感差异之类的自然语音现象等因素的干扰（与本文 3.1.3 进行比较）。虽然我们常常能够在控制良好、条件完善、录音设备高级的实验室研究中得出很多能够反映不同说话人之间重要区别的声学特征，但是迄今为止，还没人知道这些特征中到底有多少能够经得起办案实践的检验（Nolan 1983 第 11 至 14 页；Rose 2002, 51f），从而成为真正的创新。在众多方法中，测量语音、语音特征和韵律单位的时长是一个不错的选择。由于发音速率或其他言语音速参量会因人而异，因此除了那些附属于言语音速以外的时长数据应该进行归一化处理。Allen 等学者（2003）曾经提出这样疑问，即已知的说话人在嗓音起始时间（塞音除阻到后接元音振动起始点之间

经历的时间）上的差别是否仅仅是说话人在产生言语音速差别时伴随的一些副效应？他们先对音速参数进行了归一化处理，发现归一化后的言语音速仍能体现不同说话人之间的个性差异。Pfitzinger（2002）也有类似的研究发现：即使在对言语音速差异进行归一化之后，说话人在不同音段时长之间依然存在个性差异。

3.2.3. 整体语音感知

正如前文所述，一般人都具有一定辨别语音的能力，因为他们没有经历过系统的语言学或语音学的分析训练，所以他们对语音的感知基本上都是整体性的。尽管语言学或语音学专家能够对语音进行解析式感知，但对语音进行整体性感知也未尝不可。毕竟，格式塔理论（Gestalt theory）的主要原则——“整体大于部分之和”——也适用于语音鉴定。而且有研究显示，在某种程度上语音学家对说话人进行整体分析的能力比一般人要好（欲查看进一步的讨论及更多参考文献，可以参考 Hollien 2002 第 37 至 39 页的内容）。

Schiller 与 Köster（1998）曾通过实验证明了语音鉴定专家在感知方面的特殊优势。实验基本设计如下：由 6 位母语为德语的德国人作为发音人；4 至 8 秒不等的语音样本 108 个；17 位说德语且没有受过语音学训练的普通人和 10 位说德语的语音鉴定专家担任被试；首先让被试熟悉发音人的语音，5 分钟后进行听音测试，要求被试逐一判断实验中播放的语音是否与测试前听到的语音相同。结果显示没有受过训练的普通被试的正确命中率（当语音一致时，回答“一致”）为 92%，误报率（当语音不一致时，回答“一致”）为 2%，相比之下，语音鉴定专家的正确命中率为 98%，而误报率仅为 1%。可见，语音鉴定专家在认定个人和降低错误率方面表现得更好。然而，在对说话人的整体感知鉴定方面，目前关于普通语音学家或语音鉴定专家为何具备超出一般人的感知优势还没有系统的研究。

对此，我们或许可以在*样例理论（Exemplar Theory）*中找到一种可能的解释（Johnson 1997）。样例理论和其他的非解析认知理论（比较 Pisoni 1997）与那些将感知过程视为一种解析活动（类似于科学家的分析活动）的理论不同。比如说，传统的语音感知分析理论将说话人识别和语音识别视为一种循序处理的过程：听话人首先对说话人的特点进行评估，然后利用得到的信息对说话人进行归一化处理，最后将处理后的信息用于语音识别。然而，在样例理论中，说话人识别和语音识别是同时进行而不是循序进行的。在解析理论中，只有说话人的纯语音参数以及一些对语音感知（如特征检测）、词汇提取和语音识别中的其他重要方面都很必要的信息才被储存记忆，但在样例理论中，整个语音事件是以包含所有语音学信息的“样例”形式进行储存的。按照 Johnson （1997）的说法，这种“样例”是一种在听觉特性——外周听觉系统的听辨结果——与范畴标签之间的关联形式。这些范畴标签可以对任何有用的通信信息加标记，包括样例的语言学实体、说话人的性别和身份等。比如说，同一说话人所有语音样本的标记都是相同的，在识别该说话人的时候，会再次用到这些被存储的信息。通过上文的简单介绍我们可以看出，样例理论不仅为整体说话人鉴定，同时也为说话人分类特性的整体感知提供了一种简单明了的解释。比如，有研究表明普通听话人在某种程度上能够对说话人的年龄（例见 Braun 1996）或方言特点（Clopper 与 Pisoni 2005）进行区分。

样例理论有很多优点，比如说，它可以解释语音识别与说话人识别之间的相互作用问题（见 Nygaard 2005），而解析理论就很难做到。另一方面，样例理论同样还面临着“脑容量存储受限的问题”（Johnson 1997）。尽管相关的心理学研究表明人类大脑的储存容量比之前想象的大很多，但肯定存在一个极限范围，在这个范围内仅保存了说话人鉴定/分类和语音识别的基本信息，而使用整体处理方式会比使用解析处理方式更早达到这个界限。比如，当要区分大量语音的时候，一个范畴类型就会有非常多的标签需要辨别区分，这个时候就可能会出现问题（源于与 Henning Reetz 的个人沟通）。同时还要注意，听话人对说话人进行整体辨别的能力（不自觉地使用或者主动利用语言学和语音学进行分析的能力）并不是对样例理论本身的一种证明。其实，这种能力背后的认知处理过程仍可能是解析性的，只是与专家的分析相比，它是在无意识状态下进行的。有关此观点还有一种说法，即认为它或许可以解释上文提到的语音学家在语音整体感知方面强于非语音专家的原因。关于“有意识和无意识”的问题，一种可能的解释是，语音专家在整体感知语音的过程中进行无意识的解析处理时，受到了来自他/她们（进行语音分析时所用到的）意识能力的

帮助，而这种能力是经过先前的训练得来的。

另一方面，研究人员在语音识别中还发现一种“词语遮蔽”效应（“verbal overshadowing”effect），意思是说在语音辨认实验范例中，当要求被试在开始识别之前先对目标语音进行描述，与控制良好不要求被试事先进行语音描述的情况相比，前者的识别率更低（Perfect 等 2002）。在人脸识别技术中同样存在这种效应，这也暗示了整体分析相对于特征处理机制的重要性。然而，就“词语遮蔽”效应对于法庭说话人鉴定的实际意义来说，要具体问题具体分析：与对专家鉴定相比，该效应很可能对由受害人和证人进行说话人鉴定的情况来说有更强的作用，因为有关该效应的部分解释并不适用于语音比对过程。例如，Perfect 等（2002）曾指出事先进行的言语描述任务会损害受害人和证人的原始记忆痕迹，但这种解释与鉴定专家对语音材料所进行的整体听觉（或其他形式的）检验并无联系，因为鉴定专家在分析过程中可以根据需要对材料中任何一段语音进行重复播放与审听，所以记忆不会成为问题。

Nolan（2005）曾提醒到，与其他鉴定方法一样，整体语音感知方法也要遵循贝叶斯方法的原则。他还特别指出在整体语音感知分析中，人们常常过于关注相似性的问题，而忽略了普遍性的问题，那是很危险的。他警告，如果注意不到很多说话人都会引发相同的完形感知（gestalt perception）的话，鉴定专家就可能会受到“相似性”的蒙骗（第 401 页），因为从广义上说，整体语音感知可以分析利用任何类型的语音特征——包括方言和由方言决定的嗓音音质特征，所以，根据 Nolan 的观点，鉴定专家可能会对一定范围内具备某些相同方言背景的人产生某些相似的感知特性。

3.2.4. 说话人自动识别技术

说话人自动识别技术（automatic speaker identification，也有人经常使用 automatic speaker recognition 的叫法）（译者注：在国外法庭语音学相关学者的著作和文章中，关于鉴定、识别或确认对应的英文术语有很多，比较有代表性的有“identification”、“recognition”和“verification”等，国内司法鉴定领域的专家多使用“identification（鉴定）”一词，而在广义或法庭说话人自动识别领域内，上述三个术语则都见有人使用，但具体含义可能稍有不同，比如“identification”与“recognition”多对应为“识别”，而“verification”则多对应为“确认”。作者在原文的写作中对于“speaker identification by victims and witnesses”、“speaker identification by experts”和“automatic speaker identification”等术语中均采用“speaker identification”一词的原因是，作者认为在法庭语音学这个领域内最好采用一个统一的较为广义的术语“speaker identification”，然后再根据“speaker identification”的类型、任务和方法等因素的不同，对其进行进一步的划分。比如在本文的第 2 部分，作者就是根据“speaker identification”类型/任务的不同将其分为“语音比对”、“语音画像”和“受害人和证人进行的说话人鉴定”，本文的第 3 部分则介绍了语音比对的 4 种常见方法。尽管在语音比对与受害人和证人进行的说话人鉴定中“identification”的含义并非完全一样（两者的方法不同），但二者的终极目标却是一样的，即确定说话人具体是谁，在这个意义上二者是一致的。因此，作者认为最好使用一个术语“speaker identification”，不要同时使用“identification”、“recognition”和“verification”三个术语，因为那样很可能会引起一些不必要的歧义，然而那些没有争议、完全定型化的术语则可以使用，但在这里显然不适用（上述部分内容源于译者与原作者的个人交流）。鉴于国内对于“自动识别”的概念使用非常广泛，译文中并未将“automatic speaker identification”译为“说话人自动鉴定”。）在法庭说话人鉴定方面的应用只有不到 10 年的时间，但在法庭科学领域以外，它（有人称之为广义说话人自动识别，general automatic speaker identification）作为一门学科已经有很长时间了，至少可以追溯到 20 世纪 60 年代。该学科的特点是语音与技术相结合，而且目前已经发展应用到多个领域，像对敏感信息、高安全性建筑物或设备等场所和物品的访问控制，另外在生物测量应用方面也有一定的作用。与法庭说话人自动识别相比，广义说话人自动识别在很多方面都要简单一些，比如说，其接入被检测者的语音和样本库中已有的声音，无论在质量上还是在数量上都可以做到最优化处理，同时也不存在像本文 3.1.3 中提到的一系列的错配问题。另外这两种自动识别技术在概念上也不尽相同。比如说像在广义说话人自动识别中经常出现的闭集识别、可接受的漏报率或误报率条件下的阈值设置等概念，除了一些特殊的专业应用之外，在法庭科学领域就不常见。另一方面，在法庭科学

领域中最基本的贝叶斯方法（见本文 3.1.2）在广义说话人自动识别领域并不是绝对必需的。

近几年来，广义说话人自动识别的效率不断提高，计算机技术发展迅速，贝叶斯方法在法庭科学领域中的应用也有很大进步，很可能也正是得益于这三个方面的进步，才最终能够将说话人自动识别技术应用到法庭语音与音频分析领域中来。

法庭说话人自动识别系统主要包括三个部分：参数提取、参数建模和距离计算，相当于说话人自动识别处理过程中三个连续的阶段。在第一阶段的参数提取过程中，系统会从语音信号中自动提取声学参数。至于哪些声学参数对语音识别来说最有效，目前有许多不同的观点。其中，最常用的参数是美尔倒谱系数（MFCC），该特征向量的产生有以下几步：首先要通过离散傅里叶变换得到语音的功率谱（power spectrum），这是声学语音学中常见的分析方法；其次，再让频谱经过一组基于非线性美尔刻度的滤波器组的调制，对频谱形状进行平滑滤波，从而使输出结果更加符合心理声学的实际；最后，使用离散余弦变换法将滤波系数的对数转换成倒谱形式，最终得到的向量参数就叫做谱系数（详见 Bimbot 等 2004 及更多的文章）。

我们采用 20ms 的窗长，提取 20 个 MFCC 参数，并且窗移 10ms（即有 50%的窗口重合），这样就能得到说话人在整个录音过程中 MFCC 的分布情况。需要注意的是，该方法并没有将语音流分割划分到不同的语言学范畴中，比如辅音、元音或者音节等。这也正是笔者在表 1 中将说话人自动识别方法划分为整体分析方法的原因之一。

值得一提的是，MFCC 参数的一个重要特点是它能够较好地实现声源特性与声道特性的分离。在语音学中，声源-滤波模型涉及喉部的嗓音声源与对声源能量起到滤波作用的喉上结构的贡献之间的区别。研究证明，如果忽略那些声源信息—特别是基频特征—的 MFCC 维度分量，只用那些能够获取声道特性的 MFCC 维度分量的话，会提高自动识别的识别率。因为共振峰是反映声道特性的重要参数，这种选择似乎意味着对于说话人识别来说共振峰比基频更重要。然而，大家应该看到 MFCC 参数的提取不同于共振峰的自动追踪（正如本文 2.2.2 中提到的）。针对共振峰追踪的算法主要是为了寻找确定共振峰，因为无论是关于从语音产生轮廓和结构上看共振峰的产生机制，还是关于共振峰对语音感知的影响，语音学中对共振峰的研究都已经比较成熟了，对于在处理低质量的语音信号中经常出现的共振峰自动追踪失败的情况，语音学家都能够进行确认（此时，语音学家可以使用手工纠错）。相比之下，提取 MFCC 参数永远不会出错，因为对于声学信号中预先确定的结构来说无所谓追踪的对错之分。然而，由分析 MFCC 得到的参数值却不像共振峰那样容易得到解释。MFCC 参数与语音产生和语音感知之间到底有怎样的具体联系，人们对此还知之甚少。或许语音学家应该关注一下这个问题，但迄今为止还鲜有尝试，只有 Mokhtari 等（2007）做过类似的研究（可以比较 Rose 2002 对共振峰与倒谱分析之间的区别进行的类似讨论）。倒谱模式缺乏语音学/语言学上的可解读性，这也正是笔者在表 1 中将说话人自动识别划分为整体分析法的第二个原因[12]。

由于 MFCC 提取模式无法在语音学上得到解释，因此在 MFCC 参数包含的所有信息中，哪些是有关说话人的，哪些是有关信道、背景噪音以及其

[12]相比较而言，说话人自动识别产生了一种新的方法，即除了利用像 MFCC 等“声学特征”以外，还利用一些“高层特征（higher-level features）”来进行识别。麻省理工学院的林肯实验室在其 Super-SID 项目中率先使用了这种方法（相见网站 http://www.clsp.jhu.edu/ws2002/groups/supersid/SuperSID_Closing_Talk_files/frame.htm）。最近，Shriberg（2007）对此“高层特征”进行了研究综述。所谓“高层特征”就是那些可以从语音学和语言学角度进行解释的特征，比如说音段的语音配列、音段时长和语调特点等。说话人自动识别技术可以将语音流分割成音段、音节和其他的语言学单位，很多“高层特征”的获取都得益于此。由于可以进行语言学分割，同时使用可以从语音学得到解释的特征，该方法应该算是一种解析方法而不是整体分析的方法。然而，与使用 MFCC 参数的方法不同，这种使用高层特征的方法仍然是完全自动的，因此，它会在语言学/语音学分类上犯一些客观错误。特别是在处理实际案件语料并要出具鉴定意见的时候，完全依靠自动识别系统的分析能力是十分危险的。为此，笔者建议在使用自动识别系统时，至少应该有一名专家要对该系统的分析能力进行一些系统监管。当然并不是说该高层特征识别方法的研发人员常常推荐应该在法庭说话人鉴定领域中使用他们的技术。就目前而言，该方法更适合在广义说话人自动识别中使用，而不是在法庭说话人自动识别领域。

他无关的干扰声学影响的都不是很清楚。对说话人自动识别技术来说，将语音信息与其他无关信息分离开来的确是一件十分重要又十分困难的事情，特别是与一些语料质量较好的情况相比，该技术在鉴定领域的应用更具挑战性。实现信息的分离经常需要若干步骤，其中一个步骤是在提取 MFCC 参数之前要进行语音增强；另一个步骤是倒谱均值相减（cepstral mean subtraction），即在假设整个录音中的信道信息和噪声问题保持一致的前提下，每个局部 MFCC 都是减去整个录音文件的平均 MFCC 得到的。

既然 MFCC 参数是对语音信息进行整体性采集/捕获，那么它们也应该同样包括超语言方面以及语体/风格方面的信息，比如朗读与自然说话，或者是否带有特殊感情等。如果未知说话人、嫌疑人和参考人群（reference population）中的所有说话人都使用同样的超语言特征或风格背景，那就没有什么问题，但实际情况是涉及这些行为模式的时候经常会出现错配的问题。为此，研究人员在说话人自动识别中引入了“错配条件补偿”技术（Botti 等 2004），它会在第三个步骤（相似性计算）中进行，下文将做进一步解释。该技术同样可以在技术错配或多次循环组合的技术和行为错配中运用。

另外，语言不同也是一种可能出现的错配类型。在法庭说话人自动识别时，检验人员使用的参考人群中的说话人与未知说话人和嫌疑人所说的语言可能是不同的（或使用一个包含多种语言的参考人群），这种情况并不罕见。原因可能是，之前有经验显示，语言不同对MFCC模式的影响非常小。一直以来，与其他方法相比，与语言无关的特点经常被视为法庭说话人自动识别技术的主要优点。然而，最近的研究结果显示，在说话人自动识别中语言不同*确实*是有影响的，特别是对双语者的语音来说，当其训练语音与测试语音的语言不同时，与语言相同时相比，识别率会更低（Przybocki等2007）。

现在我们来看说话人自动识别的第二阶段——参数建模，本阶段的目的是为原始 MFCC 的分布情况进行建模。建模的方法有很多，但最常用的要数高斯混合模型（GMM）了，见图 3 所示。

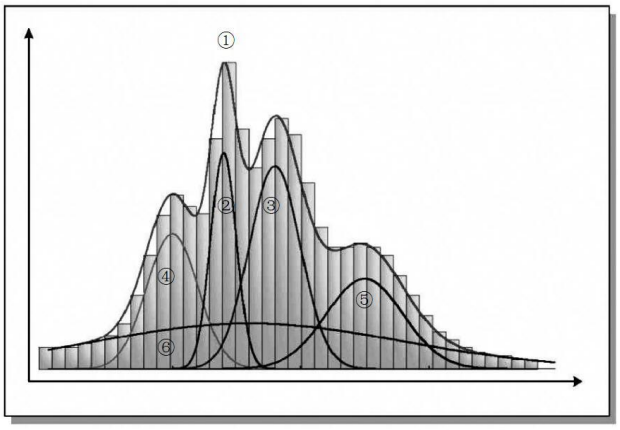


图3 高斯混合模型的抽象表示

图3显示了本例中一组高斯函数（见图中5条高斯曲线②至⑥）对MFCC分布进行建模的情况。图中包络线①表示GMM的模拟结果。理论上说，X轴表示不同的MFCC特征向量值，Y轴表示一段语音材料在给定区间内发现特征向量值的次数。图中演示的只是一维的情况，实际上需要计算大约20个特征向量数据的多维高斯函数。高斯函数也被称为“分量（component）”，这些分量的数量作为变量之一，可供用户选择使用，并且可以在说话人自动识别程序的优化处理中进行多种选择。GMM模型的最终结果是一种说话人模型（即一个特定说话人语音模式的模型）。

第三个也是最后一个阶段是距离计算。这里我们将讨论一下*相似度计算*的问题，它是通过反求透视图得到的，也比较好理解。过程见图4所示。

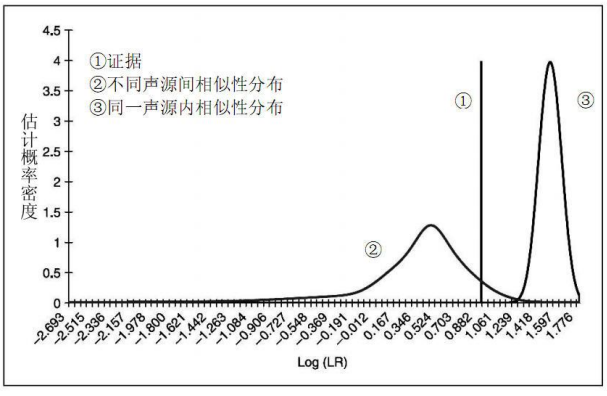


图4 法庭说话人自动识别中基于相似度的似然比计算：图中给出的是没有比对成功的例子。由此方法，相似度本身（X轴）就是以似然比的形式呈现的，但是不能将此数值与文中讲到的通过对比同一说话人内和不同说话人间的相似度分布得到的似然比相混淆。

图 4 中显示的两种分布曲线都是二阶分布（此二阶分布的建模方法有很多，其中一种方法是通过核密度函数（kernel density function）实现的），不要和 GMM 弄混了。右侧的③号分布曲线代表在一次语音比对过程中嫌疑人（说话人）的二阶模型，该模型是通过计算嫌疑说话人的 GMM 与该说话人的多个不同语音样本间的多种相似性得到的。在理想情况下，该说话人应该在不同状态下说话。我们可以从 X 轴上的具体数值读出语音样本与 GMM 的相似程度。既然③号分布曲线所含的语音样本都是同一说话人说的，那么可以确定的是这些语音样本的相似度是很高的，换句话说，③号分布曲线处在 X 轴上相似范围更高端的位置上。左侧②号分布曲线是通过计算未知说话人的语音样本与所谓的参考人群中每一个说话人的 GMM 之间的相似性得到的。这里说的参考人群指的是一组说话人，通常包括 30 个人或更多。对于通过贝叶斯方法确定除了嫌疑人之外是否还有一些说话人可能被识别为未知说话人来说，维持这个数量的参考人群是必要的。既然我们可以放心假设未知说话人与参考人群中的任何说话人都不一致，那么可以确定的是实际的相似度要比比对同一说话人的情况更低，因此在相似度范围上也就更靠近更低的位置。总的说来，③号分布曲线描述的是同一说活人内的相似性，即当嫌疑人与其自身进行比对时所得到的相似值，而②号分布曲线描述的是不同说话人间的相似性。最后在说话人自动识别程序的这个阶段还有一个重要成分，叫做“证据”（E），如图中的①号竖线所示。这里的“证据”是当匿名说话人的语音样本与嫌疑人的 GMM 进行比对时得出的相似值，具体是根据本文 3.1.2 中的 LR 公式计算得出的。如果假设此“证据”是同一说话人内分布曲线的一部分，那么 LR 的分子就表示获得此“证据”相似度的似然性（参见图中的竖线在 X 轴上的具体位置）。在图 4 中，①号竖线与③号分布曲线相交点处的似然值极低（见此点对应的 Y 轴数值）。如果假设此“证据”是不同说话人间分布曲线的一部分，那 LR 的分母就表示获得“证据”相似度的似然性，即①号竖线与②号分布曲线的相交点，很明显此点的似然值要高于同一说话人内分布时的似然值。分子除以分母得到的 LR 值小于 1。正如图 4 中所示，LR 小于 1 证明嫌疑人与未知说话人不是同一人。如果是认定同一的情况，①号竖线的位置就会向右移动，其在同一说话人内分布曲线中的位置会多于在不同说话人间分布曲线中的位置。

如果大家想了解更多有关说话人自动识别程序，特别是其第三阶段相关内容的话，可以参考 Drygajlo（2007）的相关文章。Gonzalez-Rodriguez 等（2006）与 Müller （2007）也曾对该领域的最新发展情况进行过研究综述，Bijhold 等（2007）文章则提供了更多的参考文献。该领域发展迅速，大家若要了解最新的技术发展状况，可以关注每两年召开一次的“Speaker Odyssey”会议的会议成果。虽然该会议主要立足于广义说话人自动识别技术，但每次开会都设有说话人自动识别技术法庭应用的常规会议。

4. 结论

本文综述回顾了目前法庭语音学领域中一系列具有概括性的主题、任务和科学挑战，以期待能够让读者对法庭语音学产生有一个整体印象。正如笔者在引言中提到的，尽管法庭语音学还包含许多其他领域的内容，但本文只讨论了该学科的中心科学任务——说话人鉴定方面的内容。而在说话人鉴定领域中，大部分现行的实际进展和理论问题都涉及语音比对的内容。语音比对的任务很简单，即对于给定的两段不同录音，判断它们是同一人还是不同人说的。问题虽不难，但要在鉴定实践中做出正确结论却需要丰富的办案经验和科学的知识储备作为支撑。

随着法庭科学领域中贝叶斯方法的发展，很明显，语音比对已经不仅仅是对不同录音材料中说话人语音模式的异同点进行比对的问题，同时还要弄清楚这些语音模式在相关的参考人群中的常见程度。如果其相似性和普遍性两方面都可以进行定量表示的话，那么似然比是可以计算出来的。但即使目前对这两方面的了解还更多的停留在定性化（如使用听觉语音学法）或绝对化（如使用整体语音感知法）的程度上，似然比的概念以及贝叶斯方法在法庭说话人鉴定领域仍然应该作为一项指导原则来对待（Rose 2002）。

这些年来还有一点是比较明确的，即说话人鉴定不应仅仅使用一种单一的方法，而应该使用多种检验方法。本文就介绍了四种检验方法（见表 1）。使用多种方法进行检验，可以使做出的同一（identity）或非同一（non-identity）的结论更加可靠，同时也能够降低错误认定或错误拒绝的风险。

之所以这样做，主要是因为每种方法只能对说话人全部个性特征范围内的某些不同方面的信息进行检验。比如说，一方面与解析法相比，说话人自动识别方法对语音信号的检测范围更加全面，几乎不会遗漏任何潜在的说话人个性成分；然而另一方面，使用解析法时，鉴定专家可以依靠其必备的科学知识将那些对于说话人鉴定过程来说至关重要的信息识别出来，但若通过自动方法采集的话，（计算机）则有可能将其遗失在海量的信息当中。比如说，目前的自动分析系统还常常识别不出说话人的方言信息类型，而这很有可能就会导致得出不能认定两个说话人同一的结论。

法庭语音学仍然是一门比较年轻的学科。尽管很多法庭科学实验室以及许多私人从业人员已经积累了丰富的经验，也做了很多重要研究，但该学科中还有很多领域仍然值得做进一步的研究。其中一个重要的研究领域就是声学语音学。众所周知，共振峰结构携带了说话人的重要个性信息，尽管如此，我们却仍然不知道如何最大限度地利用共振峰特征。对此，已经有学者开展了新的研究，如讨论将不同的共振峰和元音信息结合起来，或者获取共振峰的动态信息来进行鉴定。另外一个重要领域就是去进一步探索人类（或许也包括非人类）辨识语音的感知能力。对此，样例理论可以提供一种相关的理论解释。另外还有一个问题也值得做更深入的探索，即不同的检验方法在何种程度上获取类似的说话人信息（由此也包含了一定的冗余信息，这些信息还可以增加结论的可信度），或者是否获取的是一些在概率意义下可以被视为是没有关联性的不同信息（这些信息有助于提高似然比，得到强有力的结论）。然而当不同检验方法之间产生冲突时怎么办呢？实践中确实存在这种情况，比如使用听觉分析方法要认定同一时，通过声学语音学方法得到的证据却指向相反的结论（参见 Nolan1990），或者当使用说话人自动识别方法不能认定时，所有其他方法却都支持认定。还有一个重要方面是共振峰分析与说话人自动识别之间的关系问题，我们知道这两种方法都集中研究声道的特点，但二者在认定或否定说话人时是否会得出相似或不同的结论，得出这些相同点或不同点的原因何在（见 Rose 等 2003 对此方向的一项重要研究）？这都需要做进一步的研究。

最后还要强调一点，尽管看起来针对说话人特点的研究好像应该是鉴定实践中真正的任务，但实际并非如此，其真正的任务是一些下位问题，像移动电话技术如何影响共振峰、如何处理语音伪装或如何对一些错配情况进行补偿等问题。说话人特点（speaker characteristics/talker characteristics）同时也是普通语音学的研究对象之一。比如说，在许多语音学研究中，说话人都是很重要的影响因素之一，并且由此得出的众多研究成果对于法庭语音学来说也是很有意义的。法庭语音学家不能将自己完全局限在实践办案当中，而应该时刻关注普通语音学关于说话人特点的最新研究成果，为我所用。同样道理，普通语音学家也不应该认为说话人特点就仅仅事关鉴定实践，与自己无关，相反应该看到这对他们自己的研究领域来说也是十分重要的。普通语音学家同样可以对法庭语音学领域做出重要的贡献，比如说针对多个说话人展开实验研究，探索自然语音条件，或简单展示一些可以从某种程度上看到说话人变化而不是对说话人进行集中交叉研究的实验结果等等。

致谢

感谢 Paul Foulkes 和另一位审核人对本文提出了许多宝贵意见，特别要感谢他们向我推荐 Perfect 等（2002）的文章。同时还要感谢 Franz Broß 与 Timo Becker 对本文中有关说话人自动识别部分提出的意见，感谢 Franz Broß 允许我使用文中列出的第 3 和第 4 张图示。

参考文献

[1] Alexander, A., Dessimoz, D., Botti, F. et al. Aural and automatic forensic speaker recognition in mismatched conditions [J]. International Journal of Speech, Language and the Law, 2005, 12: 214–34.

[2] Sean, A.J., Miller, J.L. and DeSteno, D. Individual talker differences in voice-onset-time [J]. Journal of the Acoustical Society of America, 2003, 113: 544–52.

[3] Baldwin, J. and French, P. Forensic phonetics [M]. London, UK: Pinter, 1990.

[4] Bijhold, J., Ruifrok, A., Jessen, M., et al. Forensic audio and visual evidence 2004–2007: A review. 15th INTERPOL Forensic Science Symposium,

October 2007, Lyon, France, http://www.forensic.to/webhome/enfsidiwg/Interpol-review-2007-audio-visual-evidence-paper.pdf

[5] Bilton, M. Wicked beyond belief: The hunt for the Yorkshire ripper London [M], UK: Harper Collins, 2003.

[6] Bimbot, F., Bonastre, J.-F., Fredouille, C. et al. A tutorial on text-independent speaker verification [J]. EURASIP Journal on Applied Signal Processing, 2004, 4: 430–51.

[7] Blatchford, H. and Foulkes, P. Identification of voices in shouting [J]. International Journal of Speech, Language and the Law, 2006, 13: 241–54.

[8] Botti, F., Alexander, A. and Drygajlo, A. On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition [J]. Forensic Science International, 2004, 146S: 101–6.

[9] Braun, A. Age estimation by different listener groups [J]. Forensic Linguistics, 1996, 3: 65–73.

[10] Braun, A. Fundamental frequency – How speaker-specific is it? In: Braun, A. and Köster, J.-P., eds. Studies in forensic phonetics [M]. Trier, Germany: Wissenschaftlicher Verlag Trier, 1995: 9–23.

[11] Broeders, A.P.A. Forensic speech and audio analysis, forensic linguistics 1998 to 2001: A review [A]. In: 13th INTERPOL Forensic Science Symposium [C], 16–19 October 2001, Lyon, France, http://www.interpol.int/Public/Forensic/IFSS/meeting13/Reviews/ForensicLinguistics.pdf

[12] Broeders, A.P.A. Forensic speech and audio analysis, forensic linguistics. A review: 2001 to 2004 [A]. In: 14th INTERPOL Forensic Science Symposium [C], 19–22 October 2004, Lyon, France, 171–88, http://www.interpol.int/Public/Forensic/IFSS/meeting14/ReviewPapers.pdf

[13] Broeders, A.P.A. Phonetics, Forensic. In: By Brown, K., eds. Encyclopedia of language and linguistics [M]. 2nd ed. Amsterdam, the Netherlands: Elsevier, 2006: 460–3.

[14] Broeders, A.P.A. and van Amelsvoort, A.G. Lineup construction for forensic earwitness identification: a practical approach [A]. In: Proceedings of the 14th International Congress of Phonetic Sciences [C], San Francisco, CA, 1999, vol. 2: 1373–6.

[15] Byrne, C. and Foulkes, P. The 'mobile phone effect' on vowel formants [J]. International Journal of Speech, Language and the Law, 2004, 11.83–102.

[16] Clark, H.H. and Fox Tree J.E. Using uh and um in spontaneous speaking [J]. Cognition, 2002, 84: 73–111.

[17] Clopper, C.G. and Pisoni D.B. Perception of dialect variation [A]. In: Pisoni D.B. and Remez, R.E., eds. The Handbook of speech perception [M]. Oxford, UK:Blackwell, 2005: 313–337.

[18] Darby, J.K. (ed.) Speech evaluation in psychiatry [M]. New York, NY: Grune and Statton, 1981.

[19] Douglas, J. and Olshaker, M. Mind hunter: Inside the FBI's elite serial crime unit [M]. New York, NY: Simon and Schuster, 1995.

[20] Drygajlo, A. Forensic automatic speaker recognition [J]. IEEE Signal Processing Magazine, 2007, March: 132–5.

[21] Ellis, S. The Yorkshire ripper enquiry: Part I [J]. Forensic Linguistics, 1994, 1: 197–206.

[22] Eriksson, A. and Lacerda, F. Charlatanry in forensic speech science: a problem to be taken seriously [J]. International Journal of Speech, Language and the Law, 2007, 14: 169–93.

[23] Fisher, J. The ghosts of Hopewell: Setting the record straight in the Lindbergh case [M]. Carbondale, IL: Southern Illinois University Press, 1999.

[24] Fisher, J. The Lindbergh case. Paperback revision of the first 1987 edition, New Brunswick, NJ: Rutgers University Press, 2000.

[25] Fitch, W.T. and Giedd, J. Morphology and development of the human vocal tract: a study using magnetic resonance imaging [J]. Journal of the Acoustical Society of America, 1999, 106: 1511–22.

[26] Foulkes, P. and Barron, A. Telephone speaker

recognition amongst members of a close social network [J]. Forensic Linguistics, 2000, 7: 180–98.

[27] Foulkes, P. and Docherty, G. The social life of phonetics and phonology [J]. Journal of Phonetics, 2006, 34: 409–38.

[28] French, P. Analytic procedures for the determination of disputed utterances [A]. In: Kniffka, H., eds. Texte zur Theorie und Praxis forensischer Linguistik [M]. Tübingen, Germany: Niemeyer, 1990: 201–13.

[29] French, P. An overview of forensic phonetics with particular reference to speaker identification [J]. Forensic Linguistics, 1994, 1: 169–81.

[30] French, P. and Harrison, P. Investigative and evidential applications of forensic speech science [A]. In: Heaton-Armstrong, A., Shepherd, E., Gudjonsson, G., et al. eds. Witness testimony, psychological, investigative and evidential perspectives [M]. Oxford, UK: Oxford University Press, 2006: 247–62.

[31] French, P. and Harrison, P. (eds.) Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases [J]. International Journal of Speech, Language and the Law, 2007, 14: 137–44.

[32] French, P., Harrison, P. and Windsor Lewis, J. R vs John Samuel Humble: the Yorkshire Ripper hoaxer trial [J]. International Journal of Speech, Language and the Law, 2006, 13: 255–73.

[33] Gfroerer, S. Auditory-instrumental forensic speaker recognition [A]. In: Proceedings of EUROSPEECH 2003 [C], Geneva, Switzerland, 2003: 705–8.

[34] Goldman-Eisler, F. Psycholinguistics: Experiments in spontaneous speech [M]. London, UK: Academic Press, 1968.

[35] Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D. et al. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition [J]. Computer Speech and Language, 2006, 20: 331–55.

[36] Hammersley, R. and Read, J.D. Voice identification by humans and computers [A]. In: Sporer, S.L., Malpass, R.S. and Koehnken G. eds. Psychological issues in eyewitness identification [M]. Mahwah, NJ: Lawrence Erlbaum Associates, 1996: 117–52.

[37] Hirano, M. Clinical examination of voice [M]. Berlin, Germany: Springer, 1981.

[38] Hollien, H. The Acoustics of crime. The new science of forensic phonetics [M]. New York, NY: Plenum, 1990.

[39] Hollien, H. Forensic voice identification [M]. San Diego, CA: Academic Press, 2002.

[40] Hudson, T., de Jong, G., McDougall, K. et al. F0 statistics for 100 young male speakers of Standard Southern British English [A]. In: Proceedings of the 16th International Congress of Phonetic Sciences [C], Saarbrücken, Germany, 6–10 August 2007: 1809–1812. http://www.icphs2007.de/conference/Papers/1570/1570.pdf

[41] International Phonetic Association. Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet [M]. Cambridge, UK: Cambridge University Press, 1999.

[42] Jessen, M. Forensic reference data on articulation rate in German [J]. Science and Justice, 2007a.47: 50–67.

[43] Jessen, M. Speaker classification in forensic phonetics and acoustics [A]. In: Müller,C. eds. Speaker classification I: Fundamentals, features, and methods [M]. Berlin, Germany: Springer, 2007b: 180–204.

[44] Jessen, M., Köster, O. and Gfroerer, S. Influence of vocal effort on average and variability of fundamental frequency [J]. International Journal of Speech, Language and the Law, 2005, 12: 174–213.

[45] Johnson, K. Speech perception without speaker normalization: an exemplar model [A]. In: Johnson, K. and Mullennix, J.W. eds. Talker variability in speech processing [M]. San Diego, CA: Academic Press, 1997: 145–165.

[46] Kaplan, E. and Kaplan, M. Chances are . . . : Adventures in probability [M]. New York, NY:

Viking Books, 2006.
[47] Keating, P.A. Phonetic representations in a generative grammar [J]. Journal of Phonetics, 1990, 18: 321–34.
[48] Köster, O. and Köster, J-P. The auditory-perceptual evaluation of voice quality in forensic speaker recognition [J]. The Phonetician, 2004, 89: 9–37.
[49] Köster, O., Jessen, M., Khairi, F. et al. Auditory-perceptual identification of voice quality by expert and non-expert listeners [A].In: Proceedings of the 16th International Congress of Phonetic Sciences [C], Saarbrücken, Gertmany, 6–10 August 2007: 1845–1848. http://www.icphs2007.de/conference/Papers/1152/1152.pdf
[50] Kreiman, J., Gerratt, B.R., Precoda, K. et al. Individual differences in voice quality perception [J]. Journal of Speech and Hearing Research, 1992, 35: 512–20.
[51] Künzel, H.J. Sprechererkennung: Grundzüge forensischer Sprachverarbeitung [M]. Heidelberg, Germany: Kriminalistik Verlag, 1987.
[52] Künzel, H.J. Field procedures in forensic speaker recognition [A]. In: Lewis, J.W. eds. Studies in General and English Phonetics. Essays in Honour of Professor J. D. O'Connor [C]. London, UK: Routledge, 1995, 68–84.
[53] Künzel, H.J. Some general phonetic and forensic aspects of speaking tempo [J]. Forensic Linguistics, 1997, 4: 48–83.
[54] Künzel, H.J. Effects of voice disguise on speaking fundamental frequency [J]. Forensic Linguistics, 2000, 7: 149–79.
[55] Künzel, H.J. Tasks in forensic speech and audio analysis: a tutorial [J]. The Phonetician, 2004, 90: 9–22.
[56] Ladefoged, P. Preliminaries to linguistic phonetics [M]. Chicago, IL: University of Chicago Press, 1971.
[57] Ladefoged, P. and Maddieson, I. The sound of the world's languages [M]. Oxford, UK: Blackwell, 1996.
[58] Laver, J. The phonetic description of voice quality [M]. Cambridge, UK: Cambridge University Press, 1980.
[59] Laver, J. Principles of phonetics [M]. Cambridge, UK: Cambridge University Press, 1994.
[60] McDougall, K. Speaker-specific formant dynamics: an experiment on Australian English /ai/ [J]. International Journal of Speech, Language and the Law, 2004, 11: 103–30.
[61] McDougall, K. Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies [J]. International Journal of Speech, Language and the Law, 2006, 13: 89–126.
[62] McDougall, K. and Nolan, F. Discrimination of speakers using the formant dynamics of /u:/ in British English [A]. In: Proceedings of the 16th International Congress of Phonetic Sciences [C], Saarbrücken, Germany, 6–10 August 2007: 1825–1828. http://www.icphs2007.de/conference/Papers/1567/1567.pdf
[63] McGehee, F. The reliability of the identification of the human voice [J]. Journal of General Psychology, 1937, 17: 249–71.
[64] Meuwly, D. Voice analysis [A].In: Siegel, J. A., Saukko, P. J. and Knupfer, G.C. eds. Encyclopedia of forensic sciences [M]. San Diego, CA: Academic Press, 2000: 1413–21.
[65] Mokhtari, P., Kitamura, T., Takemoto, H. et al. Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients [J]. Journal of Phonetics, 2007, 35: 20–39.
[66] Müller, C. (ed.). Speaker classification, vol. I and II [M]. Berlin, Germany: Springer, 2007.
[67] Nawka, T. and Anders, L.C. Die auditive Bewertung heiserer Stimmen nach dem RBH-System [M]. Stuttgart, Germany: Thieme (2 CDs), 1996.
[68] Nolan, F. The phonetic bases of speaker recognition [M]. Cambridge, UK: Cambridge University Press, 1983.
[69] Nolan, F. The limitations of auditory-phonetic speaker identification [A]. In: Kniffka, H. eds.

Texte zur Theorie und Praxis forensischer Linguistik [M]. Tübingen, Germany: Niemeyer, 1990: 457–79.

[70] Nolan, F. Auditory and acoustic analysis in speaker recognition [A]. In: Gibbon, J. eds. Language and the law [C]. London, UK: Longman, 1994, 326–45.

[71] Nolan, F. Speaker recognition and forensic phonetics [A]. In: Hardcastle, W.J. and Laver, J. eds. The handbook of phonetic sciences [M]. Oxford, UK: Blackwell, 1997: 744–67.

[72] Nolan, F. Speaker identification evidence: its forms, limitations and roles. Proceedings of the conference 'Law and Language: Prospect and Retrospect', 12–15 December 2001, Levi (Finnish Lapland). http://www.ling.cam.ac.uk/francis/LawLang.doc

[73] Nolan, F. A recent voice parade [J]. International Journal of Speech, Language and the Law, 2003, 10: 277–91.

[74] Nolan, F. Forensic speaker identification and the phonetic description of voice quality [A]. In: Hardcastle W.J. and Beck, J. M. eds. A figure of speech. A festschrift for John Laver [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2005, 385–411.

[75] Nolan, F. and Grigoras, C. A case for formant analysis in forensic speaker identification [J]. International Journal of Speech, Language and the Law, 2005, 12: 143–73.

[76] Nygaard, L.C. Perceptual integration of linguistic and nonlinguistic properties of speech [A]. In: Pisoni, D.B. and Remez, R.E. eds. The handbook of speech perception [M]. Oxford, UK: Blackwell, 2005: 390–413.

[77] Ormerod, D. Sounds familiar? – Voice identification evidence [J]. The Criminal Law Review, 2001: 595–622.

[78] Perfect, T.J., Hunt, L.J. and Harris, C.M. Verbal overshadowing in voice recognition [J]. Applied Cognitive Psychology, 2002, 16: 973–80.

[79] Pfitzinger, H. Intrinsic phone durations are speaker-specific [A]. In: Proceedings of the International Conference on Spoken Language Processing[C], 2002, 2: 1113–6. http://www.phonetik.uni-muenchen.de/~hpt/

[80] Pisoni, D. Some thoughts on 'normalization' in speech perception [A]. In: Johnson, K. and Mullennix, J.W. eds. Talker variability in speech processing [M]. San Diego, CA: Academic Press, 1997: 9–32.

[81] Przybocki, M.A., Martin, A.F. and Le, A.N. NIST speaker recognition evaluations utilizing the Mixer corpora – 2004, 2005, 2006. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15: 1951–59.

[82] Robertson, B. and Vignaux, G.A. Interpreting evidence [M]. Chichester, UK: Wiley, 1995.

[83] Rose, P. Forensic speaker recognition at the beginning of the twenty-first century – An overview and a demonstration [J]. Australian Journal of Forensic Sciences, 2005, 37: 49–72.

[84] Rose, P. Technical forensic speaker recognition: evaluation, types and testing of evidence [J]. Computer Speech and Language, 2006a, 20: 159–91.

[85] Rose, P. Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination [A]. In: Proceedings of the Speaker and Language Recognition Workshop [C]. Puerto Rico: IEEE Odyssey, 2006b.

[86] Rose, P. and Duncan, S. Naïve auditory identification and discrimination of similar voices by familiar listeners [J]. Forensic Linguistics, 1995, 2: 1–17.

[87] Rose, P., Osanai, T. and Kinoshita, Y. Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold [J]. International Journal of Speech, Language and the Law, 2003, 10: 179–202.

[88] Rose, P. Forensic speaker identification [M]. London, UK: Taylor and Francis, 2002.

[89] Schiller, N.O. and Köster, O. The ability of expert witnesses to identify voices: a comparison between trained and untrained listeners [J]. Forensic Linguistics. 1998, 5: 1– 9.

[90] Shriberg, E. Higher-level features in speaker recognition [A]. In: Müller, C. eds. Speaker classification I: Fundamentals, features, and methods. Berlin, Germany: Springer, 2007: 241–59.
[91] Shuy, R.W. Language in the American courtroom [J]. Language and Linguistics Compass, 2007,1: 100–14.
[92] Sporer, S.L., Malpass, R.S. and Koehnken, G. (eds.). Psychological issues in eyewitness identification [M]. Mahwah, NJ: Lawrence Erlbaum Associates, 1996.
[93] Stevens, K.N. Acoustic phonetics [M]. Cambridge, MA: MIT Press, 1998.

---

iMichael Jessen，德国联邦刑事侦查局。曹洪林，证据科学教育部重点实验室（中国政法大学）；北京大学中文系。王英利，广东省公安厅刑事科学技术中心。原文刊载在Language and Linguistics Compass，2008 (2/4): 671–711。本文的翻译已经得到作者Michael Jessen及Wiley-Blackwell出版社的授权。感谢北京大学中文系汪高武博士和中国科学院自动化研究所朱磊博士对本文部分内容提出的宝贵意见。此文已发表在《证据科学》2010年12月 第18卷 第6期 第712-738页

韩国学习者的汉语嗓音音质加工方式

Voice Quality Processing Strategy of Korean Speakers in Chinese

吴韩娜

Oh Hanna

摘要 ：本文从嗓音音质的发声音质和调音音质的角度，以男女韩国学习者发的汉语和韩语两种语言的元音为研究对象，分别采集语音信号和喉头仪信号，并计算基频、开商、速度商，谐波振幅差值及共振峰参数，分析了两种语言的不同嗓音音质特征，从而初步了解了韩国学习者以什么样的定势来体现汉语的嗓音音质。研究发现，韩国学习者确实由与母语不同的方式来体现汉语的嗓音音质特征，尤其是发声音质的变化比较明显。值得注意的是其表现趋势和韩语的松/紧对立非常相似，据此，本文认为调音音质特征，加工发声音质的时候，也会将母语已有的特征来对应于目标语。

Abstract ：This paper investigates whether the Korean native speakers in advanced level of Chinese product Chinese with specific voice quality, and which voice quality parameters have significant difference. The study includes comparative analysis between Korean and Chinese vowel /a/ for Korean speakers, it can make certain by measuring F0, OQ, SQ and H1-H2 parameters for phonation quality and formants parameters for articulation quality by means of EGG.

The results showed that the Korean learners use specific voice quality features for improving Chinese nativeness, especially the phonation quality changes are obvious. In addition, this study found that there are similarities between phonation quality features of Korean/Chinese contrast and Korean lax/tense contrast features. It suggests, in the target language processing, phonetic features of mother tongue (L1) contribute “phonation quality” perception and production of the second language (L2) as well.

关键词 ：第二语言语音，汉语，嗓音音质，松/紧音，L2 策略

Key word ：second language, voice quality, Chinese, phonation, lax/tense

0 引言

第二语言学习者说目标语时，是否考虑目标语的发声音质特征？

众所周知，学习第二语言的最大目的无非是“交际”，而学习第二语言的学生的目标往往是达到其母语者的水平。因此，可能大多数人都经历过，说第二语言的时候，为了接近母语者的发音而以自己特定的发音方式来模仿。研究表明，双语者说两种语言时,与母语者一样,会分别产生显著不同的嗓音音质（Stockmal et al.,2000）。而在以往的第二语言语音研究以及语音学研究中，都以语音的发音性质为研究的主要内容，即都集中在口腔中。如，元音的音色、辅音的发音部位和方法等（孔江平，2001）。

Esling and Wong(1983)指出，要评价第二语言学习者的话语，母语者不能忽视说话者的嗓音特征和语言特征。收听者可能在某种程度上依靠嗓音定势和展唇、鼻音等的调音习惯。据此，Stockmal 等(2004)再提出如果这样的话，可以说语言的“声音”是音系特征和嗓音定势特征结合成的综合体。该综合体就是收听者要解码的。可见，嗓音发声音质在说话的过程中占有非常重要的地位。

“嗓音音质（Voice Quality）”通常指人在说话时一直在出现的习惯性、准永久性的口音特征（Esling，2006）。 该音质主要分成两类（孔江平，2001）：

1.发声音质（phonation quality）定义为通过不同调声发声产生的不同音色，即决定正常嗓音、假声、气嗓音等的声带定势；2.调音音质（articulation quality）定义为通过共鸣调制产生的不同音色，即唇、舌、腭以及喉头高度等的不同定势。

本文将以调音音质数据为参考数据，主要探讨韩国学习者说汉语时的发声音质是否与韩语不同，进一步了解韩国学习者以什么样的嗓音加工方式来体现汉语的嗓音音质特征。

1 实验方法

1.1 发音人

为选好此次研究的被试，事先对学生做了问卷调查。本问卷调查的主要目的为：1.控制母语背景包括文化背景（韩国，首尔标准语）；2.控制年龄（在25-35岁之间）；3.控制性别（男10，女10）；4.控制外语水平和背景（被试学汉语的期间均为4年以上并均有汉语水平考试高级证书，除了英语之外没有熟练的外语），以提高本次实验的准确性。

根据问卷调查的结果，本实验选取了男女韩国学生各10名，一共20名，分别予以考察。为方便起见，在本文中韩国男发音人标为M；韩国女发音人标为F。

1.2 录音材料

录音材料为内容一样，语言不同的两种文件：即，一种是汉语版，另一种是韩语版。汉语和韩语的“/a/”分别为“啊”和“아”，二者都没有具体词义，只表示感叹。具体内容如下：

汉语：这是汉语“啊”的发音。

韩语：이것은 한국어 “아” 발음 입니다.

1.3 实验设备和软件

本次实验所使用的主要设备和软件有：

1. Kay 公司生产的PENTAX Model 6103电子声门仪（也称喉头仪或EGG）。该仪器主要测量的是喉头间的电阻变化，即采集声门阻抗信号，以能够观察声门的变化、声带的振动方式等。

声门仪的基本工作原理是把一对电子感应片(electrodes)分别固定在喉结两边，贴紧甲状软骨。发声时，一个高频信号以非常微弱的流量从一个电子感应片发送，被另一个接收。采集到的信号就反映了声门阻抗的变化。当声带完全接触，即声门完全关闭时，阻抗值最小；当声带分开，即声门完全开启时，阻抗大大增加。从声门仪提取出来的参数可以很好地用来描写不同语言的发声类型，因而自从Fabre（1957）[1]试用以来，被语音学家广泛使用。

下图是一个典型的EGG信号参数定义示意图(孔江平，2008)，具体的定义如下：

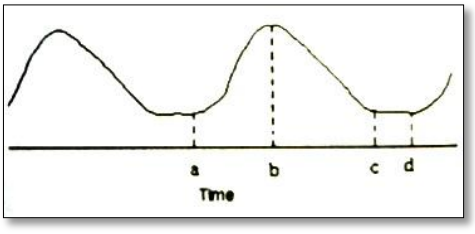


图1.1 EGG信号参数定义示意图

其中ad为周期，ac为闭相，cd为开相，bc为声门正在打开相，ab为声门正在关闭相。

基频 = 1/周期(ad)

开商 = 开相(cd)/周期(ad)×100%

速度商 = 声门正在打开相(bc)/声门正在关闭相(ab)×100%

2. Adobe Audition 1.5。这是一款功能比较齐全的录音和音频编辑软件。它支持128条音轨、多种音频特效、多种音频格式，可以很方便地对音频文件进行修改、合并。其中本文主要应用了录音和切音功能。该音频设置提供不同参数的双通道模型，本文以左通道为语音信号，右通道为EGG信号进行录音，分别采集声学信号和生理信号。

3. Voicelab 程序。这是北京大学中文系语音乐律实验室编写的一个基于Matlab软件平台的GUI（Graphic User Interface）参数提取程序。Voicelab程序主要提取EGG基频、开商、速度商参数，此外，也可通过菜单选项进行快速博立叶变换（FFT）以及线性预测系数（LPC）计算等。它的方便之处在于能够同时实现嗓音参数的提取以及保存功能。以Excel文件格式，即可以保存参数的原始数据，同时还可进行参数的归一化处理。

1.4 录音和数据处理

录音是每个发言人在北京大学中文系语音乐律实验室单独进行的。录音时，使用EGG和麦克风可以把发音人的嗓音信号和语音信号同时采集到电脑上的音频处理软件内。每个发音人的录音采样频率为22050Hz，保存为wav格式。录音材料每个发音人读两遍，等于从每个发音人可分别采集到汉语和韩语各两个/a/样本，男女各10名，总共80个。将80个原始信号文件用Audition的功能经过切音和降噪后，利用Voicelab软件来提取出所需的参数，参数包括基频、开商、速度商、共振峰、谐波幅值。将参数保存到Excel表格中，然后对每个人的汉语和韩语参数进行对比，并对男声、女声、汉语和韩语参数分别计算平均值。最后使用Excel的统计功能对汉语和韩语的参数进行标准偏差、成对双样本均值分析及相关系数分析，从而考察韩国学习者的汉语和韩语参数之间是否有显著差异以及参数之间的相关度。

2 实验结果

下文根据孔江平（2001）的音质分类 - 发声音质（Phonation Quality）和调音音质（Articulation Quality），对整个数据给予分析并探讨韩国学习者的汉语嗓音加工方式。

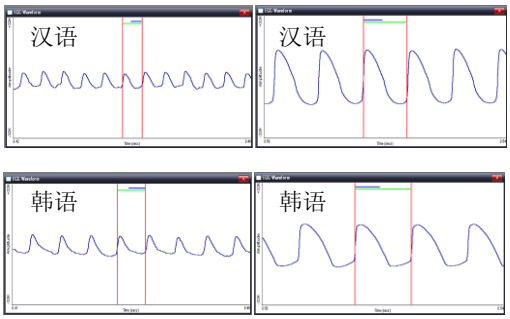
[1] Fabre,P., (1957) Un procédé électrique percutané d’inscription de l’accolement glottique au cours de la phonation. Bull. Nat. Méd.

对于发声音质，本文采用的方法是：1）基频、2）开商、3）速度商分析和 4）第一、二谐波的差值。至于调音音质，本文采用共振峰参数量化调音音质的特征。

2.1 发声音质（Phonation Quality）

“发声音质”反映的是声带振动的不同方式。上文所提到的许多研究表明，不同语言的声带可以以不同的方式振动。在语音学研究中发展出了许多研究嗓音发声类型的方法，其中采用 EGG 信号是获取声带振动参数是最方便的方法（孔江平，2001）。

下面给出的图，就是分别使用 EGG 设备采集到的男女韩国学习者发的汉语/a/和韩语/a/的 EGG 信号波形图。从图能看出他们之间有些差别，汉语的周期短，开相时间也较短，而韩语的整个周期和开相时间都较长。从 EGG 波形表现我们可以大致了解到二者的区别，下文集中探讨汉、韩发声音质的具体特征。



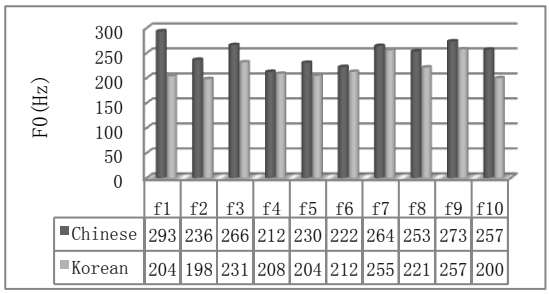
（a） （b）

图 2.1 （a）女声的 EGG 信号波形图 （b）男声的 EGG 信号波形图

嗓音发声类型的声学参数有许多，其中常用的参数有：1）基频(F0)；2）开商（OQ）；3）速度商（SQ）和 4）谐波振幅差值。

2.1.1 基频参数比较

不同发声类型会有不同的声带振动频率，因此，基频的高低是体现嗓音特征的一个重要方面。图 2.2 是女声的汉语和韩语基频参数比较图，显示每个人的具体情况。图中深灰色、浅灰色柱子分别表示汉语基频值、韩语基频值的数据。



	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Chinese	293	236	266	212	230	222	264	253	273	257
Korean	204	198	231	208	204	212	255	221	257	200

图 2.2 女声基频参数比较

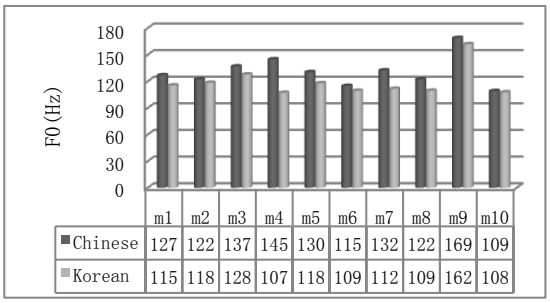
一般女声的正常嗓音为 150-300Hz，从上图 2.2 显示的个人数据而言，可以说韩国学习者的汉语和韩语的基频值都属于正常嗓音。

从整个数据来看，无论是个人数值还是平均值汉语的基频值都高于韩语的基频值。虽然也有汉、韩基频相差极小的女声数据，如，f4、f6 和 f7，但所有女声的汉语基频变化方式是一致的，无一例外。

表 2.1 女声基频参数统计

	平均	变化率	显著水平
汉语	251	14.4 （%）	0.004 (p<0.05)
韩语	219		

女声的汉语基频平均值为 251Hz，而韩语平均基频值为 219Hz。汉-韩差值平均为 31.6Hz，此数据表示基频变化率为 14%。上面给出的统计结果显示，女声的汉语基频值和韩语基频值之间的显著水平为 0.004，即，二者之间的差异极其显著。该结果表明，女韩国学习者发汉语时，基频与韩语差距明显，会明显高于韩语。



	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10
Chinese	127	122	137	145	130	115	132	122	169	109
Korean	115	118	128	107	118	109	112	109	162	108

图 2.3 男声基频参数比较

上图 2.3 给出的是男声的汉语和韩语基频参数数据。男声的正常嗓音为 100-200Hz，从整个数值范围来看，和女声一样，汉、韩都属于正常嗓音的范围之内。

从个人数据来看，虽然也有汉、韩基频相差极

小的男声数据，如，m2 和 m10，但所有男声的汉语基频值都高于韩语的基频值，也无一例外。汉语和韩语基频平均值分别为 130.7Hz 和 118.6Hz。汉-韩平均差值为 12.03Hz，和女声的数据相比其差值不算很大，但男女声的基频变化方式是一致的。

表 2.2 男声基频参数统计

	平均	变化率	显著水平
汉语	130.7	10.1 （%）	0.006 (p<0.05)
韩语	118.6		

上面给出的男声基频参数统计结果也表明，显著水平为 0.006，小于 0.05。即，男声的汉语基频值和韩语基频值之间存在显著差异。

从上文给出的男女声汉、韩基频参数分析结果得知，韩国学习者说汉语时，无论男声还是女声，在基频的加工方式上都呈现基频变高的趋势，其变化率男女分别为 14.4483% ，10.1439%，统计结果显示男女声基频均发生显著变化。

2.1.2 开商参数比较

开商定义为：开相和周期之比（开商=开相/周期）。在一般的情况下，基频和开商成反比关系（孔江平，2001），从上文的基频分析结果来看，我们可以预测汉语的开商会低于韩语的开商数值。女声的汉语和韩语开商参数比较图如下：

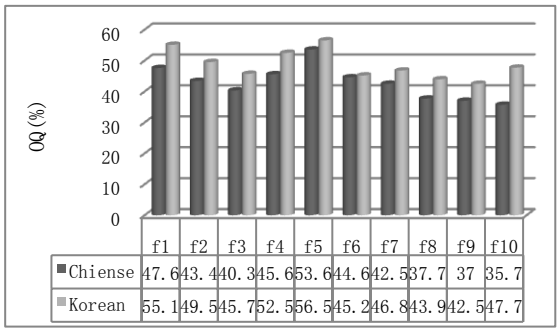


图 2.4 女声开商参数比较

一般来讲，气嗓音开商值为 65%，正常嗓音的开商值为 50%，紧喉音为 30%左右（Laver，1980；孔江平，2001）的话，韩国学习者发的韩语大多数属于正常嗓音，个别参数接近于气嗓音的发声性质。汉语的参数大多数也属于正常嗓音，而有的参数接近于紧喉音的发声性质。

表 2.3 女声开商参数统计

	平均	变化率	显著水平
汉语	42.8	12.5 （%）	0.0002 (p<0.05)
韩语	48.5		

上图给出的统计结果表明，女声的汉、韩开商参数之间变化有显著性差异。其显著水平为 0.0002，显著性非常明显。即，她们说汉语时以拉高基频，将声门开的时间变短的嗓音加工方式来体现汉语元音/a/，10 个数据中无一例外，见图 2.4。

表 2.3 显示，韩-汉变化率为 12.5%，与上文分析过的女声基频变化率（14%）比较接近。这意味着基频值变高，开商值要变低，符合已有的研究结果，即，基频和开商参数呈反比关系。使用统计工具算出实际基频-开商之间相关度，可以得到汉语为-0.67,韩语为-0.62。从此得知，无论汉语还是韩语，均有较大的负关联，呈反比关系。另外，开商大的气嗓音性质带有漏气现象，从开商分析结果来看，女韩国学习者说汉语的时候开相时间变短，随之漏气可能会变小。总之，女韩国学习者说汉语时嗓音加工方式会发生变化。

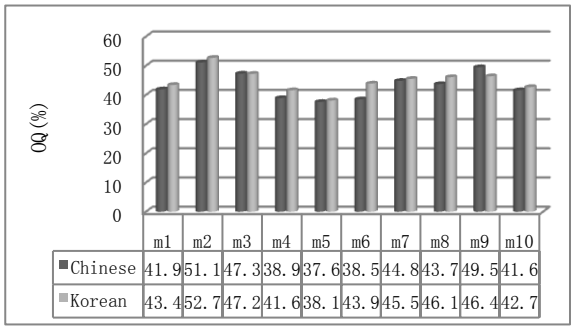


图 2.5 男声开商参数比较图

上面给出的图是男声的开商参数比较图。一般来说，男女开商平均值差别不大，本次实验结果也是如此。

从数值分布来看，无论是汉语还是韩语，都接近于正常嗓音或紧喉音嗓音性质，只是汉语数值更接近于紧喉音嗓音性质，但其差异不大。

从本文的前段分析结果已知，基频和开商有反比关系。女声的汉、韩基频变化率和开商变化率比较一致，分别为 14%和 12.5%，而男声的汉、韩开商变化率与基频变化率不一致，分别为 10%和 4.5%。其原因从图 2.9 中可以找到。从上面给出的男声开商个人数据中可以发现，在 m3 和 m9 的开商数据上出现变异。他们的汉-韩基频差分别是 7.444 和 6.7，在正常情况下，他们的韩-汉开商差值也应该表现为正值，而 m3 和 m9 的数据分别是-0.008 和 -3.18，

均为负值。从男声开商数据的整体表现来说，此变异情况应该属于个人嗓音特征或者计算中的问题，这一点可以通过下面其他参数的数据来校正。

表 2.4 男声开商参数统计

	平均	变化率	显著水平
汉语	42.3	4.4526 (%)	0.009847 (p<0.05)
韩语	44.3		

从统计分析结果来看，男声韩-汉差值变化率为4.452642%，在男声的汉语和韩语开商之间也存在显著差异。与女声不同的是男声的基频参数和开商参数之间没有大的关联，男声汉语和韩语基频-开商相关度分别为-0.15，0.22，结果表明男声基频-开商之间没有明确的相关关系。

从整体变化趋势来看，虽然从韩语到汉语男声的开商变化率为 4.5%左右，与女声的变化率 12.5%相比要低很多，但结果显示男女声的汉、韩开商数值之间均有显著差异。就是说无论男声还是女声，都以不同的发声音质来体现汉语的发声音质。尤其对女声的表现而言，基频-开商之间的相关系数也较高，呈现规律性的嗓音变化趋势。

2.1.3 速度商参数比较

速度商定义为：开启相和关闭相之比（速度商=开启相/关闭相）。速度商是一个很重要的嗓音参数，因为它一方面和音调高低关系密切，另一方面它又体现了男女之间嗓音的差别。开商和速度商在不同语言中差别很大，要根据语种来定。(孔江平，2001)。下文将根据速度商参数探讨汉、韩之间的不同嗓音音质，女声汉、韩速度商实际分析结果如下：

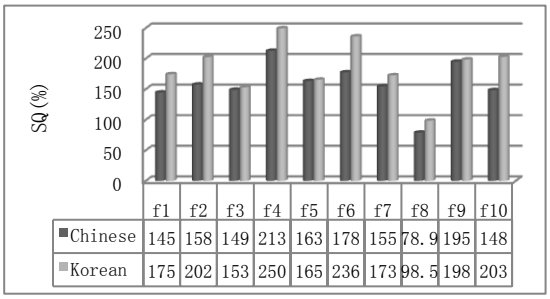


图 2.6 女声速度商参数比较图

整体而言，和女声开商分析结果一样，汉语速度商数值也低于韩语的速度商数值。从数据分布来看，汉、韩大多数数值均在 100-250 之间，属于正常嗓音或气嗓音嗓音音质。女声速度商分析结果和开商参数显示的嗓音性质有所不同，该问题在下文中结合其他参数分析结果继续讨论。

表 2.5 女声速度商参数统计

	平均	变化率	显著水平
汉语	169	13.5204 (%)	0.007182 (p<0.05)
韩语	195		

上文分析过的汉、韩基频平均数值分别为 251Hz，219Hz，其变化率为 14.4%。上表 2.5 显示，速度商的和基频的变化率非常接近。考虑基频和速度商的总体变化率，可以说女声的速度商数据也有所反映和基频之间的关系。实际统计分析结果显示，女声汉语和韩语的基频-速度商相关度分别为-0.56 和-0.26，均呈较弱的反比关系，只是汉语的基频-速度商相关度高于韩语的相关度。

在上图和统计分析给出的结果显示，女声的汉语和韩语速度商之间存在显著差异。更重要的是，上文已分析过的几项参数都显示出韩国学习者用显著不同的嗓音音质来说汉语。即，她们以基频提升、开商和速度商降低的加工方式来体现汉语的嗓音音质特征。

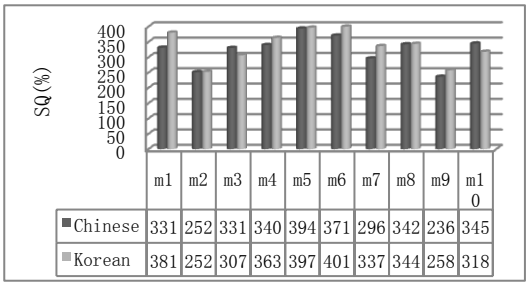


图 2.7 男声速度商参数比较图

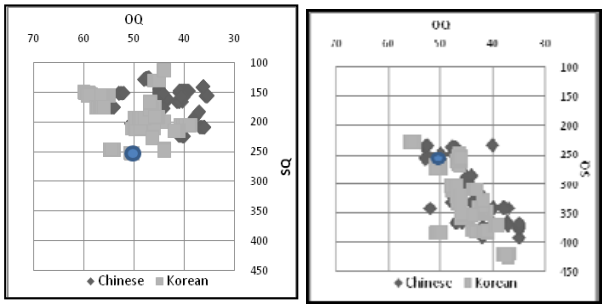
从上图给出的男声速度商个人数据来看，平均汉语速度商数值小于韩语的数值，与女声的结果一致。其中 m3 和 m10 的汉语表现和其他男声数据相反，汉语速度商数值大于韩语的，而韩-汉差值呈负值。基于上文的基频和开商数据分析来看，m3 可能属于个人特征，因为 m3 的开商数值也呈现出和别人相反的变化趋势。至于 m10 的速度商数据，也很可能属于计算问题或者个人嗓音加工方式的问题。另外，m2、m5 和 m8 的数值汉、韩之间的相差不大，差值不到 3%。该结果可以说明，在一般情况下，男生采用基频提升，速度商变低的方式来体现汉语的嗓音特征，但是与女生的结果相比，个人差异比较大。总体上，男女对发声音质的变化趋势比较一致，但其变化程度有所不同。

表 2.6 男声速度商参数统计

	平均	变化率	显著水平
汉语	323.7	3.55 （%）	0.17 (p<0.05)
韩语	335.7		

从上表 2.6 的数据来看，男声汉语速度商低于韩语的速度商数值，从韩语至汉语的变化率为 3.55%左右。与男声基频的变化率相比，此变化率较低，和男声开商数值的情况比较相似。统计结果也表明，汉语和韩语速度商之间没有显著差异。至于基频-速度商相关度，男声的基频和速度商之间呈较弱的反比关系，还是与开商的结果比较一致。

下图 2.8 是用开商和速度商画出的韩国学习者的汉、韩声学发声图，横坐标为开商，纵坐标为速度商。声学发声图可以较直观地显示男女发声音质之间的差异以及汉-韩发声音质的不同。



（a） （b）

图 2.8（a）女声嗓音声学图，（b）男声嗓音声学图

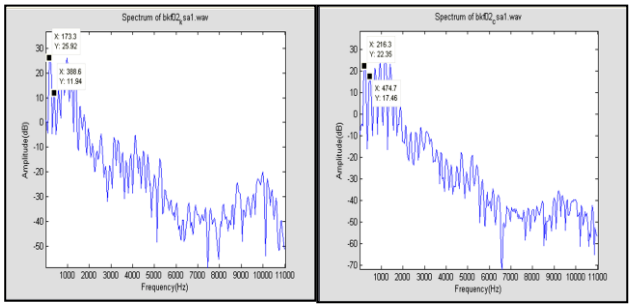
整体来看，男女声的开商分布差异不大，而男声的速度商数值大大高于女声的数值，即，速度商可以把男女的嗓音特征很好的区别开来。另外，男女相比，女声的汉、韩嗓音性质的不同比较明显，而男声汉语和韩语分布呈现基本重合的趋势。

值得注意的是，无论男女，和韩语参数相比，汉语的其他参数更接近于紧喉音的嗓音性质，而男女汉语速度商的数值都低于韩语，不符合其他参数的结果。但，有研究表明（汪锋，2008），在速度商上，紧音不一定比松音大，其原因是否在于韩国学习者的特定发声性质还是其他因素，要进一步的研究。

2.1.4 谐波振幅差值（H1-H2）参数分析

在发声类型的声学研究上，频谱分析是最常用的一种方法，频谱分析主要是测量第一谐波和第二谐波的振幅，其谐波的振幅差一般情况下就能反映声带振动时的紧张程度，这一方法可以用于大多数语言嗓音紧张程度的分析（孔江平，2001；Kirk et al， 1984）。

下图中，H1：第一谐波的振幅，H2：第二谐波的振幅。一般来说，松音第一谐波能量较强，相反较紧的音第一谐波能量较弱。因此，H1- H2 的数值越大嗓音在高频的能量就越大，在生理上表现为声带越紧，反之越松。



（a） （b）

图 2.9(a)韩语和(b)汉语的功率谱对比图

从上面给出的示意图（图 2.9）上看，无论是汉语还是韩语第一谐波的能量都强于第二谐波，但汉语的 H1 和 H2 之间的差异较小。虽然上面给的示意图是一个例子，而结合前文的几项参数分析结果和上面给的汉、韩功率谱示意图，可以预测到韩国学习者说汉语时声带会变紧。下表显示的是韩国男女声谐波振幅差值的具体分析结果：

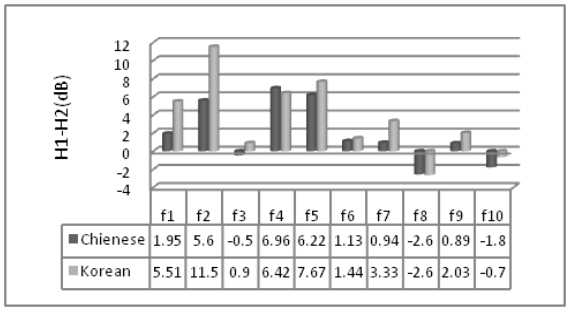


图 2.10 女声谐波振幅差值（H1-H2）比较

上面给出了每个人的数据显示，大多数人的汉语 H1-H2 数值都小于韩语的数值，只有两个例外。对于两个变异而言，在 F4 和 F8 的其他参数中都没有发现相应的特征，因此，有可能属于个人的嗓音特征或者计算上的问题，有待考察。

表 2.7 女声谐波振幅差值数据统计

	平均	变化率	显著水平

汉语	1.9	46.8 (%)	0.02 (p<0.05)
韩语	3.6		

从上表的数据来看，汉语的 H1-H2 平均数值为 1.9，韩语为 3.6Hz，汉语的平均值小于韩语的平均值。显而易见，和韩语相比，说汉语时的声带更紧。其变化率也高达 46%，汉、韩松紧定势表现之间的差异也非常明显，统计结果表明，汉、韩之间存在显著差异。该结果显示，一般来讲，女韩国学习者说汉语的时候，由声带的紧张度要变高的加工方式来体现汉语的嗓音特征。

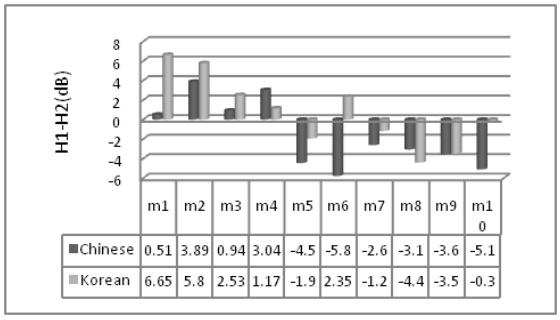


图 2.11 男声谐波振幅差值（H1-H2）比较

表 2.8 男声谐波振幅差值数据统计

	平均	变化率	显著水平
汉语	-1.6	325.7 (%)	0.004 (p<0.05)
韩语	0.7		

表 2.8 的分析结构显示，和女声的结果一样，汉语和韩语之间的变化率非常大，高达 326%左右。整个统计结果表明汉、韩之间谐波振幅差值有显著差异，即，男声说汉语的时候，声带的紧张度要增高，从而体现不同嗓音音质特征。

总之，结合上文的几项分析结果看，不管男声还是女声，韩国学习者采用较紧的嗓音音质来体现汉语。

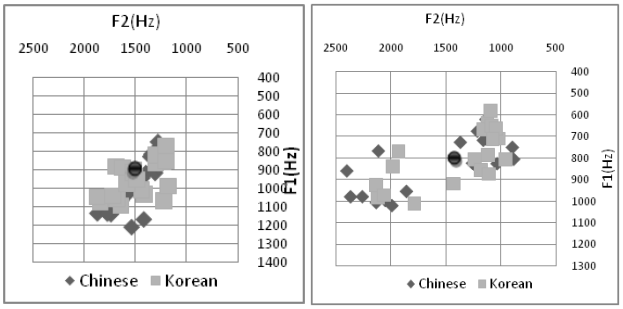
2.2 调音音质（Articulation Quality）

由于发音器官是一个整体，发声和调音之间难免相互影响（方特，1994）。因此，从调音参数不仅可以了解发声对调音的影响，同时也能反过来证明发声类型的不同。对于调音音质，现有的语音学从舌位的高低和唇型的圆展来定义，在声学上主要是从共振峰结构上来量化（孔江平，2001）。

上文通过对体现发声音质的参数，即，基频、开商、速度商和谐波振幅差值进行分析，从而了解了韩国学习者说汉语时的嗓音加工方式。下面将要给出的就是韩国学习者说汉语时的共振峰结构（F1 和 F2），以了解他们说汉语时的喉调音音质特征。

2.2.1 共振峰参数

下面给出的声学元音图可以反映汉、韩两种语言的共鸣特性。从元音共振峰的数据来看，发声音质的变化对调音的影响确实有一定的规律性。发不同的元音时，舌位变化较大会影响到喉头位置的高低，从而影响到发声。随着喉头的提高，所有共振峰数值呈现提升的趋势，其基频也会升高。相反，喉头降低的话，共振峰数值和基频值也随之变低（Laver，1980）。从以上的研究结果来看，可以预测韩国学习者说汉语时的共振峰数值会高于韩语的共振峰数值。男女韩国学习者的汉、韩实际共振峰结构如下：



（a） （b）

图 2.12 （a）女声声学元音图 （b）男声声学元音图

一般来说，韩国女性的/a/元音 F1 平均为 900Hz，F2 为 1500Hz 左右，男性 F1 平均值为 800Hz 左右，F2 为 1400Hz 左右（Chung. et al.，2010）。

从图 18 和下面给出的表 2.9 和 2.10 来看，除了女声的汉语 F1 数值和男声的汉语 F2 数值比韩语的稍微高一点之外，汉语和韩语的共振峰分布比较重合，看不出汉、韩共振峰之间的明显差异。

从个人女声数据而言，汉语 F1 和 F2 数值都高于韩语的达到 50%，F1 或 F2 数值高的各 20%。至于男声，45%的数据的汉语 F1 和 F2 数值高于韩语的，F1 高的有 10%，F2 高的为 25%。许多研究表明，紧元音的 F1 和 F2 数值比松元音要高，两个共振峰高分别表征开口度大及发音部位靠前的调音特征（孔江平，2001；Laver，1980）。Sundberg and Nordstrom（1976）也曾指出，喉头提升定势会引起提高基频和整个共振峰数值。可见，虽然其变化不太明显，但大多数韩国学习者说汉语时，或多或少还是出现开口度变大，喉头提升及舌位靠前的变化。虽然，影响 F2 变化的因素不仅是舌位前后一个，如，唇的

突出或圆展等。但是考虑元音/a/的调音特征，本文只用喉头升降和舌位前后的定势来描写汉、韩发声音质之间的不同。

表 2.9 （a）女声（b）男声共振峰数据统计

		平均	变化率	显著水平
F1	汉语	977	2.7	0.3
	韩语	952	（%）	(p<0.05)
F2	汉语	1497	0.7	0.7
	韩语	1486	（%）	(p<0.05)

（a）

		平均	变化率	显著水平
F1	汉语	822	1.1	0.7
	韩语	813	（%）	(p<0.05)
F2	汉语	1536	6.9	0.4
	韩语	1482	（%）	(p<0.05)

（b）

上面给出的平均数据表明，不管男女，汉语 F1 和 F2 的平均数值都大于韩语的共振峰数值。但是，由于其汉、韩之间的变化率较小，个人差异也大，故统计结果表明汉、韩之间还是没有显著差异。有趣的是，从平均变化率来说，女声倾向于以开口度变大的方式来体现汉语的调音音质，男声多用舌位前拉的方式来实现汉语的元音音质。

但是，结合前文的发声音质的情况，总的来说，韩国学习者主要由不同的发声音质来体现汉语的嗓音音质特征，而调音音质的影响不是特别明显。

3 讨论

由于汉语普通话的发声类型不具有明确的语音学意义，学习汉语时发声类型一般不是学习的重点。但，上文的分析结果表明，韩国学习者感知汉语语音时，还注意到发声音质的不同，因此，在表现上，以基频变高，而声带变紧的方式来体现汉语。

韩国学习者怎么能观察到发生音质的不同？从第二语言学习的角度，第一个考虑的是母语的影响。值得关注的是，韩语辅音体系中有松紧的对立。相关研究表明，韩语紧音的表现为基频高于松音，喉头比松音要紧(Cho et al.,2002; Kim et al.,2002)。众所周知，在语音学习领域，Flege(1995)提出的语音学习模型(SLM—Speech Learning Model)和 Best 等人提出的语音感知同化模型(PAM—Perception Assimilation Model)(Best，1994)是关于 L2 语音感知与产出的最具影响力的模型。语音感知同化模型和语音学习模型都认为语言学习者主要是参照母语的语音体系来学习第二语言语音的。

下面给出的就是包含韩语松紧辅音的 CV 音节中 V 的参数和上文已分析过的韩国学习者的汉、韩参数比较。由于谐波振幅差值是一个很稳定且能比较全面反映声源性质的参数，用此参数和基频参数组合也能很好地表示嗓音的性质。

表 3.1 韩语松紧音对立参数[2]和韩国学习者的汉、韩参数（男声）

	松音	紧音	变化率	韩语	汉语	变化率
基频	124	138	11%	119	131	10%
H1-H2	4.2	- 4.8	214%	0.72	-1.62	326%

显而易见，韩语松紧音的基频数值和松紧之间的变化率接近于本次分析的韩国学习者汉、韩参数数值。韩语松紧音的 H1-H2 数值差比较大，相比，韩国学习者的韩语和汉语数值差异较小，但二者之间的变化趋势还是一致的。

结合语音学习和感知模型来看，由于韩语中有松紧嗓音对立，韩国学习者有可能对其嗓音音质特征比较敏感。因此，感知汉语的嗓音音质时，有可能参照韩语的松紧嗓音音质，把母语的紧音音质对应为汉语的嗓音音质。不同母语背景的学习者对目标语会有不同的表现，母语对目标语发声音质的影响值得进一步研究。

4 结论

通过本次实验和讨论，对韩国学习者的汉语嗓音音质加工方式可得出以下结论：

韩国学习者确实由不同的嗓音音质定势来体现汉语的嗓音音质特征。从韩语到汉语，整体嗓音音质变化趋势男女声比较一致，女声的嗓音音质变化比男声更为明显。具体由基频提升，开商和速度商降低及声带的紧张度增高的方式来体现汉语的嗓音音质特征，相比调音音质的变化不太显著。

从第二语言语音学习的角度而言，从韩国学习者的汉语嗓音音质定势的表现来看，母语迁移现象除了调音音质之外，对于目标语的发声音质也会起作用。因此，不管目标语的嗓音音质变化有无区别意义，第二语言学习者到了一定水平，目标语本身的语音特征引起或者为了提高目标语的母语性，自

[2] 韩语松紧音数据来自于Cho et al.,2002; Kim et al.,2002

然关注到目标语的嗓音音质特征，并试图以特定的方式来体现。

本文所采用的方法和样本有限，未必反映全部情况，基于本次实验的结果，希望可以进一步研究，最后能够运用到汉语语音教学领域。

参考文献

[1] 孔江平（2001）《论语言发声》，中央民族大学出版社

[2] 汪锋，孔江平（2008）武定彝语松紧音研究，《语音乐律研究报告》，北京大学中文系语言学实验室

[3] Best, C.T. 1994. The emergence of native language phonological influences in infants : A perceptual assimilation model. In: Goodman, J., Nusbaum, H.C. (eds), *The Development of Speech Perception*. Cambridge : MIT Press, 167-224.

[4] Cho Taehong, Jun sun-ah, Ladefoged, P. 2002. Acoustic and Aerodynamic Correlates of Korean Stops and Fricatives. *Journal of Phonetics.* 30, 193-228.

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

[6] Fant, G. , Gauffin, J. 1994. *Speech Science and Speech Technology.* Shangwu publishing house. Beijing.

[7] Flege, J.E. 1995. Second Language speech learing: Theory, findings and problems. In: Strange, W. (eds), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.* Baltimore: York Press, 233-277.

[8] Kim Soohee, Emily Curtis. 2002. Phonetic Duration of Korean /s/ and its Borrowing into Korean. *Japanese /Korean Linguistics*. 10, 406-419.

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

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

[11] Stockmal, V., et al. 2000. Same talker, different language. *Applied Psycholinguistics*. 21,383–393.

[12] Stockmal, V., et al. 2004. Judging voice similarity in unknown languages. *In Proceedings of the 17th Congress of Linguists*, Prague.

汉语普通话双音节(C1)V1#C2V2 音节间音段协同发音研究

Cross-syllable segmental coarticulation in (C1)V1#C2V2 sequences in Standard Chinese: An Electropalatographic and acoustic study

李英浩 孔江平

摘要： 本文使用动态电子腭位技术研究汉语普通话双音节后字音节 C2V2 对前字韵母 V1 的发音过程和 F2 轨迹的方向以及变化幅度的影响。实验结果发现 C2 发音部位对 V1 后过渡段的腭位参数和 F2 轨迹的方向以及变化幅度有显著影响。第二，C2 发音方式对 V1 后过渡段的腭位参数和 F2 轨迹也有显著的影响。第三，V2 能够跨越 C2 影响 V1 后过渡段的腭位参数和 F2 运动轨迹以及变化幅度。此外，V2 的圆唇特征也有可能跨越 C2 影响 V1 后过渡段 F2 轨迹。实验结果表明 C2 对 V1 的影响以及 V2 对 V1 的影响与 C2 的舌位发音限制条件密切相关，而 V2 的圆唇特征对 V1 的影响有可能不受 C2 舌位发音限制条件的制约。

Abstract: This article uses electropalatography to investigate the anticipatory coarticulation of C2V2 to V1 and the resultant F2 transition for V1. Results show that both place and manner of intervocalic consonants affect the linguapalatal contact profile and direction as well as magnitude of the F2 transition of V1. The anticipatory vocalic effect is also found significant for EPG parameters and F2 transition for V1. The rounded feature of V2 tends to affect the F2 transition of V1. The article contends that the anticipatory consonantal and vocalic effects are constrained by the articulatory constraints for intervocalic consonants, and the rounding feature associated with V2 has an independent status in the coarticulatory process

关键词：动态电子腭位；协同发音；

Key words: electropalatography; coarticulation;

协同发音不仅是言语产出理论的核心内容，也是言语工程中的难点问题。由于发音器官的连续运动以及不同发音器官运动的相位关系受多种因素的影响而产生偏离目标动作的情况，从而导致声学层面上的参量存在大量变异现象。

声学层面上的协同发音现象实际上是多个发音器官相互配合产生的结果，不同发音器官的动作之间既存在严格的动作相位关系，也会因表达的需要以及其他随机因素的影响而发生变化，前者决定了产出了什么，而后者则决定了如何产出[1]。生理层面的发音器官的运动与声学层面的声学参量之间的共变关系异常复杂，两者往往是一对多或者多对一的关系。然而，跨语言的协同发音的感知研究表明，听音人能够准确定位自身语言中产生协同的根源以及声学效果，并且能够从复杂的声学信号中恢复说话人语音信号中包含的信息结构，但是对其他语言中协同发音的根源并不敏感[2]。从目前的言语产出的研究结果来看，协同发音过程受到发音规则的支配，因此是可预测的过程[3]。

言语工程界对协同发音的生理过程的关注度不断提高。在语音识别领域，邓力的动态语音模型（dynamic speech modeling）对协同发音采用的两步分析法的第一步就是通过声道运动特性生成目标音段的共振峰目标，并根据已有的知识对连续的目标进行平滑处理[4]。邓的模型与Keating的语窗模型（window model）相似，都是通过对连续的音段的发音动作目标进行连线的方式（interpolation）来模拟协同发音的过程，区别在于语窗模型认为不同音段的发音动作对变异的容忍度不同[5]，而邓力的模型并没有考虑到音段的发音动作限制条件。第二，邓的模型并没有考虑到隔位音段相互影响的情况。这也是言语工程界忽视的问题[6]。语音合成领域对协同发音的处理主要基于三音子模型，建立与上下文相关的语音单元数据库，在合成过程中对不同单元的谱包络进行拼接。这种方式的缺点在于需要有较大的语音数据库，同时不能灵活地合成出自然的语流。

1. 汉语普通话协同发音的相关研究

吴宗济和孙国华对汉语普通话C1V1#C2V2 序列中音节内和音节间的协同发音做了的细致的声学分析。他们的实验结果表明音节间C2 对V1 的协同发音程度主要取决于两个音段的发音部位。V2 可以跨越C2 使得V1 的声学目标发生偏离（参见该文图 2 中tu#tV2 和tu#kV2 的例子），其变化幅度与V1 和V2 的目标值的距离有一定的比例关系。V1 对后字音节的顺向协同发音只表现在后字音节的前过渡的幅度，这是通过元音环境对C2 的影响而使得C2 的发音部位发生变化产生的。最后，V1 并不影响V2 的声学目标[7]。

陈肖霞基于更大的语料库分析了V1#C2V2 序

列中音段间的逆向协同发音。她发现相同发音部位、不同发音方式C2 条件下的V1 的后过渡段的F2 运动轨迹方向相同。C2 为唇音的时候存在元音间的逆向协同发音，而在舌尖音和舌面后音的时候则不存在这种现象。此外，后音节对前字韵母的影响范围至多占前字韵母时长的 17%[8]。

Ma等使用EMMA分析了汉语普通话V1#C2V2 序列的协同发音。他们发现V2 的舌位动作只影响C2 舌体在高低维度上位置，但是不能跨越C2 来影响V1 的目标动作。他们认为这与普通话的言语规划的最小单元为音节有关系[9]。

本文采用动态电子腭位技术分析汉语普通话双音节(C1)V1#C2V2 中后字音节音段对前字音节韵母 V1 的逆向协同发音的生理过程和声学表现，并为下一步建立结构化的模型奠定基础。基于已有研究结果，本文提出如下的研究问题：汉语普通话双音节 V1#C2V2 中不同 C2 和 V1 组合条件下的协同发音生理过程以及产生的声学结果有何差异？第二，C2 发音方式对 V1 的影响是否有差异？第三，V2 对 V1 的跨音段协同发音产生的条件和结果是什么？

2. 语料和标注

语料来自国内某大学语音学实验室的汉语普通话动态电子腭位数据库的 982 个双音节样本。大多数样本的 C1 位置上为不送气舌尖中音/t/， V1 为单元音/i, a, u, ɿ, ʅ/，C2 为 21 个声母，V2 除了上面五个单元音外还包括/y/。

发音人为一名女性播音员。每个双音节用正常语速朗读 2-4 遍。语音样本信号包括 EPG 信号（采样频率为 100Hz），喉头仪信号（EGG）和声学信号（后两种信号的采样频率均为 22kHz）。对语音样本按照生理和声学的标准进行标注，主要包括如下内容：使用 EGG 信号确定 V2 的起点，如果 C2 为浊声母，则从波形和语图判断 V2 的起点。V1 结束点一般通过波形和语图来判断，同时也参考 EGG 和 EPG 信号。比如 V1 后接塞音或者塞擦音的时候，V1 的结束点设置在 EPG 出现阻塞帧的附近，此时，虽然声带仍然保持振动状态，但是 V1 的 F2 基本上消失。塞音和塞擦音的起点为口腔除阻帧之后，语图上出现冲直条之前的位置上。如果 C2 为清擦音或者浊音，C2 的起点时间即为 V1 的结束时间。使用 LPC 协方差方法获取元音段的头三个共振峰。具体做法：声学信号下采样到 10kHz，LPC 系数选 10-12 阶，窗长为 20ms，步长为 5ms。最后对照倒谱语图对 F2 轨迹进行手动调整。本文只选取 V1 后半段的 F2 运动轨迹作为分析的对象。

3. 方法

腭位参数包括前腭接触面积（ANT）和后腭接触面积（PST），ANT 为前腭区四行电极中接触的电极数与这个区域电极总数的比值，PST 为后腭区四行电极接触情况，计算方法同 ANT（参见图 1）。为了考察 C2V2 对 V1 的逆向协同发音影响，计算 V1 声学结束点之后一帧与 V1 中点附近一帧的腭位参数的差，分别为 delta_ANT 和 delta_PST。

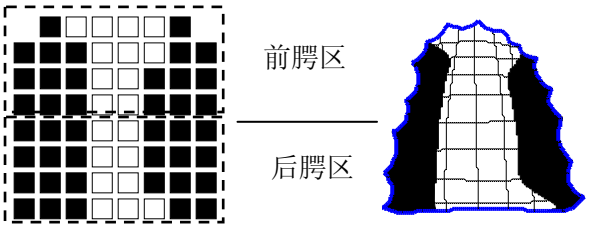


图 1　假腭的前腭区和后腭区

在声学方面，计算 V1 段 F2 轨迹结束点附近的 F2 和 V1 中点附近的 F2 的差（delta_F2）作为衡量 C2V2 对 V1 的逆向协同发音的影响程度。delta_F2 的符号代表 F2 运动轨迹的方向，如果符号为正，则说明 F2 轨迹上升，否则下降。delta_F2 的绝对值表示 V1 后过渡段 F2 轨迹的变化幅度。

研究变量有三个：首先，C2 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）。其次，C2 发音方式包括塞音（S），塞擦音（T），擦音（F），鼻音（N）和边音（L）。最后，V2 类型与声母条件有关，一般情况下 V2 的类型包括单元音/a, i, u/。如果声母无法与/i/相配，则使用/ei/来代替前高元音（文中表示为/i/）。舌尖音无法与/i/相配，只能使用同部位的舌尖元音相配（舌尖前元音表示为 i1，舌尖后元音表示为 i2）。舌面前音只能与单元音/i, y/相配，本文加入了/iu, ia/的类型考察复合元音后音段是否能够影响 V1（四种情况分别表示为 C2i, C2y, C2u, C2a）。

4. 结果

4. 1. C2 发音部位对 V1 的逆向协同发音影响

图 2 和图 3 分别为发音部位对 V1 的腭位参数和 delta_F2 影响的结果。可以看出，唇音主要降低/i, ɿ, ʅ/的 ANT，/i/的 PST 的降幅超过 10%，但是/ɿ, ʅ/的 PST 基本无变化。舌面后音主要提高 PST 和降低/i, ɿ, ʅ/的 ANT，但是/i/的 PST 基本无变化。其他部位条件下，/a, u/的 ANT 和 PST 均增加；/i/的 ANT 升幅不超过 20%，而 PST 降幅则超过 10%；/ɿ/的腭位在同部位辅音条件下几乎无变化，在其他部位的

条件下，PST 基本无变化，ANT 的增幅在 10%-15%之间。

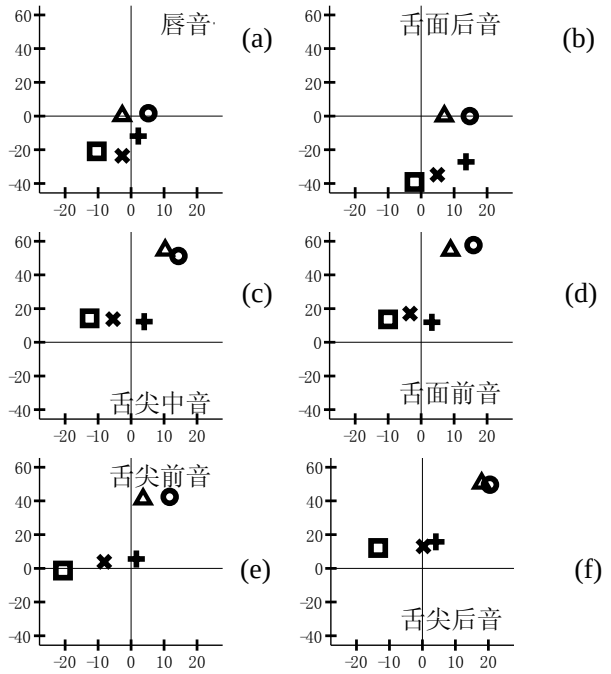


图 2 发音部位对 V1 的 delta_ANT 和 delta_PST 的影响（横坐标为 delta_PST(%)，纵坐标为 delta_ANT(%)；图中方块为/i/，圆形为/a/，三角为/u/，加号为/ɦ/，乘号为/ɣ/）

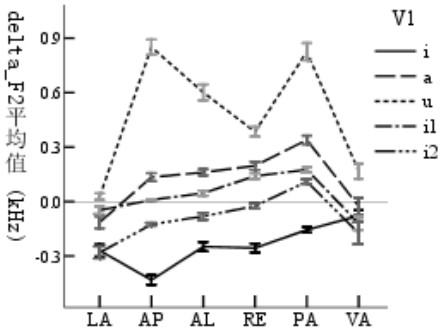


图 3 发音部位对 V1 的 delta_F2 的影响（图中误差范围为±1 个标准误）

从图 3 的结果来看，发音部位影响 V1 的 F2 运动轨迹的方向和变化幅度。固定 V1 条件下，对发音部位进行单因素方差分析的结果表明不同部位脚下的 delta_F2 具有显著差异（p<0.0001）。表 1 为使用 Turkey 方法进行事后检验的结果。

V1 为/i/的条件下，F2 过渡的方向全部为负值，F2 变化幅度在 AP 条件下最大，这与 PST 的下降有关（图 2(e)）。在舌面音的条件下最小，说明 ANT 的变化并不影响 F2 变化幅度，PST 的下降可能与 F2 的下降有关（图 2(b, d)）。其他发音部位的影响居于两者之间，并且 F2 的下降与 PST 的下降有关。相关分析结果表明 V1/i/的 delta_F2 与 delta_PST 显著相关(r=0.46, p<0.001)，而与 delta_ANT 没有关系。

V2 为/ɦ/的条件下，F2 过渡方向以 AP 为界相反。VA 条件下 F2 降幅显著大于 LA 条件下的情况，这可能与 ANT 的下降有关（图 2(a,b)）。AP 与 V2 的发音部位相似（图 2(e)），因此 delta_F2 趋近于零。RE 和 PA 条件下 F2 的升幅显著大于 AL 条件下的情况，这可能与 ANT 的上升具有关系。相关分析结果表明 V1/ɦ/的 delta_F2 与 delta_ANT 显著相关（r=0.577, p<0.001），而与 PST 的变化关系不大。

表 1 V1 固定，C2 发音部位对 delta_F2 的影响

V1	不同发音部位的 delta_F2 比较	p
i	AP(-)<LA RE AL(-) RE — < PA(-) AL — < VA(-)	<0.07
ɦ	VA(-)<LA(-) AP(+) LA — <AL(+)<RE PA(+)	<0.09
ɣ	LA(-)<VA AP AL(-) VA —— <RE(-)<PA(+)	<0.09
a	LA VA(-)<AP AL RE(+)<PA(+)	<0.01
u	LA(+)<VA<RE<AL<PA AP(+)	<0.01

注：delta_F2 均值按照实际数值由小到大排列，斜线之后括号内的符号表示 F2 轨迹变化方向（在不影响理解的情况下省略）。每行中“—”表示在此符号之上的项目与此符号所在行小于号右侧的项目没有显著差异，显著性水平选择多重比较的最大值。

V2 为/ɣ/的条件下，只有在 PA 的条件下 V1 的 F2 过渡方向为正号，这可能与 ANT 的上升有关系（图 2(d)）。LA 条件下 F2 的降幅显著大于 VA、AP 和 AL 条件下的情况，ANT 和/或 PST 的下降都可能导致这种情况。相关分析结果表明 delta_F2 与 delta_ANT 和 delta_PST 均显著相关（r=0.497, r=0.28, p<0.001）。

V1 为/a/的条件下，LA 和 VA 使得 F2 下降，然而前者 F2 降幅较大，而后者趋近于零。其他部位条件下，F2 的上升与 ANT 和 PST 的增加有关系。相关分析结果表明 delta_F2 与 ANT 和 PST 的变化显著相关（r=0.581, r=0.517, p<0.001）。

V1 为/u/的条件下，F2 的过渡方向均为正号。C2 为 LA 的时候，F2 过渡的升幅趋近于零。C2 为 VA 条件下出现的 F2 过渡的上升可能与 C2 的发音方式有关系，将在 4.2 中做分析。其他 C2 部位条件下，F2 上升可能主要与 ANT 有关系。相关分析结果表明 delta_F2 与 ANT 变化的相关程度要大于 PST 的变化（r=0.568, r=194, p<0.001）。

固定发音部位，对不同 V1 条件下的 delta_F2 的单因素方差分析结果发现 V1 的效应显著（p<0.0001）。表 2 为使用 Turkey 方法进行事后检验的结果。C2 为唇音的时候，除了/u/，其他 V1 条件下的 F2 过渡方向都为负号。在舌尖元音条件下，F2 的下降与 ANT 的下降有关（图 2(a)），而在/i/的条件下则与 ANT 和 PST 都可能有关。相关分析

结果表明唇音条件下的 delta_F2 与 ANT 和 PST 的变化都相关（r=0.531, r=0.375, p<0.001）。

C2 为舌尖音的条件下不同 V1 的 F2 过渡方向和变化幅度相对平行。/i, ɥ/情况的 F2 的下降可能与 PST 的降低相关，而/ɻ, a, u/条件下 F2 的上升可能与 ANT 和 PST 都有关系。对三个部位的腭位和 F2 参数的相关分析结果表明，ANT 和 PST 的变化与 delta_F2 均显著相关（r=0.573, r=0.605, p<0.001）。舌面前音的情况与舌尖音的相似，只是 F2 下降只出现在 V1/i/条件下。

舌面后音情况下，delta_F2 的变化幅度不大。/u/条件下 delta_F2 为正的情况可能与 PST 的增加有关，而在其他 V1 条件下 delta_F2 为负则可能与 ANT 的降低有关。相关分析结果表明 delta_F2 与 ANT 和 PST 都有关系（r=0.342, r=0.332, p<0.001）。

表 2 C2 固定，不同 V1 条件下 delta_F2 的差异

发音部位	不同 V1 的 delta_F2 比较	p
唇音	i ɥ(-)<a ɹ (-) a - < u(+)	≤0.001
舌尖前音	i (-)<ɻ (-)<ɹ (+)<a (+)<u (+)	<0.01
舌尖中音	i (-)<ɻ (-)<ɹ (+)<a (+)<u (+)	<0.01
舌尖后音	i (-)<ɻ (-)<ɹ a (+)<u (+)	<0.0001
舌面前音	i (-)<ɹ ɥ (+)<a (+)<u (+)	<0.0001
舌面后音	ɥ ɹ i a (-)<u (+)	<0.05

4. 2. C2 发音方式对 V1 的逆向协同发音影响

单因素方差分析结果表明同部位的 PA，AP 和 RE 的不同发音方式条件下的 V1 的 delta_PST 和 delta_F2 没有显著差异，delta_ANT 的显著差异只与舌前部与齿龈/硬腭成阻有关。

从图 4 可以看出，唇音条件下发音方式对不同的 V1 影响不同。单因素方差分析结果表明，V1 为/a, ɥ/时候的 delta_PST 和 delta_F2 不受发音方式的影响；而当 V1 为/i, u, ɥ/的时候，至少有一个参量受到发音方式的影响。V1 为/i/的时候，ST 和 FR 条件下的 V1 的 PST 的降幅显著大于 NA 条件下的情况(p<0.05)，这可能与 C2 成阻过程中口腔内压力有关。FR 条件下的 F2 的降幅显著大于其他发音方式条件下的情况（p<0.01），这个结果既与 PST 的快速下降有关系，也与擦音前 V1 结束的方式有关。V1 为/u/的时候，FR 条件下的 F2 上升，而其他条件下的 F2 的变化趋近于零。这可能与 V1 后过渡段下唇内收有关。V1 为舌尖后元音的时候，鼻音条件下的 F2 的降幅显著大于塞音条件下的情况（p<0.01），显然，这与 PST 的变化关系不大，而与 ANT 的变化有较大的关系。

舌尖中音条件下 C2 发音方式对 delta_PST 和 delta_F2 的影响相对一致。单因素方差分析结果表明 LA 条件下的 delta_PST 的降幅显著大于其他发音方法条件下的情况（p<0.05）。在声学方面，除了 V1/ɻ/以外，LA 条件下的 V1 的 F2 轨迹降幅显著大于其他发音方式条件下的情况，而 F2 轨迹的升幅则显著小于其他发音方式条件下的情况。由于边音的发音过程中舌面两侧下降，因此高元音向 LA 过渡过程中 PST 会迅速减少，而低元音或者后元音向 LA 过渡中后腭区域几乎没有电极接触。

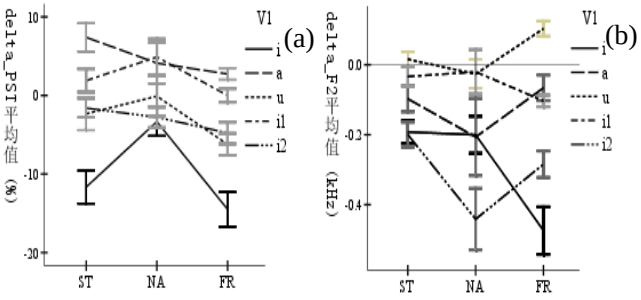


图 4 唇音条件下发音方式对 V1 的 delta_PST(a)和 delta_F2(b)的影响

舌面后音两种发音方式导致的 delta_PST 的差异是显而易见的。由于 ST 在后腭区域形成阻塞，因此 V1 后过渡段的 PST 总是增加，而同部位 FR 的 PST 要减小，以便于高速气流通过。从图 6 可以看出，擦音条件下 V1 的 F2 轨迹的降幅一般变大，这与 PST 的下降以及 V1 结束的方式有关。但是当 V1 为/u/的时候，F2 轨迹方向为正，且升幅却大于 ST 条件下的情况，这可能与 V2 的逆向协同发音有关系，这将在 4.3 中做具体分析。

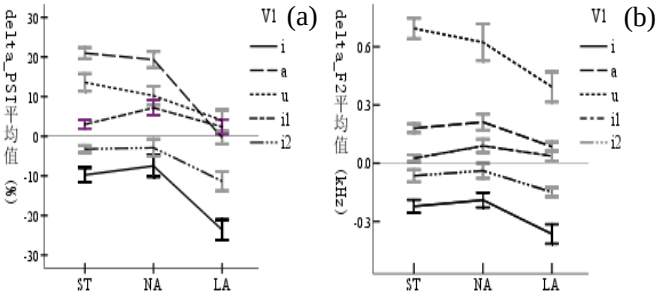


图 5 舌尖中音条件下发音方式对 V1 的 delta_PST(a)和 delta_F2(b)的影响

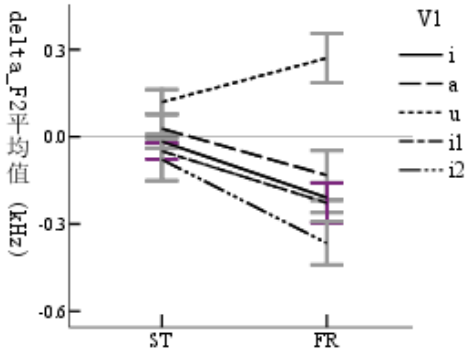


图 6 舌面后音条件下发音方式对 V1 的 delta_F2 的影响

4.3. V2 对 V1 的逆向协同发音影响

元音间的逆向协同发音不仅受到元音类型的影响，也受到中间辅音发音限制条件的影响[10]。我们按发音部位分别计算每个V1 条件下非对称元音序列V1#C2V2（V1≠V2）和对称元音序列V1#C2V2（V1=V2）的V1 后过渡段的腭位参数和delta_F2 是否具有显著差异。V1 为舌尖元音的时候，只有在同部位的舌尖辅音条件下才有对称元音序列，我们把/V1#C2i/看做对称元音序列，并作为比较的基准。同样，舌面后音条件下使用/i#C2ei/作为V1 为/i/条件下的对称元音序列。由 4.1 可知，V1 的delta_F2 在不同的C2 条件下与两个腭位参数的相关度不同。但是在分析元音间协同发音的时候，情况可能与 4.1 的结果稍有差异。分析过程中，V1 为/i, a, u/以及V1 为舌尖元音、C2 为AL和PA的时候使用delta_PST，其他情况下使用delta_ANT（表中使用“*”来表示）。

表 3 C2 为唇音条件下 V2 对 V1 的影响

V2	*i*		*a*		*u*	
V1	EPG	F2	EPG	F2	EPG	F2
i	(-)	(-)	(-)<i	-	(-)<i	(-)<i
a	(+)>a	(+)>a	(±)	(-)	-	(-)<a
u	(+)>u	(+)>u	-	(+)>u	(-)	(-)
ɪ*	(+)	(+)	(-)<i	(-)<i	(-)<i	(-)<i
ɻ*	(-)	(-)	(-)<i	(-)<i	(-)<i	(-)<i

注：表中灰色区域为每行数据比较的基准，单元格中括号内的符号表示参数的变化的方向，(±)表示基本没有变化。白色单元格内的出现“-”的时候表示没有显著差异。其他单元格中出现的括号内符号表示参数变化方向，之后的不等号表示与非对称元音序列条件的参数对称元音序列条件下的参数显著差异的方向。

表 3 是 C2 为唇音条件下 V2 对 V1 后过渡段腭位参数和 delta_F2 影响的结果。可以看出，高元音/i, ɪ, ɻ/的腭位参数和 F2 均出现下降的情况（/ɪ#C2i/情况除外），并且降幅与对称元音序列条件下的情况存在显著差异。V1 为/a, u/的时候，V2/i/使 V1 的 PST 增大，同时 F2 的轨迹方向变正；V2 为其他元音的时候，/a#C2u/的 F2 轨迹方向由高到低，而/u#C2a/表现为由低到高。

表 4 C2 为舌尖前音条件下 V2 对 V1 的影响

V2	*ɪ*		*a*		*u*	
V1	EPG	F2	EPG	F2	EPG	F2
i	(-)	(-)	-	-	-	-
a	-	-	(+)	(+)	-	(±)<a
u	-	(+)>u	-	(+)>u	(+)	(+)
ɪ*	(+)	(+)	-	-	-	(-)<ɪ
ɻ*	(+)	(+)	-	-	-	(-)<ɪ

从表 4 可以看出，不同 V2 条件下 V1 的腭位参数基本不受 V2 的影响，这说明如果 V1 向舌尖前音过渡中很少受到 V2 的影响。从 F2 轨迹的变化幅度来看，V1 为非圆唇元音的时候，V2/u/条件下的 V1 的 delta_F2 呈现下降的趋势（V1/i/的情况除外）；V1 为圆唇元音、V2 为非圆唇元音条件下，V1 的 delta_F2 显著大于/u#C2u/中的情况。这个结果说明元音的唇形有可能跨越辅音而影响辅音对侧元音的 F2 运动轨迹。

表 5 C2 为舌尖中音条件下 V2 对 V1 的影响

V2	*i*		*a*		*u*	
V1	EPG	F2	EPG	F2	EPG	F2
i	(-)	(-)	(-)<i	(-)<i	(-)<i	(-)<i
a	(+)>a	(+)>a	(+)	(+)	-	-
u	-	(+)>u	-	(+)>u	(+)	(+)
ɪ	(+)	(+)	-	(-)<i	-	(-)<i
ɻ	(-)	(-)	(-)<i	(-)<i	-	(-)<i

注：V2/y/对 V1 腭位参数和 F2 的参数与/i/类似，此处不作考虑。

从表 5 可以看出，V1/i/的 PST 一般下降，但是在 V2/a, u/条件下 V1/i/的 PST 的降幅显著增大，同时，delta_F2 的降幅也显著大于/i#C2i/中的情况。V2/i/使 V1/a/的 PST 增加。同时，V1/a, u/的 delta_F2 的升幅显著大于对称元音序列中的情况，这种现象也出现在/u#C2a/中。在/ɪ#C2i/中，V1 的 PST 增加，F2 过渡方向由低到高；而在/ɻ#C2i/中，V1 的 PST 降低，同时 F2 也下降。V2/a, u/条件下 V1/ɪ/的 F2 轨迹方向与对称元音序列异号，而 V1/ɻ/的 F2 的降幅则显著大于对称元音的情况。

表 6 C2 为舌尖后音条件下 V2 对 V1 的影响

V2	*ɻ*		*a*		*u*	
V1	EPG	F2	EPG	F2	EPG	F2
i	(-)	(-)	-	-	-	(-)<ɻ
a	-	-	(+)	(+)	-	(+)<a
u	-	(+)>u	-	(+)>u	(+)	(+)
ɪ*	(+)	(+)	-	-	-	(+)<ɻ
ɻ*	(+)	(+)	-	-	-	(-)<ɻ

从表 6 可以看出，在舌尖后音条件下，V1/i/的 PST 下降，V1 受 V2 的影响来自/u/，后者使 V1/i/的 F2 的降幅显著大于/i#C2ɻ/的情况。同舌尖前音的情况类似，V2/u/对 V1 的 delta_F2 影响显著，这说明 V2 圆唇元音有可能独立于 C2 的发音动作而对 V1 后过渡段的 F2 产生影响。

从表 7 可以看出，V1 向 PA 的过渡过程比较稳定。圆唇元音显著降低 V1/i/的 F2，同时使得 V1/a, ɪ/的 F2 的升幅降低。当 V1 为/u/的时候，V2/y/使

V1/u/的 F2 的升幅显著低于/u#C2u/的情况。V1 为舌尖后元音的条件下则没有发现 V2 对 V1 的影响。上面的结果既表明 V2 的圆唇特征有可能跨越 C2 影响 V1 的声学特征，又说明这种影响可能受到 V1 和/或 C2 唇形特征的制约。

表 7 C2 为舌面前音条件下 V2 对 V1 的影响

V2	*i*		*y*		*u*	
V1	EPG	F2	EPG	F2	EPG	F2
i	(-)	(-)	-	(-)<i	-	(-)<i
a	-	-	–	(+)<a	-	-
u	-	-	-	(+)<u	(+)	(+)
ɪ	(+)	(+)	-	(+)<i	-	(+)<i
ʅ	(-)	(-)	-	-	-	-

注：方差分析结果表明，后字音节为/C2a/的时候的腭位和 F2 的变化与后字音节为/C2i/的情况没有显著差异，因此只在 V1 为/a/的时候作为比较的基准。

从图 8 可以看出，V1/i/向 VA 过渡过程不受 V2 的影响，然而这个结果并不可靠，因为假腭无法捕捉 VA 成阻部位的确切位置。从 V1 的 F2 过渡方向和变化幅度来看，V2/u/要么使得 V1 的 F2 过渡方向发生变化，要么使得 F2 的变化幅度增大。而 V2/ei/增大 V1/a/的 F2 上升幅度，并使 V1/u/的 F2 轨迹方向发生变化。

表 8 C2 为舌面后音条件下 V2 对 V1 的影响

V2	*i(ei)*		*a*		*u*	
V1	EPG	F2	EPG	F2	EPG	F2
i	(+)	(±)	-	-	-	(-)<i
a	-	(+)>a	(+)	(+)	-	(-)<a
u	-	(+)>u	-	(+)>u	(+)	(-)
ɪ*	(−)	(−)	-	-	-	(-)<i
ʅ*	(-)	(+)	-	-	-	(-)<i

5. 结论

对汉语普通话双音节中 V1 后过渡段的动态发音过程和声学的分析结果表明，C2 发音部位和发音方式对 V1 后过渡段的腭位参数变化和 F2 轨迹的方向以及变化幅度有显著影响。同时，V2 能够跨越 C2 影响 V1 的腭位参数和 F2 运动轨迹。此外，V2 的圆唇特征也有可能跨越 C2 影响 V1 的 F2 轨迹，但是这需要结合动态唇形研究才能得到可靠的结论。实验结果表明 C2 对 V1 的影响以及 V2 对 V1 的影响与 C2 的舌位发音限制条件密切相关，而 V2 的圆唇特征对 V1 的影响有可能不受 C2 舌位发音限制条件的制约。

参 考 文 献

[1] Lindblom, B. Explaining phonetic variation: a sketch of the H and H theory[A]. In W. Hardcastle and A. Marchal（eds.），Speech Production and Speech Modeling[C]. Dordrecht: Kluwer, 1990.

[2] Beddor, P. S. A coarticulatory path to sound change[J]. Language, 2000, 85(4), 785-821.

[3] Browman, K., Goldstein, L. Articulatory gestures as phonological units[J]. Phonology, 1989, 6, 201-251.

[4] Deng, L., Yu, D., Acero, A. A bidirectional target-filtering model of speech coarticulation and reduction: two-stage implementation for phonetic recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(1),256-265.

[5] Keating, P. The window model of coarticulation: articulatory evidence[A]. In Kingston, J., Beckman, M.E. (Eds.). Papers in Laboratory Phonetics I: Between the Grammar and Physics of Speech[C]. CUP, 1990.

[6] Tatham, M., Morton, K. Speech Production and Speech Perception[M]. Palgrave Macmillian. 2006.

[7] Wu, Z. J., Sun, G. H. An experimental study of coarticulation of unaspirated stops in CVCV contexts in Standard Chinese[Z]. Annual Report of Phonetic Research, Beijing, Institute of Linguistics of CASS, 1989.

[8] 陈肖霞. 普通话音段协同发音研究[J]. 中国语文. 1997, 5, 345-350.
CHEN Xiaoxia. Segmental coarticulation in Standard Chinese[J]. Zhong Guo Yu Wen, 1997, 5, 345-350.

[9] MA Liang, Perrier, P, DANG Jianwu. A study of anticipatory coarticulation for French speakers and for Mandarin Chinese speakers. Proc. PCC 2008, Beijing.

[10] Recasens, D., Pallares, M.D., Fontdevila, J. A model of lingual coarticulation based on articulatory constraints[J]. JASA, 1997, 102, 544-561.

普通话/s/的动态发音过程和声学分析

The articulatory and acoustic analysis of apical fricative /s/ of Standard Chinese

李英浩

Li Yinghao

摘要：本文使用 EPG 分析单音节和双音节后字声母/s/的动态发音动作和频谱统计参数。研究结果表明：（1）双音节中擦音的成阻的起始方式和时间依据不同的前字韵母而不同。（2）前字韵母对擦音段动作的影响的程度和时间范围很小，擦音段主发音器官动作和舌页的耦合动作相当稳定。后字元音对擦音动作的影响主要体现在舌体动作和 sV 交界点附近舌尖动作的变化。（3）擦音段频谱不受前字韵母的影响，而与后字圆唇元音密切相关。此外，在 sV 交界点附近，擦音频谱与腭位参数的变化存在一定关系。

Abstract: The articulatory and acoustic properties of apical fricative /s/ of Standard Chinese is analyzed electropalatographically and acoustically. The results show that: (1) the approach interval of constriction and gestural timing rely on the preceding rhyme in bisyllable sequences; (2) The carryover effect of the preceding rhyme on the realization of the following apical fricative is fairly limited; The coupling effect between the tongue tip and blade is rather stable across phonetic contexts; However, the anticipatory effect of the following rhyme on the preceding apical fricative is represented in the tongue dorsum gesture and tongue tip movement at the CV juncture; (3) The spectrum of the apical fricative /s/ is not influenced by the preceding rhyme but is highly correlated with the round feature of the following rhyme. Additionally, the spectrum of the apical fricative is correlated with the EPG parameters at the CV juncture.

关键词：动态电子腭位 齿龈擦音 协同发音 频谱

Key words: electropalatography; apical fricative; coarticulation; spectrum

1. 引言

研究表明/s/的发音动作和频谱受到元音环境的影响[1-6]。X 光和 MRI 研究发现摩擦音/s/的发音部位受周边元音的影响较小，但是不参与形成阻碍的舌区易受元音的协同发音影响[7, 8]。EPG 研究发现/s/的主发音器官的舌腭接触相对稳定，发音动作限制（articulatory constraint）较强[9, 10]，因此不易受到语音环境的影响，反而对周边元音产生较强的协同发音影响[11]。元音对/s/频谱的影响主要表现在 F2 附近[12]，并且与不同元音环境的/s/的舌腭接触面积正相关[13]。此外，圆唇元音降低前腔的共鸣频率[25,26]。

汉语普通话的摩擦音/s/的 X 光和静态腭位研究表明/s/的发音动作是舌尖和齿龈前成阻，舌中线下凹[14,15]。已有的普通话 EPG 与国外研究的结果类似[16, 17]。/s/的中心频率和下限频率分别为 7 kHz 和 5 kHz 以上[18]。后接圆唇元音时候，中心频率和下限频率都降低[19]。

已有 EPG 研究大多通过分析擦音的静态腭位和声学特征研究擦音发音动作受语音环境的影响，本文将进一步考察摩擦音/s/作为单音节声母和双音节后字声母的动态发音过程以及声学特征。

2. 方法

2.1. 语料和录音

语料来自北京大学语言学实验室的汉语普通话动态电子腭位语料库。单音节包括/s/为声母，韵母首元音为/a, u, ɿ /的所有带调音节，共 31 个单音节。双音节词前字声母大多为齿龈音，便于生理信号和声学信号的对齐；前字韵母包括/i, a, u, ɿ, ʅ/以及韵尾/n, ŋ/，后字韵母为/a, u, ɿ /。单音节读 2-3 遍，共 74 个样本；双音节读一遍，共 42 个样本。

发音人为女性，曾经是北京大学广播台的播音员。录音地点在北京大学语言学实验室。发音人录音前佩戴电子假腭进行 20 分钟的适应训练，之后采集发音人的 EPG 信号（采样频率 100Hz），语音信号和 EGG 信号（采样频率为 22kHz）。

2.2. 腭位和声学分析方法

生理/声学信号的对齐和语音标记在 Matlab 上开发的腭位分析系统上进行。单音节中擦音的起点从语图判断，元音起点由 EGG 定位。双音节后字声母/s/往往与前字韵母连接紧密，由语图和 EGG 共同判断。

假腭前三行为齿龈区，中间三行为硬腭区，后两行为软腭区。腭位参数包括接触重心（COG），齿龈区面积（AC）、硬腭区面积（PC）、软腭区接触面积（VC）、齿龈区收紧点最窄宽度（MinW）。COG 定义见[20]，其他接触面积计算方法是接触电极数和区域电极总数的比率（%）。MinW 定义为第一行电极中间没有接触的电极数。擦音频谱统计参数包括平均频

率、斜度和峰度，定义参见[21]，具体算法参考了[22]。

选取 512 点窗长，256 点步（11ms）长进行分帧后计算频谱统计参数，之后进行时间归一化，选取 20 个时间点。腭位参数每 10ms 计算一次，之后进行时间归一化，选取 20 个时间点。单音节韵母选取 100ms 计算腭位参数，时间归一化后选取 20 点。双音节前字韵母和后字元音的腭位参数经时间归一化后，分别选取 20 个时间点。

3. 结果

3.1. 单音节/s/的动态腭位和声学分析

从图 1 可以看出，擦音段AC、PC和MinW只在sV交界点附近受到后接元音的影响，sa[1]、su和sɿ俩俩存在显著差异，因此可以认为sV交界点前/s/的发音动作十分稳定，AC和MinW与/s/的主发音器官动作密切相关，而PC反映了舌页与舌尖动作的耦合。VC受到后接元音影响的时间范围接近擦音的声学起点。如不考虑AC和PC，COG就与VC呈现负相关的关系，即VC越小，COG就越大。VC的变化反映了舌体（dorsum）的运动特性。由于舌体不参与形成阻碍，因此自由度较大，容易受到语音环境的影响。不同元音对/s/的影响程度也不同。/ɿ/与/s/的腭位参数基本上没有差别；在sV交接点附近，AC、PC和COG的降幅表现为/a/>/u/；/u/不断提高/s/的软腭接触面积，而/a/则反之。最后，从腭位参数的变化来看，舌尖动作的速度快于舌体动作。

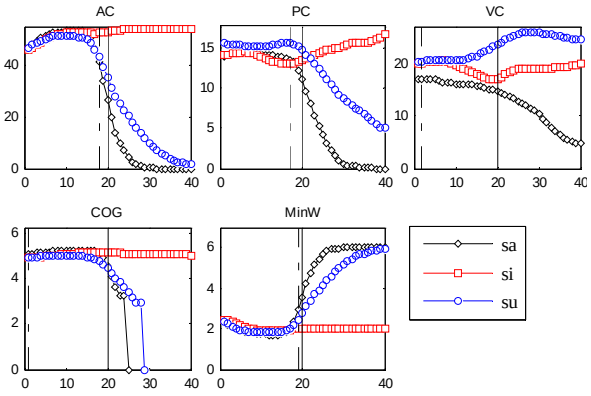


图 1 单音节/s/的腭位参数（图中竖实线为擦音结束点，点杠线为元音影响最远距离，p<0.05）

上述结果说明/s/的发音部位在成阻段非常稳定，元音对擦音动作的影响主要体现在两点：第一，擦音段软腭区接触面积的差异导致 COG 的差异；第二，/ɿ/对擦音动作影响不大，因此可以认为 sɿ在整个音节中发音动作基本保持不变。/u/对擦音的影响集中在软腭区，这种影响可能在擦音起点处就存在。在擦音接近结束点的时候，/a/的过渡起点较早，变化幅度较大。

频谱统计参数反映擦音在整个频段能量分布的特性。从图 2 可以看出，后接元音对擦音频谱的影响表现在两个方面：第一，圆唇元音降低/s/的平均频率，能量重心接近平均频率（偏度≈0），峰度最低。第二，接近sV交界点时的变化因后接元音而不同。sɿ的平均频率微降，能量重心不断远离中心频率，峰度不断升高。这与sɿ在整个音节中发音动作基本不变有关，摩擦产生的湍流往往延伸到元音段[2]，在元音段起始段形成气化嗓音。sa和su在擦音后半段的中心频率和峰度迅速下降，偏度上升。

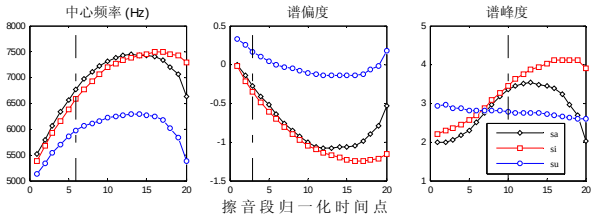


图 2 单音节/s/的频谱统计参数（图中点杠线代表后接元音影响范围的最远点）

3.2. 双音节/s/的动态腭位和声学分析

首先，我们考察 V1sV2 双音节情况。前字韵母为单元音/i, a, u, ɿ, ʅ/，后字元音为/i, a, u, ɿ/（图 3）。单因素方差分析结果表明前字韵尾对/s/的首帧腭位参数有显著影响（p<0.0001），后字元音无显著影响；后字元音对/s/的最后一帧腭位参数有显著影响（p<0.0001），前字韵尾无显著影响。这个结果表明后字元音无法跨越/s/对前字韵母产生逆向系统发音影响。

[1] 为行文方便，用 sa 代替“后接/a/的擦音/s/”。

[2] 此时的噪声源是由于声门不完全关闭产生的。

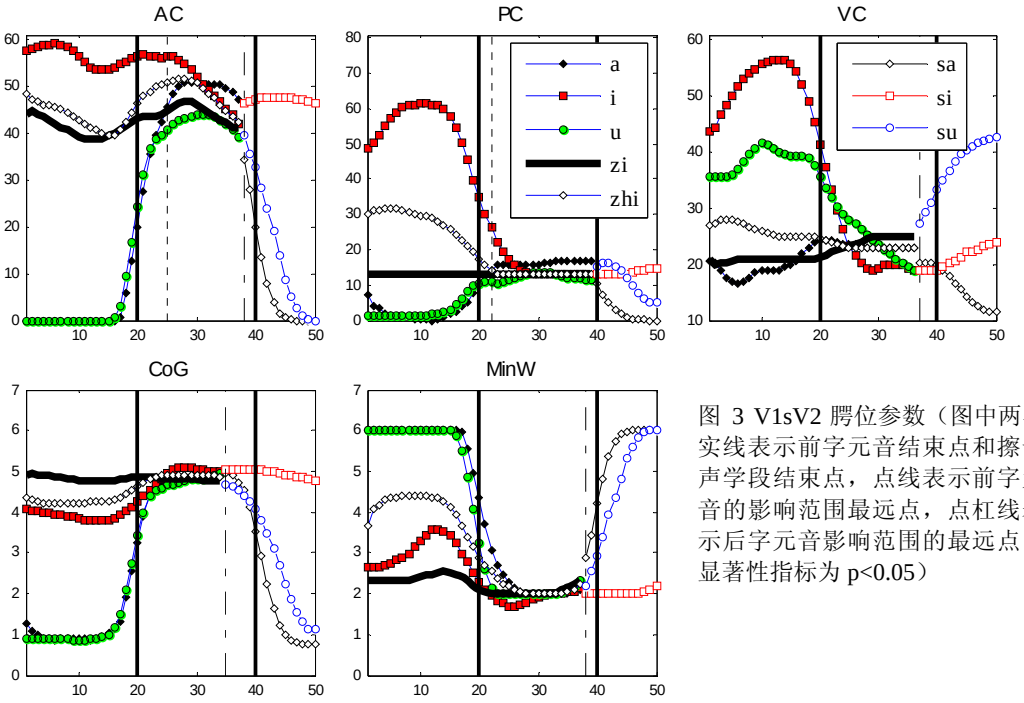


图 3 V1sV2 腭位参数（图中两黑实线表示前字元音结束点和擦音声学段结束点，点线表示前字元音的影响范围最远点，点杠线表示后字元音影响范围的最远点。显著性指标为 $p<0.05$）

/s/的成阻始于前字元音的尾渡，AC，COG 和 MinW 基本上同步变化，从前字元音 3/4 时长点（时间点 15-16）开始指向/s/的发音目标；从 PC 和 VC 上看，/s/的动作始于前字元音 1/2 时长点之后（时间点 13-15）。这个结果说明舌体动作启动时间要早于舌尖动作的启动时间。

擦音段前半部受到前字韵母的影响，但是不同腭位参数受影响的时间范围不同，一般不超过擦音段中点。AC 受到的影响最大，但是 COG 和 MinW 不受前字元音的影响，说明/s/的发音动作已经基本到位，只是由于受到前字元音动作惰性的影响，在 AC 局部区域产生差异。PC 受影响的范围较小，从第二个时间点开始就保持稳定的接触，表明舌页与舌尖动作的耦合关系。VC 受前字元音的影响没有显著差异。

擦音除阻点始于擦音段尾渡，AC、VC 和 MinW 迅速转向后接音段的发音动作，COG 也发生相应的变化，PC 基本上不受后接元音的影响。VC 的变化与单音节情况相似，但是受元音影响的时间范围小于单音节的情况。其变化是否从 V1 向 V2 平滑过渡还需进一步的研究。

从图 4 可以看出，/s/的频谱统计参数受到前字元音影响较小，在持阻阶段主要受到后字圆唇元音的影响，中心频率下降，偏度趋近零，峰度较低。接近擦音结束点的时候，sɿ的频谱参数基本不变，但是 sa 和 su 的变化同单音节情况一致。

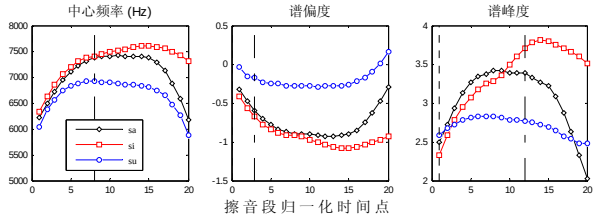


图 4 V1sV2 双音节后字声母/s/的频谱统计参数

另外一种情况是前字带有鼻尾辅音/n, ŋ/。取前字韵母为/an, in, un, ɑŋ, iŋ, uŋ/，后字为/sa, sɿ, su/的双音节分析 V1nsV2 和 V1gs2V 中/s/的发音过程和频谱参数。分析方法与 V1sV2 一样，但是加入了鼻音的声学起点。由于鼻音口腔闭合不完全，鼻音的起点依据语图和鼻音的共振峰而定。

从图 5 可以看出，/s/的成阻过程与鼻音/n/的口腔发音动作部分重合，鼻音的口腔动作先于/s/的成阻动作，但是鼻音口腔动作的目标却是指向擦音的动作。首先不同 V1 与鼻音段时长比率基本符合已有的声学研究结果[24]，不同 V1 后 AC 开始上升的时间点不同；第二，从鼻音段尾渡，AC 和 PC 不断下降，一直延伸到擦音段前半部分。第三，MinW 的变化起点随 V1 的不同而不同，且在鼻音段不为零。第四，鼻音结束帧和擦音的起点帧的腭位参数没有显著差异，表明鼻音的口腔动作的目标就是为了实现擦音的阻碍动作。对比 V1sV2 中擦音段起点帧的腭位参数，可以发现/n/后/s/的起点帧的 AC 和 PC 接近于 isV2 情况下擦音起点帧的 AC（分别为 55%和 20%左右）。

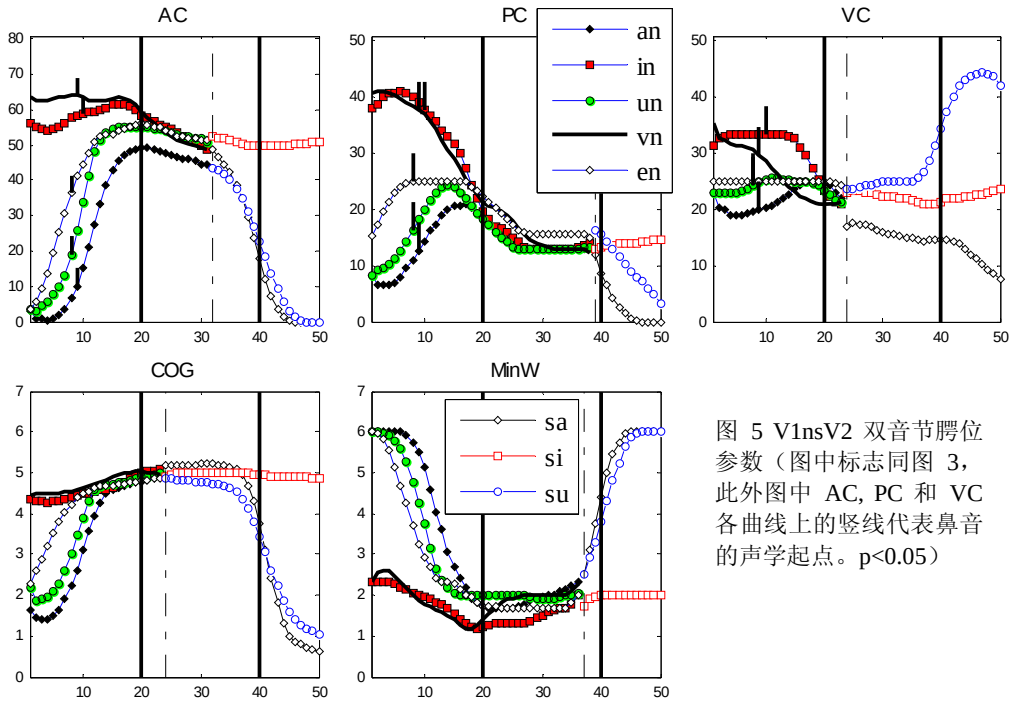


图 5 V1nsV2 双音节腭位参数（图中标志同图 3，此外图中 AC, PC 和 VC 各曲线上的竖线代表鼻音的声学起点。p<0.05）

腭位对于解释 V1snV2 的协同发音相对有限，这是因为鼻音动作不仅涉及口腔的关闭动作，软腭还要降低，从而形成鼻音。但是后接擦音需要形成较高的口腔内压，因此需要及时关闭腭咽门（velopharyngeal port）。从上面的分析来看，鼻音段 AC 和 PC 增大很有可能和口内压要求有关，接触面积增大导致收紧点后声腔体积的减小。有少量样本在接近擦音起点帧的时候形成完全的口腔阻塞。

擦音受到后接元音影响与 V1sV2 情况不同，AC 受后字元音的影响的时间范围大于 V1sV2 和 sV 情况，与前者的差异可能是 V1 的影响，和后者的差异可能来自语速和/或韵律位置的影响。PC 基本上不受影响。VC 变化与单音节情况类似，表明舌体的运动不受舌尖运动的节制。COG 与 VC 的变化密切相关，VC 越大，COG 则越小。MinW 与 sV 和 V1sV2 情况相同。

从图 6 可以看出，擦音频谱统计参数与前字韵母无关，从擦音起点就受到后接元音的影响，中心频率和谱偏度在 sV2 交界点前分别表现为 sɪ>sa>su，su>sa>sɪ，在交界点附近，sa 和 su 的差异消失。si 的谱峰度在整个擦音段都大于 sa 和 su，后两者则没有区别。

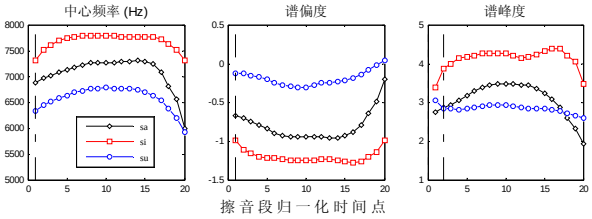


图 6 V1nsV2 双音节后字声母/s/的频谱统计参数

从图 7 可以看出，V1gsV2 情况下，/s/的成阻始于前字韵母中点之后，表现为 AC、PC、COG 和 MinW 在前字韵母段中同步指向/s/的发音动作，其在前字韵母段成阻的时间范围大于 V1sV2 情况，主要原因在于前字韵尾的发音动作与/s/的主发音器官动作并不冲突，因而产生时域上的动作叠加。

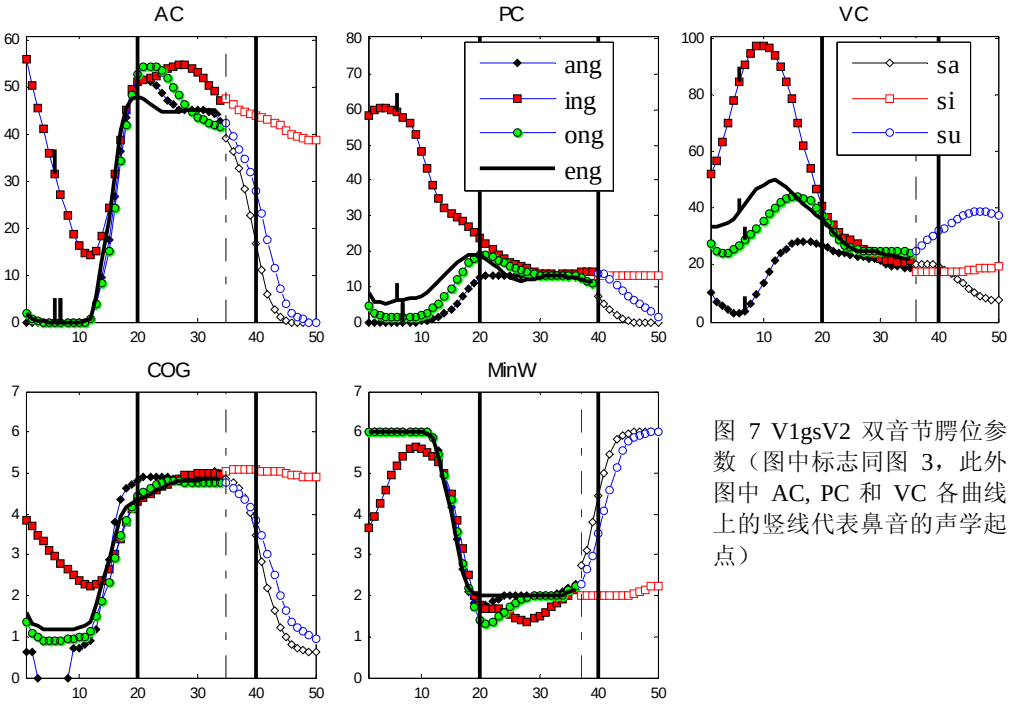


图 7 V1gsV2 双音节腭位参数（图中标志同图 3，此外图中 AC, PC 和 VC 各曲线上的竖线代表鼻音的声学起点）

擦音段腭位参数不受前字韵母的影响。后字元音对/s/的影响与 V1nsV2 情况不同。AC、PC、COG 和 MinW 受到后接元音影响的时间范围与单音节情况类似。VC 的变化与 V1sV2 相似，都是在接近 sV2 边界的时候才表现出显著差异。可能的原因在前字韵尾软腭成阻的动作速度较慢，向后接元音舌体动作过渡慢于舌尖动作。

从图 8 可以看出，同其他情况一样，擦音频谱统计参数只受到后接元音的影响，而与前字韵母没有关系。中心频率受后接元音影响的时间范围较小与样本容量有关。实际上，在擦音起点，中心频率就表现为显著差异，但是在擦音段中部，显著差异消失。谱斜率的结果与前面的结果一致。虽然谱峰度受元音影响的时间范围较小，但是从走向上看，与前面的结果也是一致的。

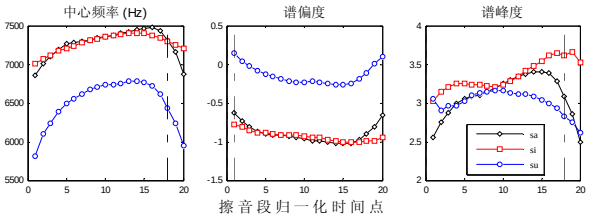


图 8 V1gsV2 双音节后字声母/s/的频谱统计参数

4. 结论

本文讨论了在单音节和双音节后字声母/s/的动态发音动作和频谱统计参数。双音节中擦音的成阻的起始方式和时间依据不同的前字韵母而不同。起始方式取决于前字韵母的发音部位，元音韵母条件下，/s/的舌体动作早于舌尖动作，舌尖成阻起点较晚；如果前字韵尾是前鼻音，/s/的发音动作与前鼻尾口腔动作融合，V1n 的动作过渡的相位关系不变，但是鼻音的口腔动作指向/s/的动作目标；如果前字韵尾是后鼻音，/s/的起始动作较早，并且与后鼻尾的软腭阻塞动作产生时域上的叠加。

前字韵母对擦音段动作的影响的程度和时间范围很小，擦音段主发音器官动作和舌页的耦合动作相当稳定，舌体动作倾向于平滑过渡。后字元音对擦音动作的影响主要体现在舌体动作和 sV 交界点附近舌尖动作的变化。

擦音段频谱不受前字韵母的影响，而与后字圆唇元音密切相关。此外，在 sV 交界点附近，擦音频谱与腭位参数的变化存在一定关系。

参考文献

[1] Heinz, G.W. On the properties of fricative consonants[J]. JASA, 1961, 33(5): 589-596.

[2] Hoole, et al. A comparative investigation of coarticulation in fricative: Electropalatographic, electromagnetic and acoustic data[J]. *Language and Speech*, 1993. 36: 235-260.

[3] Hughes, G.W. et al. Spectral properties of fricative consonants[J]. *JASA*,1956, 28: 303-10.

[4] Narayanan, S.S. et al. An articulatory study of fricative consonants using magnetic resonance imaging[J]. *JASA* 1995, 98: 1325-47.

[5] Nartey, J.N.A. On fricative phones and phonemes[R]. *Working Papers in Phonetics*, 1982, WPP 55.

[6] Subtelny, J.D. et al. Cineradiographic study of sibilants[J]. *Folia Phon*,1972, 24: 30-50.

[7] Shadle, C. et al. An MRI study of the effects of vowel context on fricatives[C]. *Proc. IOA 18:9:1*, 1996, 187-194.
[8] Shadle, C. et al. An MRI study of the effect of vowel context on English fricative[C]. *Acoustics*, 2008, Paris, 5101-6.
[9] Byrd, D. Articulatory timing in English consonant sequences[R]. *UCLA Working Paper in Phonetics*, 1994, 86.
[10] Recasens, D. et al. A model of lingual coarticulation based on articulatory constraints[J]. *JASA*, 1997, 102: 544-561.
[11] Farnetani, E. V-C-V lingual coarticulation and its spatiotemporal domain[C]. In Hardcastle & Marchal (eds.) *Speech Production and Speech Modelling*. 1990. Kluwer Academic Publications: 93-130.
[12] Soli, S. D. Second formants in fricative: Acoustic consequences of fricative-vowel coarticulation[J]. *JASA*, 1981, 70:976-984.
[13] Recasens, D., et al. Co-articulatory variability and articulatory-acoustic correlations for consonants[J]. *European Journal of Disorders of Communications*, 1995, 30: 203-212.
[14] Ladefoged, P. et al. Places of articulation: An investigation of Pekingese fricatives and affricates[J]. *JOP*. 1984, 12: 267-278.
[15] 周殿福，吴宗济. 普通话发音图谱[M]. 北京：商务印书馆，1963.
[16] 李俭，郑玉玲. 汉语普通话发音的变化[J]. 浙江工商大学学报. 2006, 5: 35-40.
[17] 郑玉玲，刘佳.普通话辅音发音部位及约束研究[C]. 第七届中国语音学学术会议论文集，北京，2006.
[18] 吴宗济，林茂灿. 实验语音学概要[M]. 北京：高等教育出版社，1989.
[19] 许毅. 音节与音联。载《实验语音学概要》（吴宗济,林茂灿主编），北京：高等教育出版社,1989.
[20] Hardcastle, W.J., et al. EPG data reduction methods and their implications for studies of lingual coarticulation[J]. *JOP*, 1991, 19: 251-266.
[21] Forrest, K. et al. Statistical analysis of word-initial obstruents: Preliminary data[J]. *JASA*, 1988, 84: 115-23.
[22] Shadle, C.H. et al. Towards the spectral characteristics of fricative consonants[C]. Proc. ICPhS, Aix-en-Provence, 1991, v. 3: 42-45.
[23] Tabain, M. Variability in fricative production and spectra: Implications for the hyper-and hypo- and quantal theories of speech production[J]. *Language and Speech*, 2001, 44: 57-94.
[24] 林茂灿，颜景助. 普通话带鼻尾零声母音节中的协同发音[J]. 应用声学，1992, 1: 12-20。
[25] Toda, M., Maeda, S., Honda, K. Formant-cavity affiliation in sibilant fricatives[C]. In *Turbulent Sounds in Speech*, Mouton de Gruyter, 2010: 1-33.
[26] Stevens, K. Acoustic Phonetics[M]. MIT Press, 1998.
[27] 郑玉玲，刘佳. 普通话 N1C2(C#C)协同发音的声学模式[J]. 南京师范大学文学院学报，2005, 3: 150-157.

普通话舌尖前擦音的动态发音过程和声学分析

An Electropalatographic and acoustic analysis of apical fricative of Standard Chinese

李英浩

Li Yinghao

摘要：传统的腭位照相技术只能静态地观察擦音的发音动作，而动态腭位技术能够实时地反映擦音在成音过程中舌腭接触的过程以及受语音环境的影响。基于北京大学语言学实验室开发的动态腭位和语音分析平台，分析了在单音节声母位置以及双音节后字声母位置上的普通话舌尖前擦音的动态发音过程以及频谱统计参数。研究结果表明：（1）在单音节条件下，声母位置上的舌尖前擦音在齿龈和硬腭区域的接触十分稳定，而软腭区域的接触受到后接元音动作的影响。在擦音段的稳态段，频谱统计参数只与后接元音是否圆唇有关；接近声韵母交界点的时候，频谱统计参数随后接元音的类型而变化。（2）双音节中擦音成阻的起始方式和时间依据不同的前字韵母而存在差异。前字韵母对擦音段动作的影响程度和时间范围很小，擦音段舌尖发音动作和舌页动作的耦合相当稳定。后字元音对擦音动作的影响主要体现在舌面以及与后接元音交界点附近舌尖动作的变化。（3）擦音段频谱不受前字韵母的影响，而与后字圆唇元音密切相关。此外，在声韵母交界点附近，擦音频谱与腭位参数的变化存在一定的对应关系。

Abstract: The articulatory and acoustic properties of apical fricative /s/ of Standard Chinese is analyzed electropalatographically and acoustically. Results show that: (1) The linguapalatal contact at the alveolar and palatal regions is relatively stable before different vowels when the apical fricative is at the onset position in monosyllables. The vocalic influence is manifested solely in the velar region. In the middle interval of the apical fricative, the spectrum is only influenced by the rounded vowels following the fricative. At the CV juncture the spectrum of the fricative starts to co-vary with the articulatory target for following vowels. (2) The approach interval of constriction and gestural timing rely on the preceding rhyme in bisyllable sequences. The carryover effect of the preceding rhyme on the realization of the following apical fricative is fairly limited; The coupling effect between the tongue tip and blade is rather stable across phonetic contexts; However, the anticipatory effect of the following rhyme on the preceding apical fricative is represented in the tongue dorsum gesture and tongue tip movement at the CV juncture; (3) The spectrum of the apical fricative /s/ is not influenced by the preceding rhyme but is highly correlated with the round feature of the following rhyme. Additionally, the spectrum of the apical fricative is correlated with the EPG parameters at the CV juncture.

关键词：动态电子腭位 舌尖前擦音 协同发音 频谱

Key words: electropalatography; apical fricative; coarticulation; spectrum

1. 引言

已有生理和声学研究表明，齿龈擦音/s/的发音动作和频谱受到元音环境的影响[1-6]。X 光和 MRI 研究发现齿龈擦音的发音部位受周边元音的影响较小，不参与形成阻碍的舌区易受元音的协同发音影响[7, 8]。对英语[9]以及加特兰语[10]齿龈擦音的动态电子腭位（electropalatography，简称 EPG）研究结果发现，该音段最大接触帧的腭位相对稳定，不易受到语音环境的影响，反而对周边元音产生较强的协同发音影响[11]。由于 EPG 只能部分地记录软腭区对应的舌面的活动情况，因此以往的 EPG 研究并没有对语音环境对擦音动态发音过程的影响做系统的研究。声学方面的实验结果表明，擦音之后的元音主要影响擦音的第二共振峰（F2）[12]。如果后接元音具有圆唇的特征，由于双唇外凸导致前腔有效长度变长，从而降低前腔的共鸣频率（即擦音的 F2）[13,14]。然而，在实际的研究中很难准确测量齿龈擦音的 F2，大多数的研究使用频谱统计参数来描写齿龈擦音的频谱特征。

普通话的舌尖前擦音/s/的发音部位比较靠前，已有的 X 光和静态腭位数据表明舌尖前擦音的发音动作特点是舌尖和齿龈前成阻，舌中线下凹[15, 16]。近年来对普通话的 EPG 研究结果发现舌尖前擦音最大接触帧的腭位相对稳定，不易受到周边元音协同发音的影响[17, 18]。然而，这些研究仅从部位描写的角度分析了普通话擦音的腭位特征，没有考察在不同语音环境下舌尖前擦音的动态发音过程。声学方面的实验结果发现舌尖前擦音的中心频率和下限频率分别为 7 kHz 和 5 kHz 以上[19]，当擦音后接圆唇元音的时候，中心频率和下限频率都降低[20]。同样，这些研究也只是选取擦音稳态段的频谱特征进行分析。

在上述研究的基础上，我们使用 62 个电极的 EPG 来研究舌尖前擦音在单音节声母位置以及双音节后字声母位置上的动态发音过程，主要考察前字韵母和后字韵母对擦音段腭位的影响。使用频谱统计参数来研究擦音段的动态声学特征，并考察擦音段频谱特征与发音动作之间的关系。

2. 研究方法

语料来自北京大学语言学实验室的汉语普通话动态电子腭位语料库。所选的单音节包括舌尖前元音为声母、韵母首元音为/a, u, ɿ /的所有带调音节，共 31 个单音节。双音节语料的前字韵母为单元音/i, a, u, ɿ, ʅ/以及以没有介音开头的鼻韵母

/an, in, un, yn, ən, aŋ, iŋ, uŋ, əŋ/，双音节后字韵母为单元音/a, u, ɿ/。单音节每字读 2-3 遍，共 74 个样本，双音节读一遍，共 42 个样本。发音人为女性，曾经是北京大学广播台的播音员。录音地点在北京大学语言学实验室。录音前发音人佩戴电子假腭进行 20 分钟的适应训练，之后采集发音人的 EPG 信号（采样频率 100Hz）、语音信号和 EGG 声门信号（采样频率 22kHz）。

在北京大学语言学实验室开发的动态电子腭位和语音分析平台上对生理/声学信号进行标注以及参量提取。单音节中擦音的起点由语图判断，元音起点根据 EGG 信号自动定位。双音节后字声母一般与前字韵母连接紧密，因此由语图和 EGG 共同判断擦音的起点。

根据发音人硬腭生理特征以及发音部位的特点把假腭分为三个功能区域，前三行定义为齿龈区，中间三行为硬腭区，后两行为软腭区（见图 1）。腭位参数包括三类：首先是各功能区域的接触面积，包括齿龈区面积（AC）、硬腭区面积（PC）和软腭区接触面积（VC），计算方法为各功能区域接触电极数目和该功能区域电极总数的比率（%）。第二类参数为接触重心（COG），主要用于衡量接触电极的分布特点，计算方法见式（1）[21]。第三类参数为齿龈区收紧点最窄宽度（MinW），主要衡量擦音成阻的特点，定义为第一行电极中没有接触的电极数。腭位参数每 10ms 计算一次，之后进行时间归一化，选取 20 个时间点。单音节条件下的韵母选取韵母起始段的 100ms 计算腭位参数，时间归一化后选取 20 个时间点。双音节前字韵母和后字元音的腭位参数经时间归一化后，分别选取 20 个时间点。频谱统计参数包括平均频率、斜度和峰度 [22, 23]。选取 512 点窗长，256 点步长（11ms）进行分帧后计算每帧的频谱统计参数，之后进行时间归一化，选取 20 个时间点。

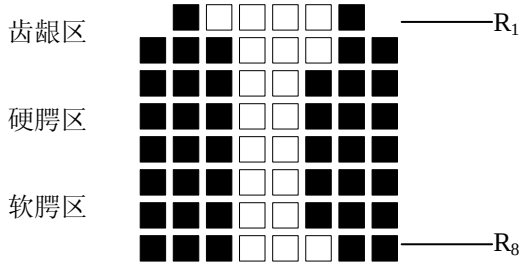


图 1 根据发音人假腭生理结构和发音特点对假腭的功能分区（该帧腭位为普通话/i/的腭位，图中每个方块代表一个电极，黑色方块代表电极接触，空白则代表没有接触）

$$COG=(R_8\times0.5+R_7\times1.5+R_6\times2.5+R_5\times3.5+R_4\times4.5+R_3\times5.5+R_2\times6.5+R_1\times7.5)/\sum_{i=1}^{8}R_i \quad (1)$$

— Ri，i=1,2,3…8 每行接触电极的数目；

3. 实验结果

3.1. 单音节声母的动态腭位和声学分析

从图 2 可以看出，舌尖前擦音的腭位参数 AC、PC 和 MinW 只在声韵母交界点附近受到后接元音的影响，而在擦音段的大部分时段内不受后接元音的影响。AC 和 MinW 反映了舌尖在齿龈前区域形成阻碍的发音动作，在擦音稳态段不易受到后接元音的影响。PC 反映舌面前区域的运动状态，由于受到舌尖/舌页动作的限制，舌面前区域的舌腭接触在擦音的稳态段也不易受到元音动作的影响。VC 主要反映舌面后的舌腭接触情况，可以看出，VC 几乎从擦音的声学起点开始就受到后接元音动作的影响，后接低元音的时候软腭区域的接触显著低于高元音条件下的情况。COG 与 VC 呈现负相关的关系，VC 越小，COG 就越大。由于舌体不参与形成阻碍，因此自由度相对较大，容易受到语音环境的影响。后接元音对/s/的影响程度不同。舌尖前擦音和舌尖前元音的腭位基本上没有差别，主要参数基本上保持一条直线。在声韵母交界点附近，AC、PC 和 COG 的降幅表现为/a/>/u/；此外，/u/不断提高/s/的软腭接触面积，而/a/则反之。

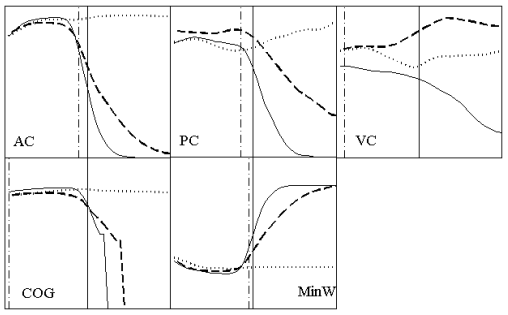


图 2 单音节声母位置上的舌尖前擦音动态腭位参数图示（图中实线代表/sa/，点线代表/sɿ/，杠线代表/su/；竖实线为擦音结束点，点杠线为元音影响最远距离，p<0.05）

上述结果说明舌尖前擦音的发音部位在成阻段非常稳定，元音对擦音动作的影响主要表现在两点：第一，擦音段软腭区域由于受到后接韵母动作的影响，因此，在不同的韵母条件下擦音段的 VC 和 COG 存在显著差异。第二，舌尖前元音对擦音段的舌腭接触基本上没有影响。/u/对擦音的影响集中在软腭区，这种影响可能在擦音起点处就存在。/a/对擦音的主发音器官动作的影响出现在声韵母交界点之前，对 AC 和 MinW 的影响早于其他韵母的情况。

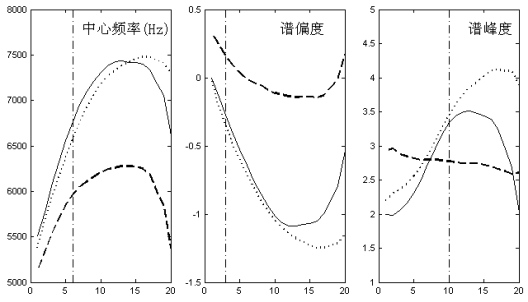


图 3 舌尖前擦音的频谱统计参数（图中曲线标识如图 2）

频谱统计参数反映擦音在整个频段能量分布的特性。从图 3 可以看出，后接元音对擦音频谱的影响表现在两个方面：第一，圆唇元音降低舌尖前擦音的平均频率，能量重心接近平均频率（因为偏度约等于 0），峰度最低。第二，声韵母交界点之前的变化取决于韵母的类型。/sɿ/的平均频率微降，能量重心不断远离中心频率，峰度不断升高。这与/sɿ/在整个音节中发音动作基本不变有关，摩擦产生的湍流往往延伸到元音段[1]，在元音段起始段形成气化嗓音。/sa/和/su/在擦音后半段的中心频率和峰度迅速下降，偏度上升。

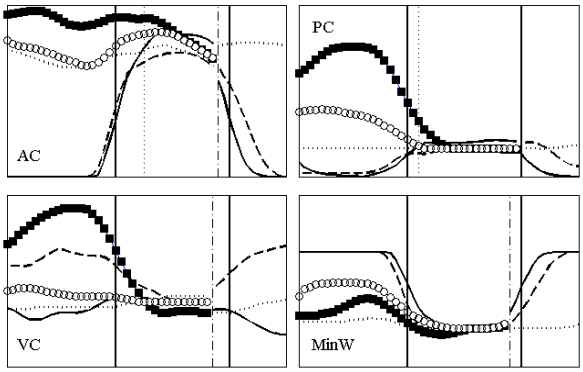


图 4 双音节/V1#sV2/中舌尖前擦音的动态腭位参数[2]（图中实线代表 V1/V2 为/a/，点线/ɿ/，杠线/u/，实心方块代表 V1 为/i/，空心圆圈代表 V1 为/ɥ/。两条纵向的黑实线表示前字元音结束点和擦音声学段结束点，点线表示前字元音的影响范围最远点，点杠线表示后字元音影响范围的最远点，p<0.05）

3.2. 双音节后字声母的动态腭位和声学分析

舌尖前擦音作为双音节后字声母的时候，前字韵母的韵尾有两种情况。我们首先考察前字韵母为单元音条件下后字声母的动态腭位和声学特点。所选语料的前字韵母为单元音/i, a, u, ɿ, ɥ/。单因素方差分析结果表明，舌尖前擦音的首帧腭位的参数只受前字韵母的影响（p<0.0001），而最后一帧腭位参数只受到后字韵母的显著影响（p<0.0001）。

从图 4 可以看出，舌尖前擦音的成阻始于前字韵母的尾渡，AC，COG 和 MinW 基本上同步变化，从前字韵母 3/4 时长处开始就指向/s/的发音目标。从 PC 和 VC 上看，舌尖前擦音的动作始于前字韵母 1/2 时长处之后。这个结果说明舌体动作的启动时间要早于舌尖动作的启动时间。擦音段前半部分受到前字韵母的影响，但是不同的腭位参数受影响的时间范围不同，一般不超过擦音段中点。AC 受到的影响最大，但是 COG 和 MinW 不受前字韵母的影响，说明舌尖前擦音的发音动作已经基本到位，只是由于受到前字韵母动作惰性的影响，在 AC 局部区域产生差异。PC 受影响的范围较小，从第二个时间点开始就保持稳定的接触，这说明舌页与舌尖动作的耦合关系相对稳固。VC 受前字元音的影响不显著。

擦音除阻点始于擦音段的尾渡，AC、VC 和 MinW 迅速向后接韵母的发音动作过渡，COG 也发生相应的变化，PC 基本上不受后接元音的影响。VC 的变化与单音节情况相似，但是受元音影响的时间范围小于单音节的情况。

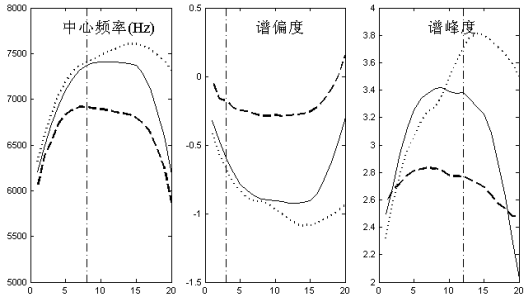


图 5 / V1#sV2/中舌尖前擦音的频谱统计参数

从图 5 可以看出，舌尖前擦音的频谱统计参数受到前字韵母的影响较小，在持阻阶段主要受后字圆唇元音的影响，中心频率下降，偏度趋近于零，峰度较低。接近擦音结束点的时候，/sɿ/的频谱参数基本不变，但是/sa/和/su/的变化同单音节情况一致。

[1] 此时的噪声源是由于声门不完全关闭产生的。

[2] COG 的曲线与 AC 的曲线非常相似，由于篇幅所限，省去 COG 的曲线。

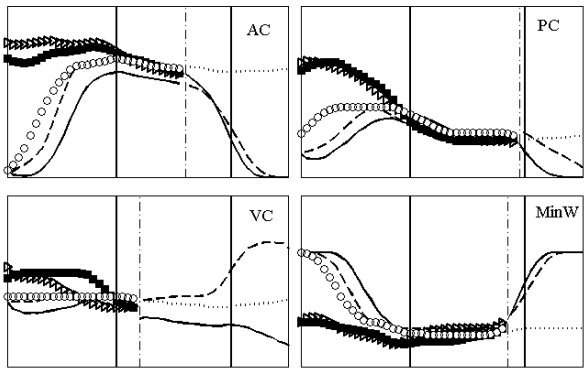


图 6 双音节/V1n#sV2/中舌尖前擦音的动态腭位参数（图中实线代表 V1/V2 为/a/，杠线/u/；实心方块代表 V1 为/i/，空心圆圈代表 V1 为/ə/，三角形代表 V1/y/。其他同图 4）

双音节的第二种情况是前字韵母为鼻韵母，鼻韵母又分为两种类型。我们先分析前鼻韵尾的情况，选取的前字韵母为/an, in, un, yn, ən/，分析方法与/V1#sV2/一样。前鼻韵尾由于受到后字舌尖前擦音的影响，鼻音的口腔闭合动作不完全，因此鼻音段的起点依据语图和鼻音的共振峰的走向而定。

从图 6 可以看出，舌尖前擦音的成阻过程与前鼻音的口腔发音动作部分重合，鼻音的口腔动作先于舌尖前擦音的成阻动作，但是鼻音口腔动作的目标却是指向擦音的动作。首先，不同 V1 的时长与鼻音段时长的比率基本符合已有的声学研究结果[24]。在生理方面，不同 V1 条件下的 AC 上升的时间起点不同。第二，在鼻音段的尾渡，AC 和 PC 不断下降，一直延伸到擦音段的前半部分，这个过程说明前鼻音的动作不断向舌尖前擦音的动作过渡。第三，MinW 的变化起点取决于 V1 的类型，且在鼻音段不为零。第四，鼻音结束帧和擦音的起点帧的腭位参数没有显著差异，表明鼻音的口腔动作的目标就是为了实现擦音的阻碍动作。对比/V1#sV2/中擦音段起点帧的腭位参数，可以发现前鼻韵尾条件下的舌尖前擦音的起点帧的 AC 和 PC 接近于/i#sV2/情况下擦音起点帧的情况（分别为 55%和 20%左右）。

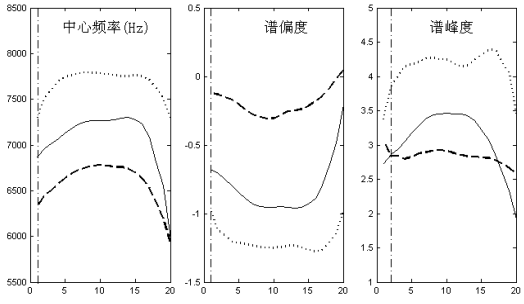


图 7 / V1n#sV2/中舌尖前擦音的频谱统计参数

腭位参数的变化对于解释/V1n#sV2/的协同发音相对有限，这是因为齿龈鼻音动作不仅涉及口腔的关闭动作，腭咽门（velopharyngeal port）也需要打开，从而形成鼻音。但是擦音需要较高的口腔内压，因此需要及时关闭腭咽门。从上面的分析来看，鼻音段 AC 和 PC 增大很有可能和口内压要求有关，接触面积增大导致收紧点后声腔体积的减小。有少量样本在接近擦音起点帧的时候形成完全的口腔阻塞。

在双音节/V1n#sV2/中，后字韵母对舌尖前擦音的影响与单音节和前字为单元音条件下的双音节情况不同，这主要表现为 AC 受语音环境影响的方式以及时间范围不同。与单音节的差异可能来自韵律边界的影响。在单音节条件下，擦音前是较长的静音段，因此舌尖前擦音处于较高的韵律边界上。此时擦音动作有所增强，后接元音对擦音发音动作的影响变小；而在双音节条件下，舌尖前擦音处于音步内，因而动作易于弱化，同时后接元音对舌尖前擦音的协同发音影响变大。与/V1#sV2/情况下的差别主要表现为擦音段前部。由于在/V1n#sV2/中擦音的动作从鼻音段就已经开始，因此前字韵母基本上不影响擦音的发音动作。

从图 7 可以看出，擦音频谱统计参数与前字韵母无关，从擦音起点就受到后接元音的影响。中心频率和谱偏度在首字音节声母交界点之前受后接元音的影响分别表现为/sɪ/>/sa/>/su/，/su/>/sa/>/sɪ/。在擦音结束点，/sa/和/su/的差异消失。/sɪ/的谱峰度在整个擦音段都大于/sa/和/su/，后两者则没有区别。

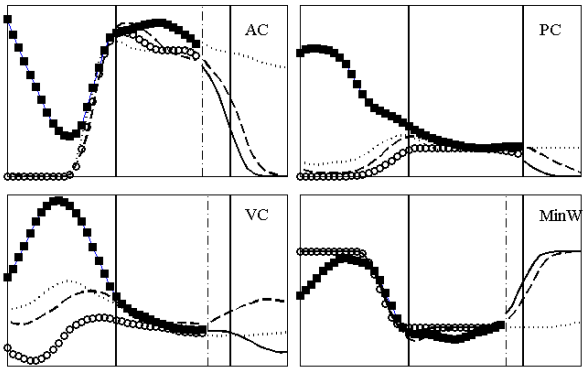


图 8 双音节/V1ŋ#sV2/中舌尖前擦音的动态腭位参数（图中实线代表 V1/V2 为/a/，杠线/u/；实心方块代表 V1 为/i/，空心圆圈代表 V1 为/ə/。其他同图 4）

图 8 是前字鼻韵尾为/ŋ/条件下的双音节腭位参数的动态变化。可以看出，舌尖前擦音的成阻始于前字韵母中点之后，表现为 AC、PC、COG 和 MinW 在前字韵母段中同步指向舌尖前擦音的发音动作。此时，舌尖前擦音的动作与前字韵母

动作产生时域上的重叠，这说明前字韵尾的发音动作与舌尖前擦音的主发音器官动作不冲突[25]。

擦音段腭位参数不受前字韵母的影响，这与/V1n#sV2/中的情况相似。后字韵母对舌尖前擦音的影响与/V1n#sV2/情况不同，AC、PC、COG 和 MinW 受到后接韵母影响的时间范围与单音节情况类似，说明擦音段主发音器官的动作非常稳定。后接韵母对 VC 的影响在声韵母边界前才表现出显著差异，可能的原因在于前字韵尾软腭成阻和除阻的动作速度较慢，从而影响了舌面向后接韵母动作的过渡。

同其他情况一样，擦音频谱统计参数只受到后接元音的影响，而与前字韵母没有关系（图 9）。中心频率受后接元音影响的时间范围较小与样本容量有关。实际上，在擦音起点，中心频率就表现出显著的差异。但是在擦音段中部，显著差异消失。谱斜率的结果与前面的结果一致。虽然谱峰度受元音影响的时间范围较小，但是从走向上看，与前面的结果也是一致的。

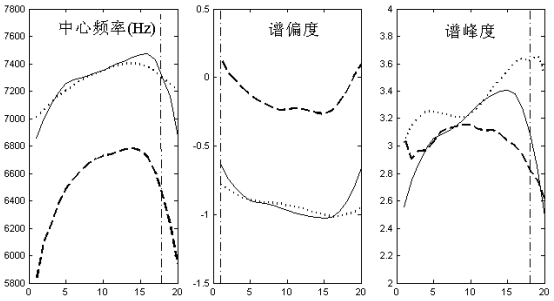


图 9 / V1ŋ#sV2/中舌尖前擦音的频谱统计参数

4. 结论

本文讨论了舌尖前擦音作为单音节声母和双音节后字声母的动态发音动作和频谱统计参数。实验结果表明：

1. 在单音节条件下，声母位置上的舌尖前擦音在齿龈和硬腭区域的接触十分稳定，而软腭区域的接触受到后接元音动作的影响。在擦音段的稳态段，频谱统计参数只与后接元音是否圆唇有关；接近声韵母交界点的时候，频谱统计参数随后接元音的类型而变化。

2. 双音节中擦音的成阻的起始方式和时间依据不同的前字韵母而不同。起始方式取决于前字韵母的发音部位，元音韵母条件下，舌体动作早于舌尖动作，舌尖成阻起点较晚；如果前字韵尾是前鼻音，发音动作与前鼻尾口腔动作融合，此时前字韵母中主元音和前鼻尾的动作相位关系不变，但是鼻音的口腔动作指向舌尖前擦音的动作目标；如果前字韵尾是后鼻音，舌尖动作与后鼻尾的软腭阻塞动作产生时域上的叠加。前字韵母对擦音段动作的影响的程度和时间范围很小，擦音段主发音器官动作和舌页的耦合动作相当稳定。后字元音对擦音动作的影响主要体现在舌体动作和声韵母交界点附近舌尖动作的变化。

3. 擦音段频谱不受前字韵母的影响，而与后字圆唇元音密切相关。此外，在声韵母交界点附近，擦音频谱与腭位参数的变化存在一定关系。

参考文献

[1] Heinz, G.W. On the properties of fricative consonants[J]. The Journal of Acoustical Society of America, 1961, 33(5): 589-596.

[2] Hoole, et al. A comparative investigation of coarticulation in fricative: Electropalatographic, electromagnetic and acoustic data[J]. Language and Speech, 1993. 36: 235-260.

[3] Hughes, G.W. et al. Spectral properties of fricative consonants[J]. The Journal of Acoustical Society of America,1956, 28: 303-310.

[4] Narayanan, S.S. et al. An articulatory study of fricative consonants using magnetic resonance imaging[J]. The Journal of Acoustical Society of America, 1995, 98: 1325-47.

[5] Nartey, J.N.A. On fricative phones and phonemes[R]. Working Papers in Phonetics, 1982, WPP 55.

[6] Subtelny, J.D. et al. Cineradiographic study of sibilants[J]. Folia Phon,1972, 24: 30-50.

[7] Shadle, C. et al. An MRI study of the effects of vowel context on fricatives[C]. Proc. IOA 18:9:1, 1996, 187-194.

[8] Shadle, C. et al. An MRI study of the effect of vowel context on English fricative[C]. Acoustics, 2008, Paris, 5101-6.

[9] Byrd, D. Articulatory timing in English consonant sequences[R]. UCLA Working Paper in Phonetics, 1994, 86.

[10] Recasens, D. et al. A model of lingual coarticulation based on articulatory constraints[J]. The Journal of Acoustical Society of America, 1997, 102: 544-561.

[11] Farnetani, E. V-C-V lingual coarticulation and its spatiotemporal domain[C]. In Hardcastle & Marchal (eds.) Speech Production and Speech Modelling. 1990. Kluwer Academic Publications: 93-130.

[12] Soli, S. D. Second formants in fricative: Acoustic consequences of fricative-vowel coarticulation[J]. The Journal of Acoustical Society of America, 1981, 70:976-984.

[13] Toda, M., Maeda, S., Honda, K. Formant-cavity affiliation in sibilant fricatives[C]. In Turbulent Sounds in Speech, Mouton de Gruyter, 2010: 1-33.

[14] Stevens, K. Acoustic Phonetics[M]. MIT Press, 1998.

[15] Ladefoged, P. et al. Places of articulation: An investigation of Pekingese fricatives and affricates[J]. Journal of Phonetics. 1984, 12: 267-278.

[16] 周殿福，吴宗济. 普通话发音图谱[M]. 北京：商务印书馆，1963.

[17] 李俭，郑玉玲. 汉语普通话发音的变化[J]. 浙江工商大学学报. 2006, 5: 35-40.

[18] 郑玉玲，刘佳.普通话辅音发音部位及约束研究[C]. 第七届中国语音学学术会议论文集，北京，2006.

[19] 吴宗济，林茂灿. 实验语音学概要[M]. 北京：高等教育出版社，1989.

[20] 许毅. 音节与音联[M] // 实验语音学概要，北京：高等教育出版社,1989.

[21] Hardcastle, W.J., et al. EPG data reduction methods and their implications for studies of lingual coarticulation[J]. Journal of Phonetics, 1991, 19: 251-266.

[22] Forrest, K. et al. Statistical analysis of word-initial obstruents: Preliminary data[J]. The Journal of Acoustical Society of America, 1988, 84: 115-23.

[23] Shadle, C.H. et al. Towards the spectral characteristics of fricative consonants[C]. Proc. ICPhS, Aix-en-Provence, 1991, v. 3: 42-45.

[24] 林茂灿，颜景助. 普通话带鼻尾零声母音节中的协同发音[J]. 应用声学，1992, 1: 12-20。

[25] 郑玉玲，刘佳. 普通话 N1C2(C#C)协同发音的声学模式[J]. 南京师范大学文学院学报，2005, 3: 150-157.

普通话语音多模态研究与多媒体教学

Multi-speech modality research in Putonghua and multimedia teaching

孔江平

香港大学香港普通话培训测试中心

北京大学中文系，汉语言学研究中心

摘要：本文根据北京大学中文系对汉语普通话进行的语音多模态研究，讨论了汉语普通话语音在多媒体教学系统中的作用和实际应用的可能性。文章对香港人学习普通话的特点及多媒体教学系统的制作也进行了初步的介绍，并展望了汉语普通话多媒体教学的前景。

Abstract: This paper introduces the multi-speech modality research of Putonghua done in the department of Chinese language and literature, Peking University and discusses the role and possibility in establishing the multimedia teaching system of Putonghua. The paper also discusses the property in Putonghua learning by Cantonese and the multimedia teaching system which is being established in Hong Kong University. Finally the paper presents the prospect of Putonghua multimedia teaching.

关键词 普通话；语音多模态研究；多媒体教学

Keywords: Putonghua; multi-speech modality research; multimedia teaching

1. 引言

随着中国力的增强和文化科技水平的提高，中国无论在经济方面还是在文化方面和国际的交流都在迅速发展和扩大，这致使汉语普通话的使用范围也迅速扩大。汉语是联合国的工作语言，在国际政治经济活动中有着重要的地位，汉语也承载着中国五千年的文明所创造的灿烂文化。普通话是汉语口语的标准，因此，对汉语普通话的学习必然会成为世界各国语言教育中的一个重要的方面。

语言的学习与其他科学文化知识的学习不同，有其自身的特点，这些特点对语言教学的十分重要，因为它直接影响到语言学习的成效。众所周知，人们在习得母语的过程中，没有人会感到是一种负担或者十分辛苦，无论学习者的智商如何，都能在母语的环境中自然掌握一种语言。然而，在第二语言或方言的学习方面，如果没有语言环境，学习过程都会感到十分辛苦，语言教学的这种特点只从经验很难解释。

在以往的教育中，对教学法的理解更多的偏重于经验，甚至认为是一种艺术，这些理念在教育中也确实起到了一定的作用。然而，随着科学的发展，特别是人类行为的科学研究和人类大脑科学研究的进展，越来越多的人已开始认识到教育的科学性质。同时，言语科学的发展，特别是言语脑科学的发展，也使人们看到了语言教学理论的方向和科学的曙光。

在语言科学研究中，语音学的地位十分重要和直接，因为它主要研究语言的物理特性和语言学意义之间的关系。语音学分为传统语音学和语音科学，虽然语音学已经完成了从传统到科学的转化，然而，传统的语音学然在语言学研究中广泛使用，只是研究的领域不同。本文将从语音科学的角度，探讨一下汉语普通话语音多模态研究和普通话语音教学的关系以及建立普通话语音教学多媒体系统科学基础。

2. 语音科学研究与教学

汉语普通话语音多模态研究主要涉及到语音声学研究、X 光动态发音动作研究、核磁共振声道研究、电子腭位发音动作研究、嗓音发声类型研究、唇形唇读研究、动态声门研究及呼吸带语音韵律研究几个方面，下面将从这几个方面讨论一下这些基础科学研究和普通话教学的关系以及应用的可能性。

2.1. X 光动态发音动作研究与教学

在传统语音学研究中，X 光的使用使语音学完

成了从传统到科学的第一次飞跃。因为在此之前，人们是靠观察和自我感觉确定发音的部位和动作，有了 X 光以后，人们可以从科学实证的角度研究发音器官、发音动作和音义的关系。在教学方面，由于 X 光的使用，人们更精确地了解了语音发音器官的运动方式（汪高武，鲍怀翘，孔江平，2008），大大地提高了语音教学的科学性，特别是推动了语音教学的方法和理论的产生，从此语音学理论进入了普通的语言教学。国内这方面的成果有《汉语普通话发音图谱》（周殿福，吴宗济，1963）、图谱教材等，都属于这一类。在我国，汉语普通话 X 光的动态资料只有鲍怀翘教授在文革后拍摄了 4 位发音人的资料，由于健康保证方面的原因，根据普通话声韵调的组合关系和轻声儿化等语音现象，每人拍摄了 200 多样本，其中大部分是单音节样本。由于信号处理技术条件的限制，直到最近这些样本才被北京大语言学实验室处理为可供研究和开发教学软件的数字化数据库。

2.2. 语音声学分析与教学

在二次大战后期，出现了频谱分析技术，这种技术可以实时实现语音信号时域到频率域的转换，对于语音学研究来说可谓又一次技术上的飞跃。这使人们对语音的性质有了质的认识，大大推动了语音学和语音教学的发展。在 19 世纪贝尔实验室出版了《可视语言》(Ralph K. Potter, 1947)一书，此书的出现极大地推动了声学语音学的发展，很快人们发现了语音声学特征和生理特征的关系，从而建立起了语音声学和生理学的基本关系以及基础理论。将语音变换成图像可以使听力有障碍者通过视觉读懂语音，为残疾人的语言教学提供了新的方法。在正常人语言的学习上，可以通过分析学习者的发音，将其转换为可以理解的发音形式，如，声调曲线可以用于教学中声调偏误纠正，通过提取共振峰画在元音图上，帮助纠正元音的学习等等(吴宗济 林茂灿主编，1989)。由于电脑速度和信号处理技术的进步，现在人们正在研究语音到生理发音器官动作的反推问题，这一问题如果能被彻底解决，自我纠正发音将会变得很容易，这将彻底改变语音学习和教学的基本模式。

2.3. 核磁共振声道研究与教学

核磁共振用于人体透视检查是近十年的事，一般认为少量拍摄对人体无害，主要用于医疗检测，核磁共振的原理与 X 光不同，对于语音学研究来说，发音器官的成像更为清晰，而且可以拍摄动态二维、静态三维和动态 2.5 维的图像，非常适合语音学的研究。目前，语音学、言语生理学和言语工程都在利用核磁共振技术进行言语科学最前沿的研究。北京大学中文系语言学实验室近几年在日本国家实验室(ATR)拍摄了 9 个人普通话基本发音动作的资料，并且已经建立起了一个全数字普通话发音数据库。基于核磁共振的普通话声道和发音动作的研究使我们对汉语普通话有了更深入的认识，目前正在努力进行汉语普通话三维立体的数字模型研究。在国际上，语音教学领域一般认为最高水平的教学系统应该是三维立体的数字发音模型，目前这样的商用教学系统英语也还没有做出来，只有瑞典语的一个博士生做了一个演示系统。我们目前正在研究制作基于二维的核磁共振普通话教学系统，这种系统可以让学习者直接看到发音部位和发音动作，提高普通话语音的学习和教学效率，而且可以帮助聋儿童学习普通话的基本发音。

2.4. 电子腭位发音动作研究与教学

动态电子腭位技术是将一个假腭放在口腔中，在发音过程中，舌头和上颚的接触可以实时地被记录下来。通过对这些数据的分析，可以研究普通话的发音动作，特别是普通话的协同发音。动态电子腭位最初的目的是为了腭裂儿童术后学习发音而设计的，这是因为腭裂儿童常常使用代偿性发音，在手术后仍然不会正确的发音，电子腭位可以用于纠正腭裂儿童的发音动作，主要是辅音的发音动作。可以看出，电子腭位主要用于研究辅音的发音动作和辅音的协同发音，在普通话的教学显然有重要的应用前景。我们这几年建立了一个很大的普通话电子腭位数据库，正在做基础研究和腭裂儿童的语言康复的应用研究，如果要应用在普通话的教学方面，除了基础研究以外，其瓶颈主要在技术的应用转化和成本方面，但理论上已经成熟。

2.5. 唇形研究与教学

从语言交际的角度看，唇形和语言的学习及教学有着密切的关系，语言学习过程中的模仿始终都离不开唇形的信息。在聋哑儿童的唇读学习中，唇形的作用更是必不可少。从语言学的角度，唇形的研究主要注重唇位的研究，即唇形的语言学意义。在言语工程研究上，往往用图位来定义唇型，更注重唇形产生的模型和模拟。这两方面的成果为语言的教学提供了可能和坚实的基础。在唇位研究方面，

主要可以从视频中检测出唇形的基本信息和动态信息进行研究。近几年，由于技术的发展，唇型检测又有了专门的设备红外三维立体唇型信号采集系统。目前我们建立了一个基于红外三维立体信号的汉语普通话唇型基础研究和教学应用研究的大型数据库，希望建立一个用于汉语普通话教学系统的基础模型，基于这个模型，开发具有唇型视频的普通话语音多媒体语音教学系统。

2.6. 动态声门研究与教学

动态声门研究是基于高速数字成像技术对声带振动频率和振动方式的研究，具体是利用高速摄像机（4000 帧/秒）拍下声带的振动过程，然后用图像处理技术检测出相关参数对声带的语言学特性进行研究（Kong J.P., 2004，2007)。如对汉语普通话声带振动的研究表明，汉语普通话的上声的低音部分是一种挤喉音的发声类型，这说明普通话上声不仅靠相对低的音调来体现声调的语言学意义，同时，特殊的发声类型是学习普通话上声的关键。许多外国人只是从音调或旋律上来模仿上声，没有意识到发声类型的不同，因此学习效果很差。在对汉语普通话发声类型的研究中发现，上声的发声在生理和声学上有五六种类型和表现形式。近期的研究也证明了汉语普通话发声类型对声调的感知有一定的贡献，特别在自然度方面更是如此。动态声门要真正用于普通话的教学就要进一步作发声类型视觉反馈的研究，即从语音中算出声带振动的方式，给学习者声带发声的视觉反馈，这样就能更有效地通过教学系统学习。

2.7. 呼吸带语音韵律研究与教学

汉语普通话的语音率是比较复杂的，这是因为普通话除了基本的语调外还有声调，这两种语音学现象在物理上主要通过基频变化来实现，因此这对普通话韵律的研究带来了很大的困难，以往的研究主要是通过语音来提取基频、时长和振幅等参数，这些参数对韵律和声调的性质不能进行充分的解释，因而长期制约了普通话韵律的研究。通过呼吸带同时采集语音和呼吸信号，可以很好地研究韵律以及韵律和声调的关系。目前我们已经建立了一个大型的汉语普通话呼吸-语音数据库，研究表明，呼吸在律诗、散文、新闻等不同文体中都有不同的表现和固定的模式（谭晶晶等，2008）。而且，通过呼吸和语音声学的研究发现，呼吸信号更能表现出朗读的个人风格和特点。这就为汉语普通话学习系统提供了更多的可应用功能。可以看出，普通话呼吸的研究可以为普通话朗读教学和普通话艺术教学提供广阔的应用前景。

2.8. 语音感知范畴研究与教学

我们都知道语音学研究对语音的教学很重要，但目前在教学中运用的语音学知识实际上是很片面和很陈旧的。如在普通话的语音教学中，通常只用到发音部位和发音方法一些基本的语音学知识，而且有些还不准确，这常常体现为一个很普通的语音学问题会有很多人在那里反复争论，不断强调自己的感觉，而不是去用科学的方法证明。然而，普通话语音研究已经从声学和生理开始进入语音感知领域，恐怕很少有普通话的老师了解普通话各个音位之间感知关系。如，在教“上声”发音时，只注重音调频率的变化，而对“上声”发声类型不了解，然而上声的感知和发声类型密切相关。最近的普通话声调感知研究表明，发声类型对汉语声调的感知范畴和自然度都有较大的贡献。因此，在普通话的教学中，将新的语言学知识特别是范畴感知和连续感知等基础知识引入教学是十分必要的。

2.9. 语音识别与测试

语音识别可以分为基础的研究和工程应用两种。基础的究主要注重声学语音学的研究，即研究语音的基本参数和语言学意义的关系，进而用于语音的判断和识别。而工程应用的主要在语音识别系统的技术实现上，在声学层面一般只用美倒谱系数，在识别层面主要用统计模型-隐马尔科夫模型和某种定义的语言学模型。在中国做基础研究识别的人很少，国家的投入和成果都较少，而做工程语音识别的人很多，国家投入很大，目前也有很多语音识别系统。在语音教学和测试方面，两种语音识别都能有广泛的应用前景。基础的识别研究元音共振峰参数和普通话音位系统以及韵母的关系、基频和声调语调的关系、辅音声学参数和辅音音位等，最终希望能从语音的声学参数识别出具体的语音音位，这些研究的成果在教学方面可以应用在语音声学视觉反推上，甚至能应用在语音生理的视觉法推上，如发一个元音就能在一个坐标上画出这个元音的相对位置，或者画出这个元音的声道形状，这样就可以在学习发音时进行自我矫正，从而达到高效自我学习的目的。聋哑人和言语障碍者也可以通过这种系统进行语言学习和语言矫正，其应用前景极其广阔。用于工程的识别在自然语言的识别上已经有了

很大进展，某些领域已尝试应用，但离实际的应用还有不少差距。在语音的测试方面，由于这种识别主要是统计模型，不太注重语音的语言学意义，因而在测试的应用上也还有一很大的困难。从目前的研究水平看，建立一个单音节的汉语普通话语音测试系统因该是可以做到的，这主要取决怎样建立训练识别系统的语音数据库和建立怎样的语音学模型。

3. 香港普通话语音多媒体教学

在第二语言或方言的学习中，语音的学习有着自身的特点，因为它涉及到两种语言音位感知系统的冲突。如果母语和所学语言或方言的音位感知系统冲突很大，学习困难就会很大，也会导致学习的效果很差，这和人的智力没有关系。例如，印度人学英语很难学会送气音，日本人学英语很难学会弱化音节，中国人学英语很难分辨清浊辅音，这些都是因为音位系统的冲突造成的。在方言的语音习得过程中，常常会受语音历史音变规律的影响，香港人在学习普通话过程中，最大的障碍就是会受到语音历史音变的影响，从而导致语音类推的失误。众所周知，汉语从中古音到各方言有一定的演变规律，可以进行语音的类推，但这种规律复杂到一定程度，类推就会常常出错。因此，在制作香港人学习普通话的教学系统时，一定要理论上考虑汉语历史音变的规律和音变规律在学习过程中的误用。

在对语言接触和语言融合的研究中发现，两种语言感知的冲突在自然语言接触和融合过程中是有规律的（陈保亚，1996）。这种规律反映了语言感知系统对冲突的自然处理方式，这种处理包括生理的、语言学的和大脑认知的，对其研究会为我们编写普通话语音教材和教学提供依据。在目前的汉语普通话对外教学和方言的教学中，人们都在研究偏误，无论经验的、声学的和生理的都对普通话的教学起了积极的作用。然而从语言学理论的角度看，偏误的内涵在大脑音位系统的冲突所反映出来的语音感知结构的差异，这方面的研究需要脑电和功能性核磁共振等高端的仪器和设备，目前在国内研究的还不多，正在起步阶段。对于香港普通话语音多媒体教学系统的研制来说，首先要搞清楚香港话和普通话在语音感知体系上的差别，才能研制出适合香港人学习的汉语普通话语音多媒体教学系统。

4. 普通话多媒体教学的展望

任何一种语言的学习都需要进行大量的操练，这是语音学习的特点。一般情况下这些操练都是在课堂上进行的，然而，课堂操练的时间是相当有限的，这就形成了语言学习和教学上的一个矛盾。从现在的语音学研究看，语音多媒体教学系统可以较好地解决这个矛盾，因为，多媒体学习系统可以不受时间和地点的限制，同时能具有音频、视频甚至判错等功能，这就为语音教学的自我学习创造了条件，使多媒体教学具有很大的优势。

在国际上，各国研究和制作语音多媒体的教学软件已经成为语音学科学研究的一个重要的方面，而且已经有了较好的系统。在国内，基于生理和物理的基础研究也已经逐渐展开，我国的语音学研究和言语科学技术也已经十分先进，但在语音大脑认知方面的研究才刚刚起步，这就阻碍了高端多媒体系统的研发。另一方面，在语音多媒体教学系统的研制方面，大家的认识有很大的差别。目前国际国内都有两种方式，一种是从语音的发音发声原理和语言学理论方面来进行语言多媒体系统的研究和制作。另一种是从技术的角度研究和制作教学系统。香港普通话培训测试中心和北京大学中文系正在联合研制的属于第一种方法，试图将基础研究应用于教学系统的制作。

随着中国国力的不断提高，汉语普通话在国际政治、经济和日常生活中的地位会不断提高，汉语普通话的教学势必越来越重要。然而，语言是有民族性的，这就要求我们不仅要对汉语进行大量深入的语音学和语言学的研究，而且也要对汉语普通话的教学进行大量的研究。为了能使汉语普通话走向世界，研究制作汉语普通话的语音多媒体教学系统是将汉语普通话推向世界的重要一环。随着汉语语音学、生理语音学、声学语音学、心里语音学、语音认知研究以及言语科学技术的发展，汉语普通话的多媒体教学系统也将会不断完善。语音脑科学的研究使我们正在不断接近普通话音位感知系统的本质，这将会对普通话的教学方式产生根本的改变。然而，我们也十分清楚，在普通话的教学中，我们只知道汉语普通话语音的认知体系是完全不够的，我们还必须研究学习者母语的语音认知体系以及生理和物理的特性，只有这样才能研发出真对世界不同语言的汉语普通话语音多媒体教学系统和汉语普通话水平测试系统。

5. 参考文献

1. Ralph K. Potter,George A. Kopp,Harriet C.Green,1947, Visible Speech, D. Van Nostrand Co., New York

2. JP Kong (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.

3. JP Kong (2004), Phonation Patterns of Tone and Diatone in Mandarin, From Traditional Phonology to Modern Speech Processing, Foreign Language Teaching & Research Press, edited by G. Fant et al.

4. JP Kong (2007), Laryngeal Dynamics and Physiological Model, Publishing house of Peking University

5. 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

6. 周殿福，吴宗济，1963,《普通话发音图谱》，商务印书馆

7. 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

8. 理论陈保亚，1996,《论语言接触与语言联盟》，语文出版社

9. 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

10. 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

普通话在香港

汪锋、孔江平

提要：本报告在前期研究的基础上，进一步分析普通话在香港的发展，从不同角度来探讨香港多语系统中普通话角色的定位；分析普通话推广与三语共存的现状，重点考察普通话当前在不同领域的角色和普通话教学与研究在香港的现状；探讨普通话教学与推广跟现代科学技术的结合及前景。本报告集中探讨在普通话发展的大背景下，如何处理好与香港社会现存的其他语言的关系；在恰当定位的前提下，探讨在现代科技背景和社会条件下普通话在香港面临的机遇和挑战，现代科学技术如何与普通话教学相结合来改善普通话的应用。

1. 普通话在香港语言生活中的位置

香港政府设有法定语文事务部，“中文和英文都是香港的法定语文。政府向公众发表的主要文件，均备有中英文本。政府与市民的书信往来，则因应对象选用中文或英文。立法会会议和政府举行的各类会议，会视乎需要而提供英语、广东话、普通话的即时传译服务。”值得注意的是，其中专门设有“普通话支援服务组”。

香港政府统计显示按惯用语言划分的人口比例如下：

	1991		1996		2001	
	数目	百分比	数目	百分比	数目	百分比
广州话	4 583 322	88.7	5 196 240	88.7	5 726 972	89.2
普通话	57 577	1.1	65 892	1.1	55 410	0.9
其他中国方言	364 694	7.0	340 222	5.8	352 562	5.5
英语	114 084	2.2	184 308	3.1	203 598	3.2
其他	49 232	1.0	73 879	1.3	79 197	1.2
总计	5 168 909	100.0	5 860 541	100.0	6 417 739	100.0

各种语言在香港语言生活中的实际地位可以通过香港特区政府的另一表述体现出来：“在政府部门以及法律界、专业人士和商界之中，英文是广泛采用的语文。在许多于香港营商或到内地和台湾经商的企业中，精通英语、广东话和普通话的三语人才都担任重要职位。”

普通话随着香港回归祖国而大幅迈进香港社会，进而形成了现在大多数香港公共服务机构采用的“两文三语”。所谓的两文是指书写上的英文和中文，三语指的是口头表达上的英语、粤语和普通话。严格来说，“两文三语”这一概括需要进一步明确，因为中文在文字形体上有繁体和简体之分，而从所记录的语言性质可以有粤语和普通话之分。

口语层面使用更多的粤语和英语，普通话所占比率很小。从另一个角度来说，这一概括还忽略了香港的其他语言变体，比如，很多人说客家话等汉语方言。因此，普通话是在多语的香港社会中寻求适当的位置，即如何顺应社会发展的规律，与其他语言在书面和口头两个方面和谐共处。

截至 2009 年，香港回归已经 12 年了。由于普通话的强势介入，导致原有语言格局的变化，而尤其造成了教学领域的一个争议性问题，即，是采用中文(粤语或汉语)教学还是用英语教学。即使采用中文授课，用粤语还是用普通话也是一个问题。2009 年前，香港的 601 所小学、527 所日间中学及 61 所特殊学校，其中 114 所规定使用英语，其余的学校都采用粤语教学。而在 1997 年香港回归以前，香港政府政策规定所有学校的官方语言都是英语。转到粤语教学的状态显然是语言生活受政治大势影响的明显标志，也就是，学校的语言生活已经从英语状态转向中文状态。另一个值得关注的变化是，从 2006 年起，香港开始引入小学生普通话水平考试，逐步加强普通话教育，这显示了在中文状态下普通话与粤语的竞争。这一状态在中国内地普遍经历过，即普通话与各地方方言或语言在学校生活中的竞争，一个普遍的结果是普通话最终战胜地方方言成为教

学语言，但地方方言或语言作为学校生活中当地学生沟通的方式仍然普遍存在，即，课上普通话，课下本地话。因此，普通话最后与地方语言之间形成了分层共处的局面。至于，粤语与普通话在香港的竞争究竟会以什么样的步伐达到怎样的结果还比较难以预料，但通过分层使用而和谐共处似乎是最理想的结果。

受普通话的冲击，英语教育在香港学校中的分量急剧减少，但这造成了非预期的结果：占少数的英文学校成了稀缺资源，并在民众心目中成为尖端和先进的代表。同时，据香港的一项调查显示，目前只有 40%的小学六年级学生能够讲英语。很多人认识到英语教育的大幅减弱将关系到香港的国际金融地位，或许会成为被北京或者上海取而代之的重要因素之一。2009 年，香港教育局开始调整中学教学语言，2 月 28 日，教育局长孙明扬与中学校长座谈，提出：“香港是国家唯一明文规定中英文同样是法定语文的地区。国家亦冀望香港担当国际金融中心的角色。所以，我们的学生既要认识民族传统，又要具备国际视野，达到中英兼擅，日后才能够为社会、为国家做出贡献。”教育局开始强调“母语教学、中英兼擅”是正确的政策目标。新的教育政策不再把中学硬性分为“中中”和“英中”，目前小学主要采用母语教学，而高中和专上教育则以英语教学为主。政府将会在未来 3 年间投入 9.8 亿港元来实施“新教学语言政策”。新政策通过多次媒体见面宣传以及召开民众咨询会等方式进行磨合。行政会议在 5 月 26 日通过让学校根据“学生为本、因材施教”的原则，采取不同模式的多元教学语言安排。5 月 29 日教育局正式宣布:“学校可以在 2010/11 学年的中一开始实施微调中学教学语言的安排，让学校以学生为本，加强校内的英语环境，增加学生运用和接触英语的机会。”之后，教育局安排多次会谈与演讲，协调与学校、家长对新政策的认识与落实。不少学校积极响应香港政府所要实施的关于加强对英语教学的方针，但应注意到这一政策的目的是实现教学语言的多元化。因此，香港教育工作者联会对此政策提出异议，担心实际会造成英语再度压制母语的情形。

在实际生活中，香港的商业语言向普通话倾斜，民众学习普通话趋于积极。很多生活现象中可以发现普通话的加入，譬如：政府网站同时使用繁体、简体中文和英文三个版本；各种公共场所的指示大都以繁体、简体中文和英文标明；火车、汽车的报站除了粤语、英语之外，也增加了普通话。随着内地的发展以及与香港联系的加强，尤其是实施“自由行”以来，内地访港游客大量增加，为了与内地来港顾客沟通，大量的商品要求雇员使用普通话或者直接聘用普通话者，这就造成了大量的日常商铺间普通话的流行，但在高端商业领域，英文的地位仍然没有改变，因为在该领域沟通交流的各方是国际性的，而不是香港对内地的关系。

主动学习普通话的香港人士无论从数量还是从职业范围来看都有增加，以香港中文大学普通话教研及发展中心为例，每年报读普通话课程的人数几乎都有两位数增长，学员来自各行各业。

综合来看，普通话是香港多语社会中的新生力量，无论在政府、学校还是日常生活领域，都越来越占据相当重要的分量，但英文和粤语也同样是香港语言生活中十分重要的组成部分，连同其他语言，共同构成了香港语言生活的多维景象。

2. 普通话推广及教研在香港的现状

香港中文大学普通话教育研究及发展中心何伟杰在总结普通话在香港的发展历程时，将复兴期定为“约 1998 – 2008”，其间，普通话科正式成为中小学的核心课程和中学会考独立科目，是香港普通话教育发展的重要里程碑。政府投放大量资源，推动师资培训，设立进修基金资助港人学习普通话；语文教育及研究常务委员会（简称语常会）资助教师进修、举办社会推普活动、支援普教中等措施，为香港普通话教育发展踏上更高台阶创造了条件。

香港政府倡导“两文三语”，在政策层面设定了普通话推广的基础。 事实上，早在 20 世纪 80 年代香港教育部门就已经开始推广普通话教学工作，将普通话科列为中小学的独立科目。目前的问题是，如何根据香港语言生活的需求情况因地制宜地稳步推进普通话。

随着普通话需求领域和人群的扩大，学校设置的普通话课程和科目就显得不够，而且学生缺少语言环境，教师力量不够，推普活动多样性不够等问题开始凸显。针对这些问题，香港社会通过与国家语言工作委员会、香港语常会、香港各大学以及民间机构的努力来进行针对性的工作。2009 年度的活动分述如下：

2.1. 教育局活动

在香港，“用普通话教中文”开始成为一种趋势，相关配套措施，如：教师进修问题、课程设置问题等，都开始逐步得到加强。2009 年 6 月 6 日香港中文大学举办了“普通话教中文研讨会”，与会者对普教中的成效予以肯定，同意建议教育局能根据香港长远发展的需要，尽快制定中小学逐步实施普教中的时间表，同时为普教中的师资培训制定相关的发展计划。至于这一趋势是否应该或者何时推广到幼儿园，目前尚未提上日程。

香港教育局语文教育支援组 2008 年开始引进内地专家增强对“普教中”学校的支援，此举无疑会直接推动普通话的教学力度，但教育局于 2009 年开始“微调中学教学语言”，引起香港教育工作者的担忧，认为可能极大影响母语教学，2009 年 7 月香港教育工作者联会致信立法会教育事务委员会，提出自己的意见：倾斜资源支持母语教学；教学语言与收生脱勾；延长教师进修宽限期；研究提升英语教学效能等。

香港教育局于 2009 年在中学实施新高中课程，中国语文科设“普通话与表演艺术”及“普通话传意和应用”两个选修单元，此举也意在加强普通话教育。教育局语文教学支援组举办“协助香港中、小学推行以普通话教授中国语文科计划”分享会，活动在培侨小学(小西湾) 及福建中学(小西湾)举行，专题是“课堂观摩及校本经验分享：说写结合”。

推广普通话，不单是教授语言，通过各种交流活动，可以促进香港与内地的文化沟通，从而潜移默化的设置普通话的背景支持。香港教育局资助实施“同根同心”-香港初中及高小学生内地交流计划（2009），作为“薪火相传”国民教育活动系列的部分。获教育局认可的非政府机构承办共十二个内地交流行程，每一行程为期 2-3 天，交流地点为广东省珠江三角洲。计划以高小（小四至小六）及初中（中一至中三）学生为对象，旨在让学生从历史、文化、经济、教育、艺术、建筑、军事、民间信仰等方面，有系统地加深对国家的认识，亲身体会粤港两地同根同心之紧密关系。

2.2. 香港高校普通话培训测试中心

2009 年 8 月 28 日，国家语言文字工作委员会普通话培训测试中心与香港教育学院在香港签订“第五次合作协议”。截至 2009 年，香港已经有 10 所高校设有普通话培训测试中心。各中心积极开展各种活动推动普通话的教育、测试及出版工作。除定期举办普通话水平测试、考试辅导课程外，还举办普通话教学及测试讲座，提供网上咨询和普通话资料等。

香港大学普通话培训测试中心于 1996 年 4 月成立，率先于 1996 年与国家语言文字工作委员会合作，在香港定期举行普通话水平测试。2009 年度提供 30 个课程，约有 400 以香港教师为主体的听众参加。课程主要包括教师培训（教育专业普通话证书、普通话诊断及正音课程 、暑期普通话沉浸课程）、普通话水平测试辅导、外国人汉语培训(基础汉语、中级汉语、高级汉语)。

2009 年度以香港中文大学普通话教育研究及发展中心最为活跃，推广普通话教研的力度也最大。1 月，中国国家话剧院著名演员来港演出，中心向普通话课程的学员推荐他们主演的中外名家名篇大型朗诵会。9 月 9 日，林建平副主任应教育局的邀请，为“协助香港中小学推行以普通话教授中国语文科”计划的“教师专业培训课程”，给中小学教师与内地教师主讲《普教中课程规划与调适》；10 月 24 日，为中学普通话科教师主讲“中学普通话评估系列”，讲题为《中学普通话科听说教学与评估》。11 月 20 日，中心应仁爱堂田家炳中学的邀请，为“普教中”的中文教师进行课程观摩及讲评活动。11 月 21 日，中心举办了第十三次“普通话教中文系列专题讲座”，讲座的主题是《普教中阅读教学的实践》。12 月 16 日，刘遵义校长代表中心跟国家语言文字工作委员会普通话培训测试中心续签了合作协议。11 年来，近 2 万人参加了香港中文大学普通话教育研究及发展中心举办的普通话水平测试。

2.3. 香港语文教育及研究常务委员会

香港语文教育及研究常务委员会（語常会）成立于 1996 年，就一般語文教育事宜及語文基金的运用，向政府提出建议。2008 年开始拨款二亿港元推行“以普通话教授中国语文科计划 ”，通过邀请内地专家及本地顾问到校支援，协助学校推行“普教中”。计划分四期推行，每期支援三十所小学及十所中学，每所参与学校将接受为期三年的协助，提供

中文科教师专业培训，并安排两地交流探访。语常会方面表示，每期均有一百多间学校报名，竞逐四十个名额。每所参与学校将接受为期 3 年的协助，第一年由内地专家及本地顾问到校协助策划，第二年及第三年则由本地顾问协助普通话教学。

语常会为鼓励大众多接触和多应用普通话，以提升普通话的水平。语常会自 2002 年起每年都举办大型的推广普通话活动。2009 年以“活学活用普通话”为主题，推出 7 项学校及社区活动，包括：与香港电台普通话台合办普通话电台广播剧训练、粤港澳高校辩论赛和全港普通话广播主持选拔赛(公开组和中学组)；与香港教育专业人员协会合办全港中小学普通话歌唱比赛；与新市镇文化教育协会合办全港中小学普通话演讲比赛；与香港理工大学香港专上学院合办“香港任我行”——学生普通话旅游大使培育计划；与新域剧团合办到校普通话戏剧训练计划；以及由新城电台制作《财经点滴》电台节目。共有 7,000 多名学生及市民参与这些活动。语常会还聘请艺人方力申为语常会 2009 年度推广普通话大使，方力申出席多项推广普通话活动，利用自己的影响，鼓励香港民众多听、多说普通话。语常会的这些活动将普通话的推广应用到实际生活层面，从课本语言中跳脱出来，对于普通话植根香港社会有进一步的推动。

2.4. 民间机构的推普

香港的一些民间机构也积极推广普通话，其作用功不可没。例如，创办于 1976 年的香港普通话研习社是一个非赢利民间慈善团体，其宗旨是在香港推广普通话。30 多年来，普通话研习社共培训 30 万人说普通话。研习社成立了各种各样的兴趣小组，包括：演讲组、畅谈组、青苗组、合唱组、乐韵组、读书组。基督教浸会也逐步引进用普通话唱圣诗，牧师用普通话讲道时不用粤语传译等措施。

3. 普通话教研

3.1. 大学测试中心

香港中文大学普通话教育研究及发展中心面向新世纪的语文教育开办普通话教育文学硕士学位课程，培养普通话骨干教师和推普专才，增强学员的科研能力，以迎合学校和社会的需要。2009 年中心举办多次普通话教育讲座，讲者包括香港、北京、广州、厦门等地的知名专家，议题涉及“普通话科教师专业发展”、“语文教师的专业进修”、“汉语拼音教学与普通话教中文”等。6 月 6 日，中心举办了“普通话教中文研讨会”。研讨会的主题是“教学语言的规划与实践——香港普通话教学的现况与发展”。二十多位专家学者就普教中的教学语言规划、师资培训、课堂实践、教学资源发展、语文能力评估、跨课程的语言策略等各方面的问题发表演讲，并跟与会者进行了深入广泛的讨论，有本地及内地专家、学者、校长、专业老师等二百多人参加了此次会议。10 月 17 日，“普通话教育文学硕士学位课程”召开检讨会议。11 月 19 日，何伟杰主任等代表中心前往深圳第二高级中学作交流访问。11 月 21 日，中心举办了“名家名师谈对外汉语教学专题讲座”，北京语言大学杨惠元教授、北京大学李晓琪教授作学术报告，并与中心硕士研究生座谈，就汉语教学和普通话教学进行了深入的讨论。

香港大学普通话培训测试中心孔江平教授在《第四届普通话水准测试学术研讨会》发表题为“普通话语音多模态研究与多媒体教学”的讲演。其研究团队开展了我针对香港语言环境的普通话教研，包括：“普通话语音发音动作教学演示（粤语版）”，孔江平、鲍怀翘、汪高武、李永宏、何向真；“普粤中古语音对应字汇”， 孔江平、方芳、张雯、黄雪妍、潘晓声、姚文礼；普通话者对粤语声调的范畴感知研究，孔江平，吴君如，姚文礼。

3.2. 现代科学技术背景下的普通话教学研究

语言学习有其自身的特点，言语科学研究表明，语言教育应基于科学的方法，并涉及到语音学、语言学、言语声学、生理学、心理学以及脑科学多个领域。而目前普通话在香港的教学研究比较局限于传统的探讨，比如，如何通过香港文化与内地文化的差异来分辨普通话词汇与香港词汇的不同；如何讲解普通话发音之于香港人的难点；某些普通话词语、句式如何使用，等等。

现代科学技术其实已经在普通话语音多模态方面有很多研究成果，包括基于 X 光和核磁共振成像的发音器官运动方式、基于电子腭位元的普通话声母发音方法、基于视频图像的唇形运动、基于呼吸信号的朗读风格以及基于声带高速成像和范畴感知

的声调本质。通过这些言语科学的研究成果，有望发展形成普通话语音多媒体教学的基本理念和一个针对粤方言者学习普通话语音的多媒体教学系统。

在香港大学普通话培训测试中心，普通话文本数据库和普通话-粤语-中古汉语对应字汇数据库正在建立，这些数据库的建成会为科学系统的推进粤语背景下的普通话学习提供有力的支撑。同时，也建立了粤语者的普通话偏误言语数据库(Mispronounced speech database of Putonghua Assessment by Cantonese speakers)，这可以促进对粤语者在普通话偏误方面进行有针对性的全面研究，以提出明确有效的解决方案。另外，普通话词汇言语数据库(Speech database of Mandarin words)也在该中心完成，为教师提供丰富系统的普通话语音教学材料。

香港中文大学语言工程实验室在著名语言学家王士元教授的带领下，集中力量运用脑波仪来研究语言和大脑的关系，探讨不同语言在大脑处理机制上的异同，目前的成果主要有普通话者和粤语者在声调的范畴感知上的异同；普通话者和粤语者对繁简体中文的识别模式的异同，等等。这些研究从脑科学的角度为普通话在香港的习得提供了很重要的参考，并进一步将普通话在香港的教研推向深入。

4. 总结

香港社会语言生活的情况可分为自然驱动与人为控制两种，商界、日常生活等领域的语言选择以自然驱动为主，人们可以根据交际的需要以及交际对象等因素自然决定语言的选用，而普通话的推广力度和定位可以随之自然确定。

普通话推广的难点在于人为控制层面，如法定语言和教学语言。法定语言已经规定了所谓“两文三语”，但法定语言的实施相对来说范围比较窄，影响更大的是教学语言层面。在教学语言层面，英文与中文之间的关系仍然是纠葛各界的一个难题，理想的目标是中英文俱佳，但如何实现这一目标却不是短时间内可以解决的，无论从资金投入，还是教研力量的配备，都显得有待完善。而要到达中文(普通话)与英文俱佳如果单纯从学校开始恐怕并不足够，现代语言学的经验告诉我们，儿童在学前掌握三种口语是完全能够胜任的，因此，将来的研究和措施或许应该提前到儿童学语时，家庭、社会、政府通力合作，造就一个和谐的香港语言生活社会，而不是普通话、粤语或者英语中的任何一种占据压倒性优势。

如果说人为控制层面还需要谨慎探讨，那技术层面加强对普通话教育的力度却无需太多争议，随着计算机技术以及各种科学手段的更新，社会应该利用现代技术和现代语言理论的成果来为民众提供方便快捷的习得普通话的途径，至于民众是否选择这一途径，当然视需要而定。再进一步说，只要有需要，就应发展更好的普通话习得方式，因此提供多维度的普通话学习系统应该是努力的方向。

新平彝语松紧嗓音研究

叶泽华

摘要：本文运用电子声门仪 EGG 信号定义基频、开商和速度商三个嗓音参数，基于这三个嗓音参数建立新平彝语单音节发声模式，并利用这三个嗓音参数区分新平彝语各个调类。松紧调类的基频参数差别不大，松音调类的开商参数明显高于紧音参数，速度商参数明显低于紧音参数。

关键词： 电子声门仪 基频 开商 速度商 调质

1. 绪论

1.1. 电子声门仪工作原理及信号计算方法

1.1.1. 电子声门仪工作原理

电子声门仪所是一种通过非侵入性方式获得嗓音信号的重要工具（Childers, 1986; Fourcin, 1971, 1974），在言语嗓音研究及病理嗓音研究领域有着重要作用，最早由 Fabre（Fabre 1957）提出。它通过测量经过声带区域的电流来间接反映声带的运动状态，同时具有操作简单，对人体无害的优点。

电子声门仪通过释放稳定的微量高频电流（300kHz 以上，10mA 以下）通过喉部来记录喉部声带开合的情况。当电流通过人体时，由于人体是导电体，所以当声带闭合时，其电阻会大大减小，导致输出电信号增强，声带继续靠拢，电阻会继续减小，输出信号持续增强；而当两片声带分开时，会导致它们之间充满导电性能较差的空气，从而使得电阻增大，输出电信号减弱。而我们根据这个性质，也就间接地掌握了声带运动的信息。图 1.1（Baken，1992）为电子声门仪工作原理简图。图中左右两侧的黑色长条标识的就是电子声门仪两个电极片所在的位置，灰色区域表示整个电流经过的区域，中间的黑色区域显示声门所在的位置。

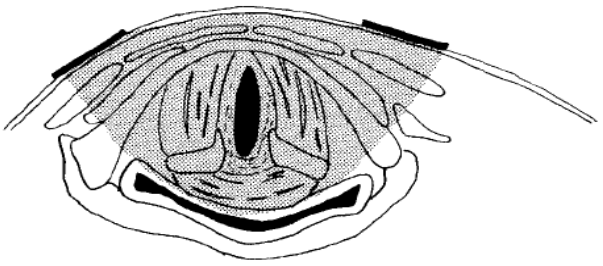


图 1.1 EGG 工作原理简图

研究证明，EGG 信号与声带接触面积呈正相关关系（Baken，1992）：当声带接触面积增大时，EGG 信号增强；声带接触面积减小时，EGG 信号减弱。关于这一点，已经被如高速声门摄影（Fourcin, 1974；Baer T et al，1983； Gilbert HRet al，1984）等许多同步的技术手段所证明，所以 EGG 信号被广泛用于言语发声和病理嗓音相关的研究。但是利用 EGG 信号来研究言语嗓音也有其不足之处。首先是信号采集问题，不是每个人都能用电子声门仪采集到很好的信号，如女性的 EGG 信号一般都会比男性要差一些，较胖的人的 EGG 信号也往往不如较瘦的人的信号稳定。另外，利用 EGG 信号研究嗓音还有一个不足之处就是，虽然在整个声带闭合阶段即闭相阶段，EGG 信号能够比较准确地反映声带接触面积的变化，但是在声带打开阶段及开相阶段，EGG 信号并不能精确地反映这一过程。这从 EGG 信号与其他嗓音信号如高速摄影图像信号、语音逆滤波信号的对比中就可发现。所以在利用 EGG 信号研究语言嗓音的时候，其开相的数据一般我们都不用，关于这一点，Baken（Baken，2000)曾提出过自己的质疑，他认为 EGG 信号不能精确地表示声门开启后的信息，不能被运用于与声门宽度有关的研究。

1.1.2. EGG 信号分析方法

本文主要的思路是通过对于 EGG 信号的计算分析来区别松紧嗓音，用到的主要参数是基频、开商和速度商参数。利用 EGG 信号来计算这些参数存在着计算方法上的差别，而各种方法又各有其优缺点。实际采用哪种方法，既有理论上的考虑，也有对实际采集信号质量的考量。

下图是一个典型的 EGG 信号，其横轴是时间轴，纵轴是声带接触面积。图中标出的 contacting event 及 de-contacting event 表示的是声门的闭合点和开启点，它们之间的区间即表示声带的闭相。

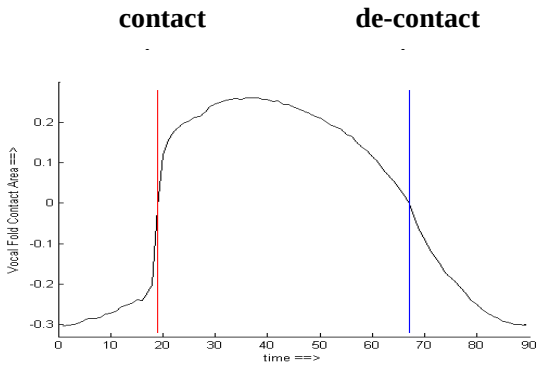


图 1.4 典型的 EGG 信号波形图

在 EGG 信号研究领域，有一个常用的概念，那就是接触商（contact quotient），简称 CQ。CQ 的定义是声带闭合段与整个声带运动周期的比例。EGG 信号的开商的定义是声带打开段与整个声带运动周期的比例。同时由于 EGG 信号只能比较精确地反映声带闭合段及闭相的运动信息，故我们把 EGG 信号的速度商定义为在闭相内的一个参数。下图是一个典型的 EGG 信号参数定义示意图（孔江平，2001），其横轴是时间轴，纵轴是声带接触面积，具体的定义如下：

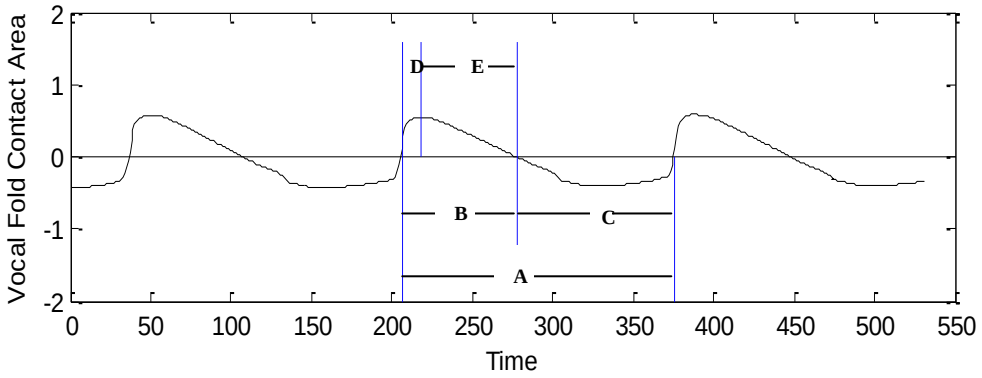


图 1.5 EGG 信号参数定义示意图

图中各字母所示时间段的意义如下：

A 表示一个嗓音周期（Cycle），定义为上一个声门关闭点到下一个声门关闭点之间的区间，及两个 contacting event 之间的区间；

B 表示一个闭相（Closed phase），定义为声门关闭点到声门开启点之间的区间，即周期内 contacting event 和 de-contacting event 之间的区间；

C 表示一个开相（Open phase），定义为声门开启点到下一个声门关闭点之间的区间，即 de-contacting event 到下一个 contacting event 之间的区间；

D 表示一个关闭相（Closing phase），定义为在闭相当中，从声门关闭点开始声带接触面积逐渐增大直到接触最大值之间的区间；

E 表示一个开启相（Opening phase），定义为闭相当中，从声带接触面积最大值开始到声带接触面积逐渐减小至声门开启点之间的区间。

对应的基于 EGG 信号的基频、开商、接触商、速度商参数的定义如下：

基频（Pitch）= 1/周期（A）；

开商（Open Quotient，简称 OQ）= 开相（C）/周期（A）×100%；

接触商（Contact Quotient，简称 CQ）= 闭相（B）/周期（A）×100%

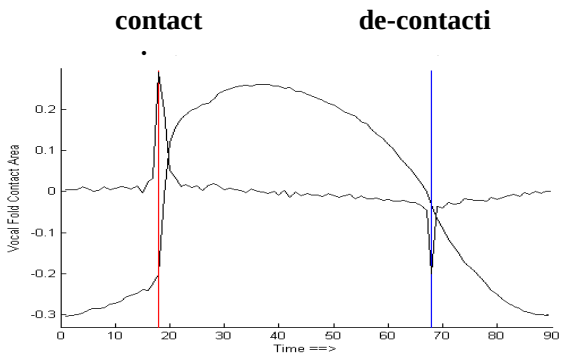
速度商（Speed Quotient，简称 SQ）= 开启相（E）/关闭相（D）×100%

由公式可知，开商和接触商是一组相对的概念，由于开相和闭相的和为周期，示意图上表示为 B+C=A，故可得 OQ+CQ=1（Henrich, 2004）。

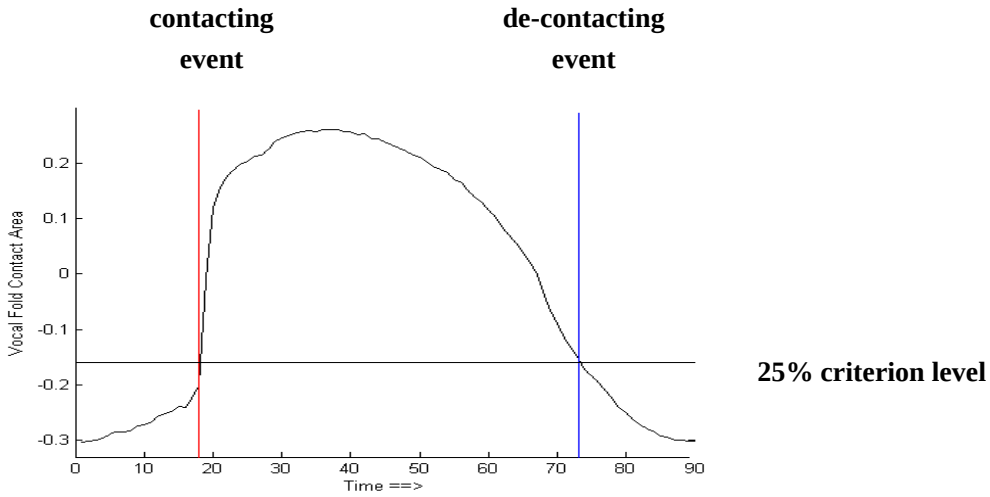
从 EGG 信号计算得到的 OQ、SQ 参数意在区别不同的 EGG 信号形式，从而反映不同的声带运动形式，对应不同的嗓音发声模式。从示意图上我们还能了解到从 EGG 信号计算基频、开商、速度商参数所需要做的工作：检测声门关闭点、声带接触面积最大值点、声门关闭点这三种重要参考点的位置。对于声带接触面积最大值点的检测一般没有异议，而对于声门关闭点和开启点的检测则有不同的考量，因此也带来 EGG 信号计算方法的不同。

基于对 EGG 信号声门关闭点和开启点的定义的不同，存在着三种计算 EGG 信号参数的方法，分别为微分（Derivative EGG Method）方法、标准线方法（Criterion-level Method）以及混合方法（Hybrid Method）。微分方法先对原始 EGG 信号进行一阶微分运算，然后根据周期内微分信号的最大值及最小

值确定声门关闭点和开启点。该方法是由 Childers et al.(1986)提出的，后来 Henrich et al.(2004)对于这种方法的准确性进行了证明。该方法的优点是符合直觉，物理意义明显。因为在声带接触或开启的瞬间，通过电子声门仪的电流会迅速增大或减小，在信号上表现为信号强度迅速增强或减弱，而一阶微分则是描述事物变化速率的数学工具。该方法的主要缺点是其抗噪性比较弱，容易受到噪声特别是高频噪声的影响，导致提取的参数不准确。下图是 EGG 信号微分方法示意图，图中黑色曲线为一个 EGG 信号周期，灰色曲线为该 EGG 信号周期的一阶微分信号，其声门关闭点由一阶微分信号的正向峰值点确定（图中 contacting event）；而声门开启点则是由一阶微分信号的负向峰值点确定（图中 de-contacting event）。



EGG 信号参数计算的标准线方法则是通过人为定义标准线（criterion-level）水平来确定声门关闭点和开启点的位置，最早由 Rothenberg（1981）提出，其优点是抗噪性比较好，缺点是没有微分方法准确，通过标准线方法计算得到的参数不具有绝对的物理意义，只是一个相对值，可以用于对 EGG 信号的比较分析，在计算时需要注明所定义的标准线水平。下图是 EGG 信号标准线方法示意图，水平线指示的是 25%标准线水平。这种方法需要检测出每个信号周期内的信号最大值和最小值，然后计算出其差值，标准线水平为 25%即表示将声门的关闭点和开启点定义在这个差值的 25%水平。



EGG 信号参数还有一种计算方法——混合方法，采取的是折中的思路来定义声门关闭点和开启点，由 Howard(1990，1995)提出。按照该方法，声门关闭点以 EGG 微分后信号的最大值(contacting peak)为标准，而声门开启点则采用的是标准线方法，一般以 42%(3/7)作为标准线水平。该方法主要是针对声门开启点运用微分方法往往不容易检测到的现象提出的。

由于本文所使用的录音材料里面存在一些不规则的情况，综合考量决定运用标准线方法来定义和计算参数，采用的标准线水平为 25%（C.Herbst 2006）。

1.2. 本文研究意义及理论背景

中国是一个幅员辽阔的多民族国家，有 56 个民族，有着丰富的语言文化资源，境内至少有一百多种语言，已确认的语言有八十多种，这为语言学的研究创造了非常有利的条件。境内丰富的少数民族语言资源中，最具特色的是拥有各种不同的发声类型，这些不同的发声类型有着很高的学术价值。国内学者也一直致力于境内的民族语言研究，早期研究内容以语言音系描写、不同语言或方言的对应比较、语言语源关系探讨为主。民族语有着很丰富的

发声类型，这在很早就受到国内学界的关注，如胡坦、戴庆厦(1964)、戴庆厦(1979)等。早期对于少数民族语特殊发声类型的研究主要集中在音系描写和历史比较方面，强调普通发声类型与特殊发声类型在音系中的对立关系，并把这种对立关系描述为“松紧音”的差别，实际上是把松、紧处理为一个附着于元音的音位特征，研究的主要出发点是音系学和音位学。

由于研究方法的限制，早期研究并未过多从言语产生的角度来定量分析这种松紧音的差别，只是模糊地把语言中存在的这种对立概括为“松紧”。随着国内实验语音学的发展，利用实验手段定量分析这种对立的论文逐渐多了起来（孔江平，2001；石锋、周德才，2005；汪锋、孔江平，2009），《论语言发声》（孔江平 2001）集中讨论了语言发声类型研究的方法和意义，对境内很多民族语进行了分析研究。该书中亦有利用 EGG 信号对于语言嗓音性质研究的实践，而本文同样也是通过 EGG 信号来对新平彝语的嗓音进行研究的。

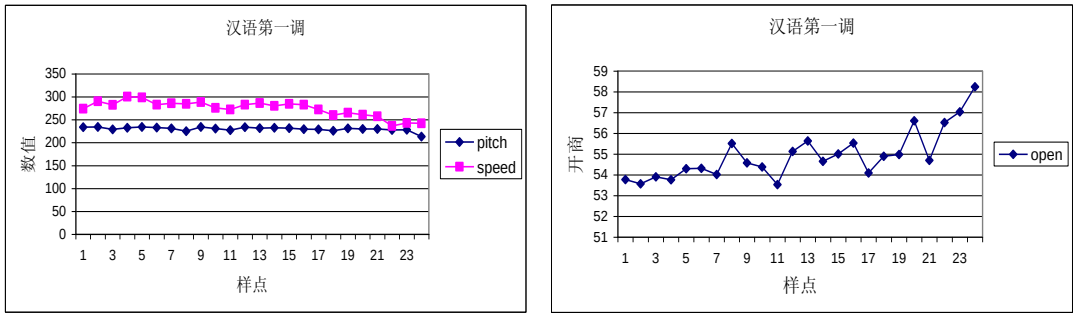
下面将简单介绍一下该书中利用 EGG 信号对语言嗓音进行研究的相关内容。书中利用男声元音[a]、[i]、[u]样本研究了嗓音参数开商速度商随着基频变化而变化的性质，结论是随着基频的提高，开商和速度商参数都会出现“转折点”的现象，在转折点的两端，嗓音参数体现为不同的性质。这个研究的意义在于说明开商速度商参数随着基频的提高的变化情况是复杂的，而不是简单的线性关系。

《论语言发声》一书区分了五种不同的发声类型，按照音调由低到高的顺序，依次是（1）气泡音（fry，F）；（2）气嗓音（breathy，B）；（3）紧喉音（creaky，C）；（4）正常嗓音（modal，M）；（5）高音调嗓音（high，H）。研究使用的元音样本同样也是元音[a][i][u]。研究得出这 5 种嗓音模式基频、开商、速度商的大小关系为，表中数字表示参数大小排序，数字越大，代表参数值越大：

表 1.1 不同发声类型嗓音参数比较

嗓音类型	气泡音	气嗓音	紧喉音	正常音	高音调嗓音
音调	1	2	3	4	5
速度商	4	1	5	3	2
开商	5	4	1	3	2

这些研究内容对本文研究都有很好的启示和参照作用，和本文研究内容直观相关的部分是《论语言发声》一书中对汉语普通话声调的发声模式的研究，本文的主体内容和该研究从方法到结构都很接近。书中给出了汉语普通话的单音节和双音节声调模式，使用参数为基频、开商和速度商三个参数。下图为书中给出的第一调的声调发声模式。



(a) (b)

图 1.8 （a）为第一调的基频和速度商参数 （b）为第一调的开商参数

本文之前还没有学者运用电子声门仪对新平彝语松紧嗓音的发声模式进行过系统研究，本文对新平彝语松紧音的发声模式进行客观描写和分析，试图利用 EGG 嗓音参数区分新平彝语的各个调类。

2. 新平彝语音位系统、录音词表及发音人信息

2.1. 新平彝语音位系统

新平彝语共有 35 个声母，18 个韵母，3 个声调，有非常整齐的松紧音对立，音系如下：

声母（35）

p	p^h	b	m	f	v
t	t^h	d	n		l ɬ
ts	ts^h	dz	s	z	
tʂ	$tʂ^h$	dʐ	ʂ	ʐ	
tɕ	$tɕ^h$	dʑ	ȵ	ɕ	ʑ
k	k^h	g	ŋ	x	ɣ
∅					

说明：（1）n 和ȵ不对立，可合并；（2）零声母∅多以喉塞音ʔ的形式出现。

韵母（18）

i	ɿ	ʅ	u	ɛ	ə	a	o	ie
i̱	ɿ̱	ʅ̱	u̱	ɛ̱	ə̱	a̱	o̱	ie̱

可以看出新平彝语的韵母系统由 9 对整齐的松紧元音组成。

声调（3）

第一调 33　　第二调 21　　第三调 55

新平彝语 33 调和 21 调都存在松紧音对立，而 55 调没有，为方便起见，本文将 33、21 调与松紧音的搭配分别处理为两个声调。从音系学和音位学上来说，这是没有问题的，因为这种标记方法同样能体现出语素音位的对立性质，且会带来记录上的方便。另外，这种标记方法还可以体现出把松紧差别与音调高低归并到一组嗓音特征上的思想，即我们把 33 松与 33 紧处理为不同的嗓音模式，它们的区别我们可以用不同的嗓音参数来区别。故本文将新平彝语声调系统一共处理为 5 个调位，即 33 松、33 紧、21 松、21 紧、55 松。分别用 1、2、3、4、5 调表示，重新排列为：

声调（5）

第一调 33 松　　第二调 33 紧　　第三调 21 松

第四调 21 紧　　第五调 55 松

2.2. 录音词表

笔者 2009 年 7 月份随北京大学中文系实验室一行人员赴云南调查民族语言，对新平彝语的音系进行录音，录音人次记 3 男 3 女。录音词表的设计充分考虑到了音系的完整性，同时也突出了该语言存在整齐的松紧元音对立的特征，特别是单音节声调表的设计，基本都是按照最小对立原则选取出来的。

单音节声调词表共计 85 个词语，在词表设计上充分考虑到最小对立原则，词表前 10 行例词体现出非常整齐的声调对立特性（包括一般意义上的松紧对立），后 10 行例词也都是比较整齐的最小对立对，共计 38 对最小对立对。单音节词表给出了例词的汉语解释及国际音标注音。

表 2.1 新平彝语单音节录音词表

33 松		33 紧		21 松		21 紧		55 松	
例词	音标	例词	音标	例词	音标	例词	音标	例词	音标
铜	dzɿ33	砍	dzɿ̱33	太阳	dzɿ21	断代	dzɿ̱21	皮	dzɿ55
借	$tʂ^hʅ$33	热	$tʂ^hʅ̱$33	臭	$tʂ^hʅ$21	削	$tʂ^hʅ̱$21	甜	$tʂ^hʅ$55
矮	ti33	浸泡	ti̱33	顶、抵	ti21	砸	ti̱21	推	ti55
打闹	bɛ33	射	bɛ̱33	掉	bɛ21	纠缠	bɛ̱21	壶	bɛ55
吃	tʂo33	撞	tʂo̱33	中（打中）	tʂo21	罩	tʂo̱21	筛	tʂo55
倒	bɤ33	围起	bɤ̱33	山	bɤ21	蹄	bɤ̱21	堆	bɤ55
虫	bu33	饱	bu̱33	背（东西）	bu21	霉	bu̱21	啼	bu55
说	ɣɯ33	喊	ɣɯ̱33	切（菜）	ɣɯ21	揉	ɣɯ̱21	舅舅	ɣɯ55
短/钝	də33	踩	də̱33	落（太阳落）	də21	结（果子）	də̱21	酒曲	də55
粘	na33	把（量）	na̱33	你	na21	唠叨	na̱21		
割（割肉）	dɛ33	淡（盐淡）	dɛ̱33	给	bi21	裂开	bi̱21		
朵	pu33	反（反面）	pu̱33	开门	k^ha21	扛	$k^ha̱$21		

放	to33	剁（剁肉）	to̱33	轻	la21	捞	la̱21		
天	mu33	吹（喇叭）	mu̱33	平/慢	p^{h}i21	吐（痰）	p^{h}i̱21		
软	nu33	羽毛	nu̱33	树枝	sʮ21	新/渴	sʮ21		
泻	fu33	蛆	fu̱33	斗（一斗）	tu21	点火/燃烧	tu̱21		
筋	dʑu33	害怕	dʑu̱33	箍儿	vɛ21	搓（搓绳）	vɛ̱21		
锄草	tʂ^{h}u33	秋	tʂ^{h}u̱33	干净	xʌ21	元（一元）	xa̱21		
				切（切菜）	ɣə21	针	ɣə̱21		
				二	nji21	压/按	nji̱21		

3. 实验设备、方法及数据处理

3.1. 实验设备简介

本次录音所使用的 EGG 设备是 The Glottal Enterprises 生产的 EG2-Standard 系列电子声门仪。图 3.1 右侧是电子声门仪电极片照片，左侧显示的是该设备的主要操作和功能按钮。

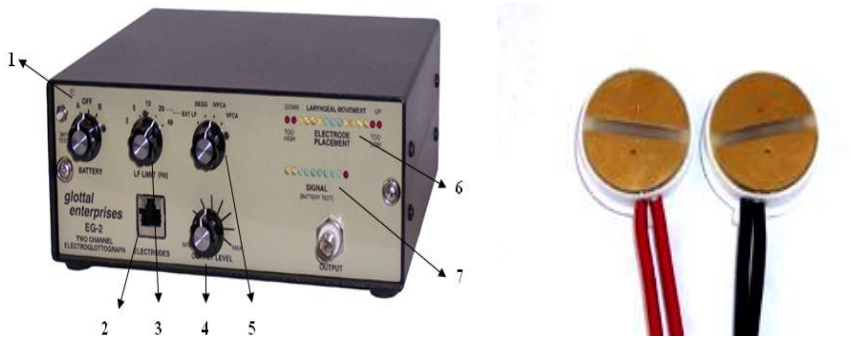


图 3.1 EG-2 Standard 系列 EGG[i]

下面给出两张图，分别是使用该 EGG 设备采集到的新平彝语的一个典型的松音 EGG 信号波形图和一个典型的紧音 EGG 信号波形图。从波形图能看出它们之间非常明显的差别，松音的开相时间明显较长，而紧音开相时间明显较短，表现在参数上，即为松音的开商要大于紧音；同为闭相区间，松音的 EGG 信号波形更多地表现出对称的特点，波形近似正弦波，而紧音的 EGG 信号波形更多地表现出非对称的特点，波形向左侧倾斜明显，表现在参数上，即为松紧的速度商参数要小于紧音。从 EGG 波形表现我们可以大致给新平彝语的松紧音做定性分析，其松音波形非常接近气嗓音 EGG 信号，而紧音波形很接近正常嗓音 EGG 波形。

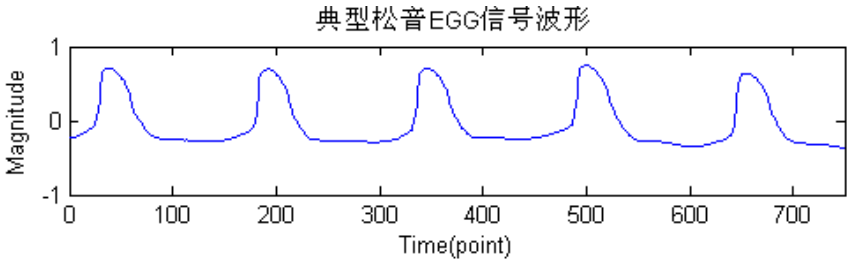


图 3.2 新平彝语典型松音 EGG 信号波形图

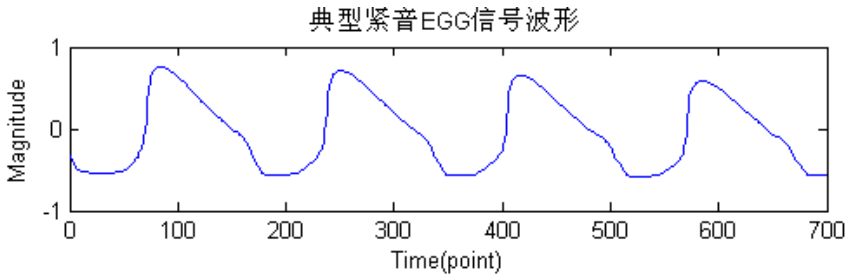


图 3.3 新平彝语典型紧音 EGG 信号波形图

3.2. 实验方法介绍

整个实验过程可以分为三部分：一是准备工作，包括查阅相关参考资料了解新平彝语音系，设计制作录音词表，寻找合适的发音人等；二是录音，包括录音前的录音设备连接调试等，本次录音使用的软件是 AdobeAudition 1.5。录音时每个录音词读两遍，如果发现发音人有读错的现象，需要及时纠正，录音文件保存时一般都是每一个发音人保存为一个数据文件；三是切音和数据分析，先使用 Audition 软件对录音文件进行切音，将每个词语都另存为一个.wav 格式语音文件，以方便后面的分析处理。然后是使用自编 Matlab 程序 Voicelab 提取文件参数，并将参数保存到 Microsoft Excel 表格中，最后在 Excel 表格中分别对男声和女声参数进行算术平均，并将平均后的参数以图表形式呈现出来。

4. 新平彝语单音节发声模式及松紧嗓音参数比较

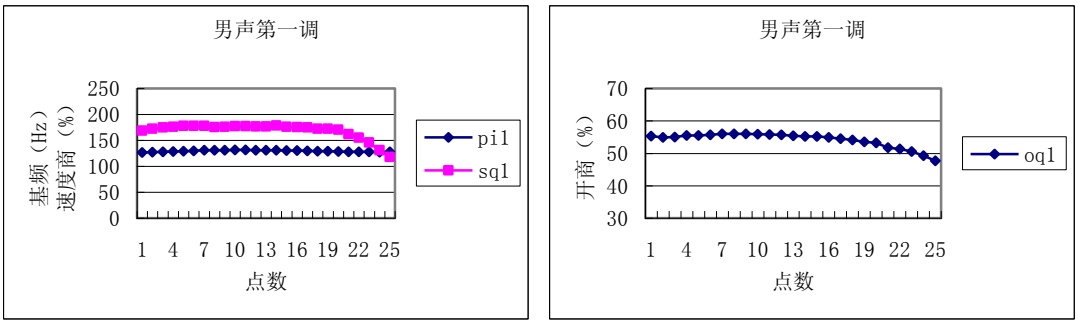
本章试图以基频、开商、速度商三个嗓音参数建立新平彝语的单音节发声模式，由于男女声发声性质存在一些区别，本文所给出的发声模式曲线会分列男女，体现男女差别。如前文所述，本文将松紧音处理为不同的调位，故将新平彝语声调系统处理为 5 个调类：33 松、33 紧、21 松、21 紧、55 松，分别记为 1、2、3、4、5 调。本章第一节将给出新平彝语单音节发生模式曲线图，第二节则会重点分析松紧音在基频、开商和速度商参数上的区别。第三节会在前两节的基础上总结出新平彝语单音节发声模式的区别性特征。

4.1. 新平彝语单音节发声模式

本节按性别对所有单音节声调文件基频、开商、速度商参数进行算术平均，将计算结果以曲线图形式呈现。图例中 pi（pitch）表示基频曲线，oq（open quotient）表示开商曲线，sq（speed quotient）表示速度商曲线，数字 1、2、3、4、5 表示不同的调类。

4.1.1. 新平彝语男声单音节发声模式

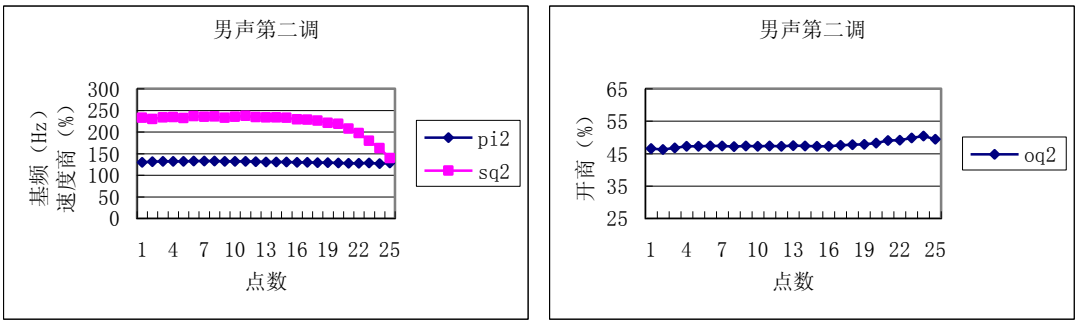
本小节将给出新平彝语的男声单音节发声模式曲线图 4.1-4.5。每一组图的左图为基频和速度商曲线图，右图为开商曲线图，基频参数单位为赫兹（Hz），开商和速度商参数单位为百分比（%）。在给出嗓音参数曲线图后，会就曲线形状和数值做简单描述，为下文总结发声模式做材料准备。



（a） (b)

图 4.1 （a）男声第一调基频、速度商参数 （b）男声第一调开商参数

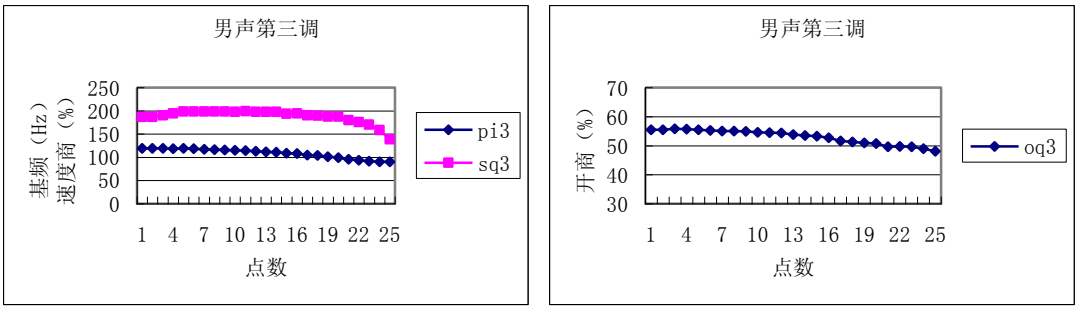
男声第一调基频曲线表现为平，数值在 125Hz 左右，与传统描写 33 中平调一致。速度商曲线前段表现为平，在末段有明显下降，前段数值在 170%左右，后段下降到 130%。开商曲线整体呈先平后降趋势，从前段的 55%下降到末端 47%。



（a） (b)

图 4.2（a）男声第二调基频、速度商参数 （b）男声第二调开商参数

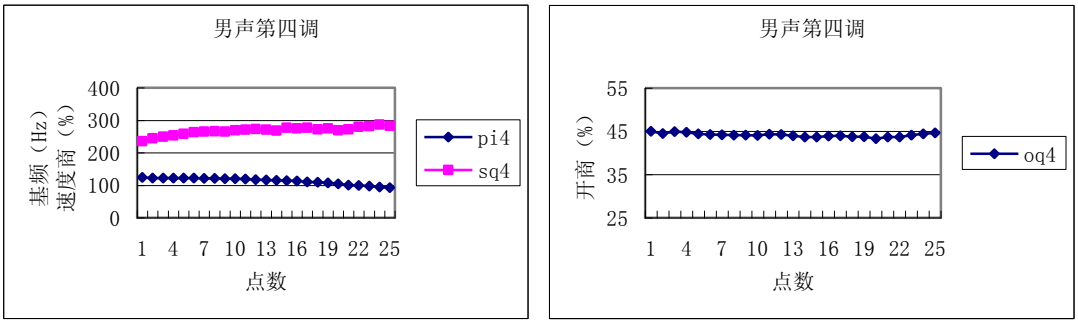
男声第二调基频曲线表现为平，数值在 125Hz 左右。速度商曲线前段表现为平，在末端有明显下降，前段数值在 235%左右，后段下降到 130%。开商曲线呈先平后升趋势，从前段的 45%上升到末端的 49%。



（a） (b)

图 4.3（a）男声第三调基频、速度商参数 （b）男声第三调开商参数

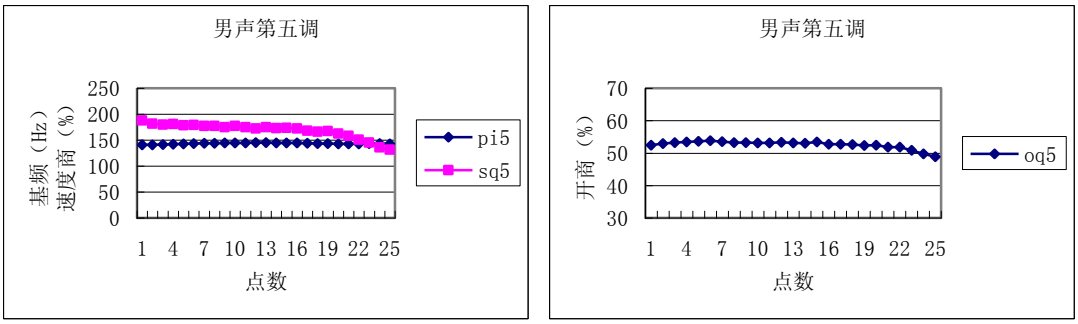
男声第三声基频曲线表现为降，从起始的 120Hz 下降到末段的 90Hz，与传统描写 21 低降调一致。速度商曲线前段较平，后段明显下降，前段数值在 195%左右，后段下降到 150%。开商曲线表现为降，从前段的 55%下降到末端的 48%。



（a） (b)

图 4.4（a）男声第四调基频、速度商参数 （b）男声第四调开商参数

男声第四调基频曲线表现为降，从起始的 120Hz 下降到末段的 90Hz。速度商曲线表现为升，从起始的 230%上升到末端的 285%。开商曲线表现为整体较平，但有小幅曲折趋势，数值在 45%下降到 43%再上升到 45%。



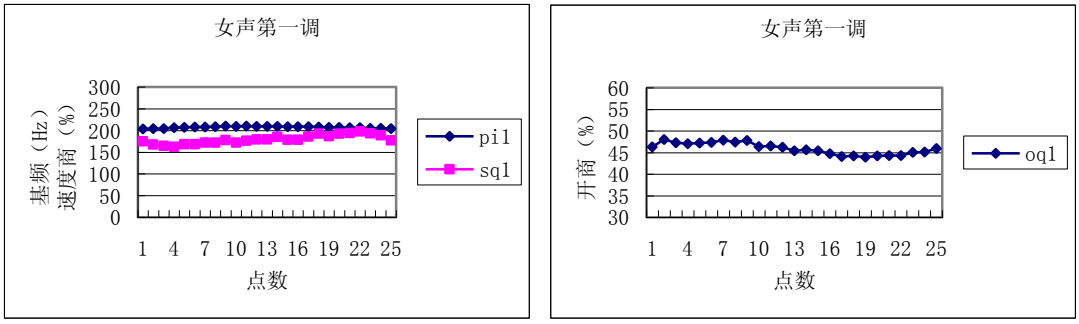
（a） (b)

图 4.5（a）男声第五调基频、速度商参数 （b）男声第五调开商参数

男声第五调基频曲线表现为平，数值在 145Hz 左右，与传统描写 55 高平调一致。速度商曲线表现

为降，数值从前段的 185%下降到末端 130%。开商曲线主体较平，但在末段有小幅下降，前段参数基本在 53%左右，尾段参数下降到 50%附近。

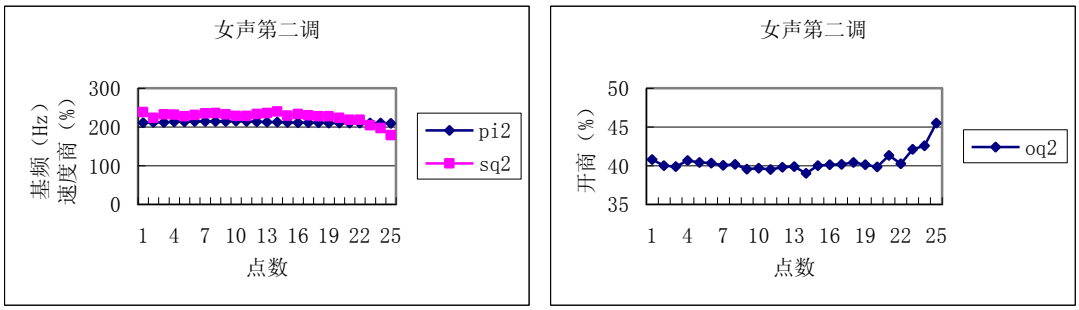
4.1.2. 新平彝语女声单音节发声模式



（a） (b)

图 4.6（a）女声第一调基频、速度商参数 （b）女声第一调开商参数

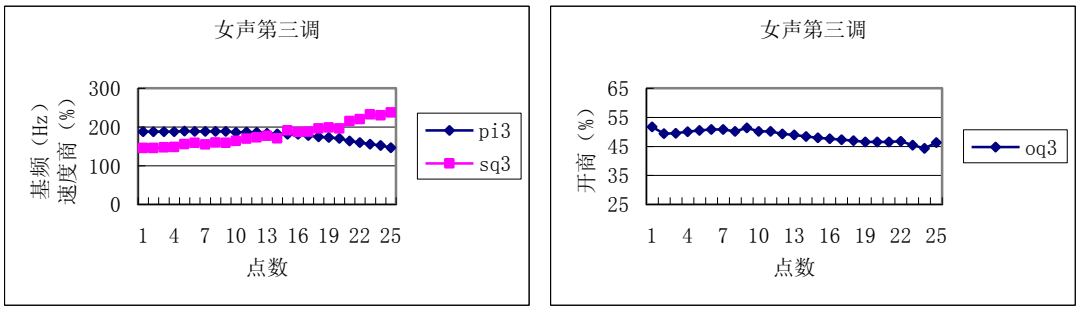
女声第一调的基频曲线表现为平，数值在 208Hz 左右，与传统描写 33 中平调一致。速度商曲线表现为小幅上升，数值从前段 165%上升到末端 190%左右。开商曲线表现为小幅下降，从前段的 48%下降到后段的 44%。



（a） (b)

图 4.7（a）女声第二调基频、速度商参数 （b）女声第二调开商参数

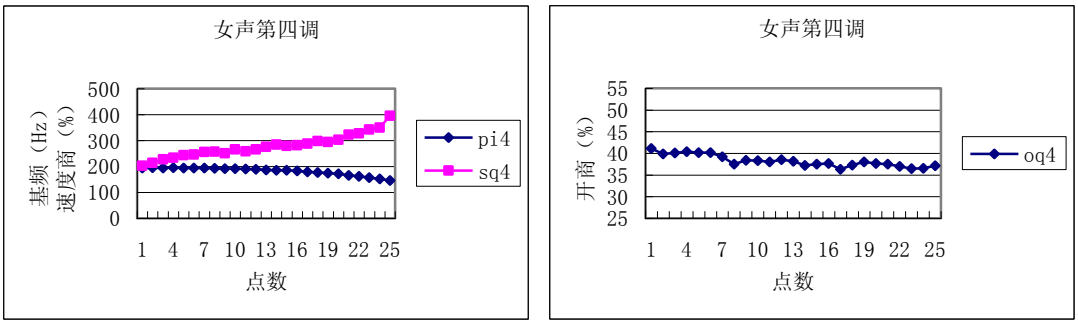
女声第二调的基频曲线表现为平，数值在 208Hz 左右。速度商曲线总体较平，但在末端有小幅下降，前段数值在 230%左右，末段数值下降到 180%。开商曲线总体较平，在末端有小幅上升，前段数值在 40%左右，末端上升到 46%。



（a） (b)

图 4.8（a）女声第三调基频、速度商参数 （b）女声第三调开商参数

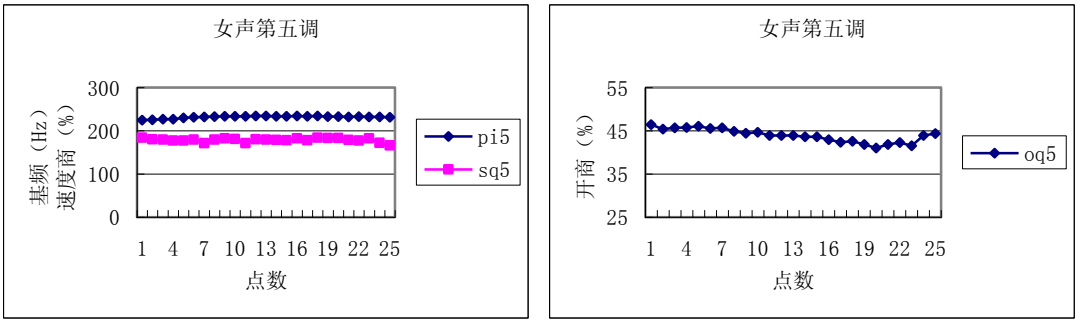
女声第三调基频曲线表现为降，数值从起始的 195Hz 下降到末尾的 150Hz，与传统描写 21 低降调一致。速度商曲线体现出明显升幅，从起始的 150%上升到末尾的 235%。开商曲线表现为降，从起始的 54%下降到末尾的 45%。



（a） (b)

图 4.9（a）女声第四调基频、速度商参数 （b）女声第四调开商参数

女声第四调基频曲线表现为降，从 195Hz 下降到末尾的 145Hz。速度商曲线表现出明显升幅，从起始的 210%上升到末端的 400%。开商曲线表现为降，从起始的 41%下降到末端的 36%。



（a） (b)

图 4.10（a）女声第五调基频、速度商参数 （b）女声第五调开商参数

女声第五调基频曲线表现为平，数值在 230Hz 左右，与传统描写 55 高平调一致。速度商曲线表现为平，数值在 180%左右。开商曲线表现为小幅下降，起始数据为 46%，最低点数据为 41%。

本小节将给出男声和女声 1-5 调基频参数曲线图 4.11-4.12，以比较松紧音基频曲线在大小和形状上的差别。

4.2. 新平彝语各声调嗓音参数比较

4.2.1. 单音节基频参数比较

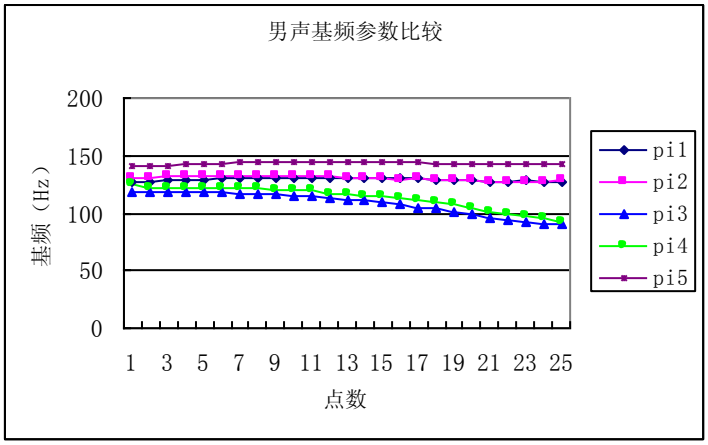


图 4.11 男声基频参数比较图

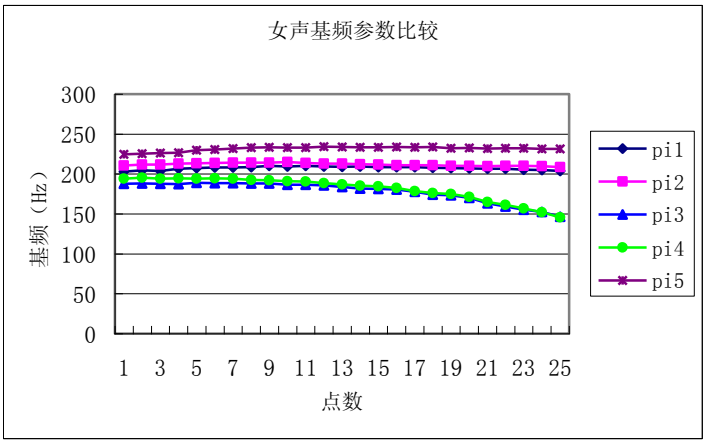


图 4.12 女声基频参数比较

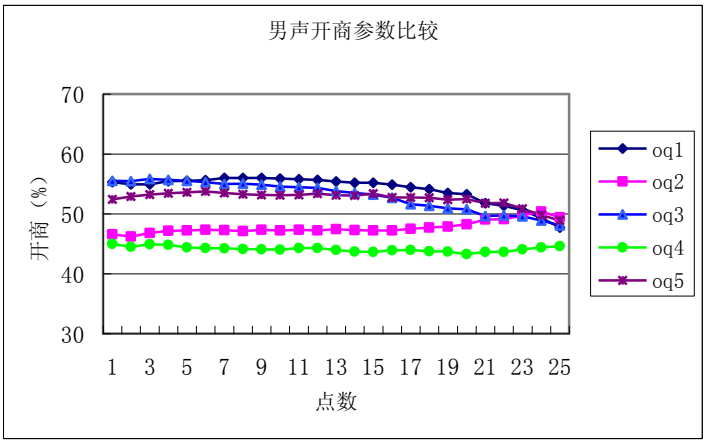
从比较图可以看出，男女声 1、2 调基频曲线和 3、4 调基频曲线都基本重合，4 调略高于 3 调。而 5 调基频曲线则明显高于其他声调。1、2、5 调为平调，3、4 调为降调。这里以 1 调基频参数为基准，计算 2 调和 5 调基频参数与 1 调参数的比值；以 3 调基频参数为基准，计算 4 调基频参数与 3 调参数的比值。然后再对所有参数点进行平均，计算结果如下（保留到小数点后 3 位有效数字）：

表 4.1 基频参数比较表

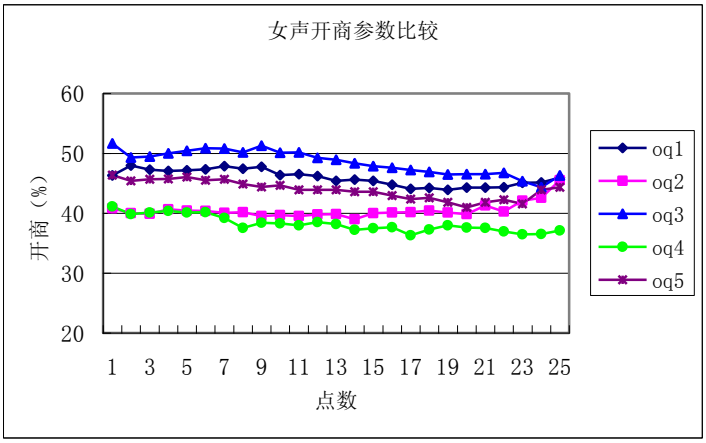
参数	pi2/1	pi5/1	pi4/3
男声均值	1.007	1.109	1.041
差值	0.007	0.109	0.041
女声均值	1.021	1.116	1.017
差值	0.021	0.116	0.017

可见，5 调基频明显高于 1 调基频，差别在 10%以上。紧音 2 调和 4 调基频分别略高于松音 1 调和 3 调，但差别都很小。

4.2.2. 单音节开商参数比较



4.13 男声开商参数比较图



4.14 女声开商参数比较图

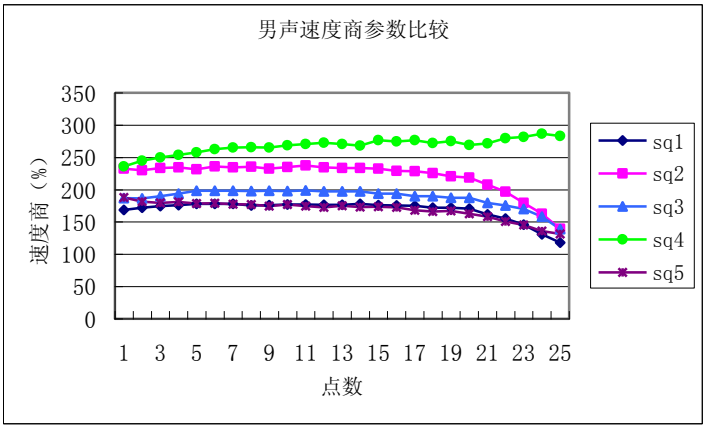
从比较图可以看出，紧音 2 调和 4 调的开商曲线都明显低于松音 1 调和 3 调曲线。5 调开商参数数值接近松音参数，大于紧音开商参数。下表为平均参数比较（保留到小数点后 3 位有效数字）：

表 4.2 开商参数比较表

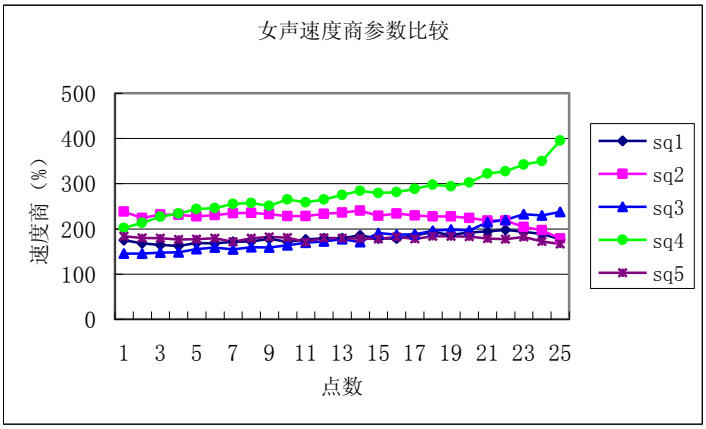
参数	oq2/1	oq5/1	oq4/3
男声均值	0.885	0.972	0.840
差值	-0.115	-0.028	-0.160
女声均值	0.882	0.956	0.786
差值	-0.118	-0.044	-0.214

紧音开商和松音差别明显，在 10%以上。而 5 调开商略小于 1 调，但差别较小，平均参数差别在 5%以内。

4.2.3. 单音节速度商参数比较



4.15 男声速度商参数比较图



4.16 女声速度商参数比较图

从比较图可以看出，紧音 2、4 调的速度商参数都明显高于松音 1、3 调。5 调速度商参数与松音 1 调参数接近，小于紧音速度商参数。下表为平均参数比较：

表 4.3 速度商参数比较表

参数	sq2/1	sq5/1	sq4/3
男声均值	1.301	1.007	1.459
差值	0.301	0.007	0.459
女声均值	1.265	1.005	1.532
差值	0.265	0.005	0.532

紧音速度商参数明显高于松音参数，2 调参数高于 1 调 25%以上，而 4 调速度商参数更是高出松音 3 调 45%以上，差别非常明显。5 调速度参数与 1 调非常接近，差别在 1%以下。

4.3. 新平彝语单音节发声模式的区别性特征

本小节试图定义新平彝语单音节发声模式的区别性特征，这是对整个嗓音参数系统的高度概括。要定义一个语言系统的嗓音区别性特征，需要以这个语言的整个嗓音参数系统作为参考。从前文给出的单音节基频、开商和速度商曲线和参数数值表现可以看出，新平彝语单音节嗓音参数曲线只出现了平或升降的变化，而并没有出现曲折的情况。故这里定义两个维度来描述发声模式的区别性特征：一个是嗓音参数值的大小，这里使用参数曲线的起点

值来代表；另外一个维度就是嗓音参数曲线的斜率。通过确定参数曲线的起点值大小和斜率，我们就可以把一条直线定义出来。本小节讨论区别性特征定义时，笔者先从单音节声调词表中挑出 9 套例词，每一套例词都包含 5 个例词，分别为第 1 至第 5 调。这里试图通过绘制这些例词的嗓音参数表现图，以确定参考空间，然后再定义新平彝语嗓音参数的区别性特征。

4.3.1. 基频模式的区别性特征

下面分别给出 9 套例词的的男声和女声的基频曲线表现，这里使用二维参数图形式，图中的每一个小图标（如* +）都代表一个二维数值，分别表示一个例词基频曲线的斜率和基频曲线的起点值。其中横轴表示的是基频曲线的斜率，纵轴表示的是基频曲线的起点值大小，单位为 Hz，图例中 T1-T5 分别代表第一至第五调（下同），分别用不同的图标表示。图中椭圆基于各调类参数拟合出来的。

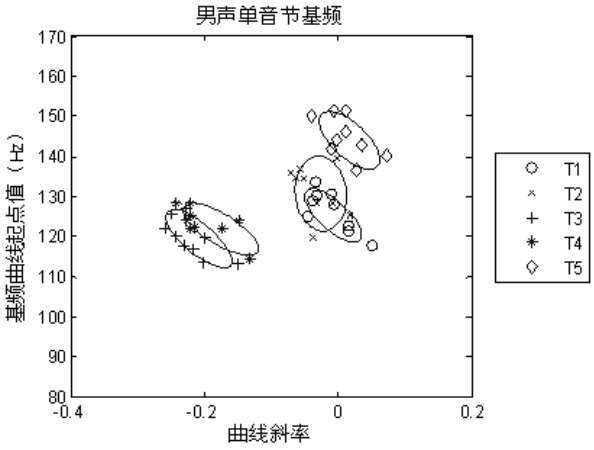


图 4.17　男声基频曲线二维参数图

从上图可以看出，男声基频 1-5 调集中为三块，1、2 调集中在一个区域；3、4 调集中在一个区域；5 调集中在一个区域，从拟合椭圆上看，就是 1、2 调拟合椭圆重叠较多；3、4 调拟合椭圆重叠较多；5 调拟合椭圆和其他调类拟合椭圆不重叠。就大小而言，3、4 调区域最低；1、2 调区域居中；5 调最高。就曲线斜率而言，1、2、5 调曲线斜率都在 0 附近上下浮动，体现出参数曲线形状较平的特点；而 3、4 调曲线斜率则集中在-0.2 附近区域，体现出参数曲线的下降表现。

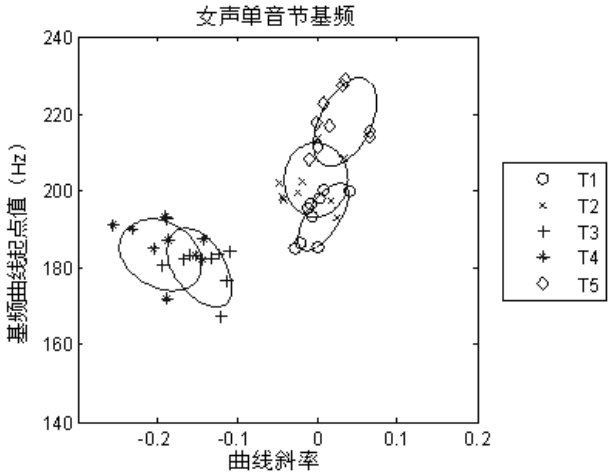


图 4.18　女声基频曲线二维参数图

从上图可以看出，女声基频参数集中程度与男声一致，按集中程度可分为三块，1、2 调集中在一个区域；3、4 调集中在一个区域；5 调集中在一个区域。就大小而言，3、4 调集中区域最低；1、2 调集中区域居中；5 调最高。就曲线斜率而言，1、2、5 调曲线斜率都集中在 0 附近；而 3、4 调曲线斜率则集中在[-0.3 -0.1]区域，体现出曲线下降趋势。

参考男女声基频曲线表现图及 4.1 节基频平均参数曲线描述，这里将新平彝语基频的区别性特征定义为三类： 1、2 调为中平，3、4 调为低降，5 调为高平。本文用 L（Low）表示低，M（Middle）表示中，H（High）表示高，F（Falling）表示降，R（Rising）代表上升。若参数曲线表现为平，则只需使用 L、M 或 H 来定义其区别性特征；若参数曲线表现为升或降，则需要同时定义参数起点位置和曲线斜率来定义区别性特征，如将男声 3 调基频定义为 LF，即表示其起点值较低，曲线形状为降。通过上述定义，得出新平彝语的单音节基频模式区别性特征如下表，表中附带数值表现栏，该栏内数字为所有单音节录音文件的参数平均值，栏内数字若为一个数字，则表示曲线形状较平，该数字为参数均值；若为两个数字，则表示曲线有升降表现，分别给出的是起点和终点的参数值，若为三个数字，则表示曲线有曲折表现，分别给出的是起点、拐点以及终点的参数值，下文区别特征模式表中出现的数值栏定义都与此一致，不一一赘述。

表 4.4 基于区别性特征的单音节基频模式

调类	1 调	2 调	3 调	4 调	5 调
男声	M	M	LF	LF	H
数值（Hz）	125	125	120-90	120-90	145
女声	M	M	LF	LF	H
数值（Hz）	208	208	195-150	195-145	230

由区别性特征表可以看出，基频参数不能将 1、2 调和 3、4 调区分开来，而只能将 5 个调类区分为 3 组，分别为：1、2 调为一组，表现为中平；3、4 调为 1 组，表现为低降；5 调为 1 组，表现为高平。

4.3.2. 开商模式的区别性特征

下面分别给出 9 套例词的的男声和女声的开商曲线表现，图中的每一个小图标（如* +）代表一个二维数值，分别表示一个例词开商曲线的斜率和开商曲线的起点值。其中横轴表示的是开商曲线的斜率，纵轴表示的是开商曲线的起点值大小，单位为%。

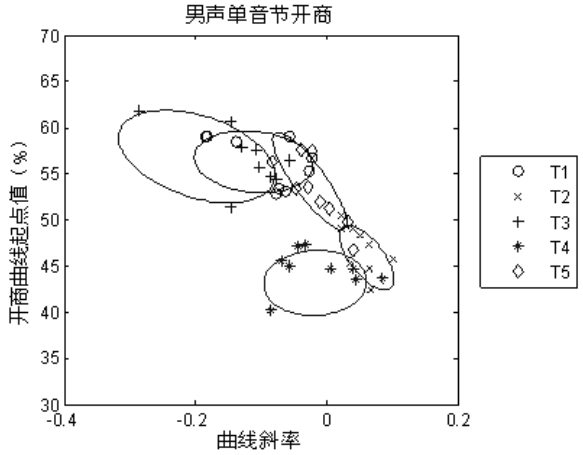


图 4.19　男声开商曲线二维参数图

男声 1、3 调开商参数重合度较高，其斜率都为负，且起点值都较高。2 调开商参数起点值较低，斜率为正。4 调开商参数起点值都较小，斜率有正有负，但斜率的绝对值都较小，升降变化不大，总体较平。5 调开商参数斜率有正有负，斜率绝对值都较小，反映出曲线总体较平的特点，平均而言，其起点值高于 2、4 调，但低于 1、3 调。从拟合椭圆分布上看，1、3 调拟合椭圆重合较多，5 调拟合椭圆和 1 调部分重合，但不和 3 调重合，松音 1、3、5 调和紧音 2、4 调拟合椭圆基本不重合，且紧音 2、4 调拟合椭圆位置明显低于松音 1、3、5 调。这里可将男声开商参数起点值定义为高（H）、中（M）和低（L）三档：开商起点参数大于等于 50%的区域定义为 H；小于等于 45%的区域定义为 L；在 45%至 50%之间的区域定义为 M。就升降表现而言，1、3 调明显下降，记为 F；2 调明显上升，记为 R；4、5 调参数升降变化不大，总体较平。

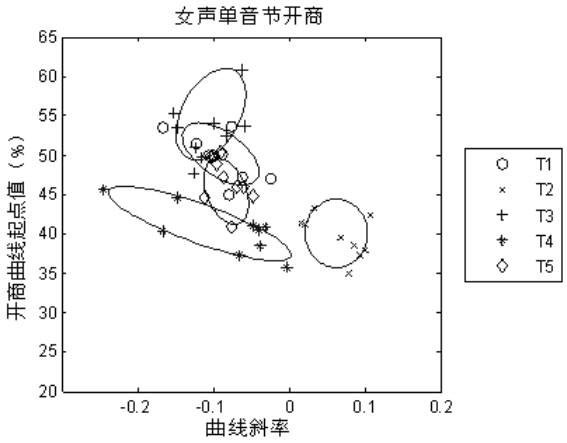


图 4.20　女声开商曲线二维参数图

整体来说，女声 3 调开商参数的起点值最高，1 调、5 调次之，2 调和 4 调开商参数最低。从拟合椭圆分布上看，同样，1 调拟合椭圆分别与 3 调和 5 调拟合椭圆都有重合，但是同为松音，3 调和 5 调拟合椭圆没有出现重合的情况，这一点与男声情况一致，表现出 3 调和 5 调在开商参数上的区别；紧音 2、4 调拟合椭圆位置明显低于松音 1、3、5 调。这里可以考虑将开商曲线起点值参数分为 H、M 和 L 三档：3 调起点开商参数主要集中在[50% 60%]区间，可定义此区间为 H；1、5 调开商起点参数主要集中在[45% 50%]区间，可定义此区间为 M；2、4 调开商起点参数基本都在 45%以下，可定义此区间为 L。就曲线斜率而言，1、3、4、5 调斜率为负，意味着开商曲线呈下降趋势；2 调表现为升，曲线表现为升。

根据这里给出的男女声开商曲线表现及 4.1 节开商平均参数描述，给出新平彝语单音节开商模式的区别性特征表如下。

表 4.5 基于区别性特征的单音节开商模式

调类	1 调	2 调	3 调	4 调	5 调
男声	HF	LR	HF	L	H
数值（%）	55-47	45-49	55-48	44	52
女声	MF	LR	HF	LF	MF
数值（%）	48-44	40-46	54-45	41-36	46-42

从开商模式的区别性特征表可以看出，新平彝语的松紧音表现出很好的区分特性。男声松音 1、3 调开商参数性质较接近，都表现高降 HF，从具体数值表现来说也很接近，其平均参数起点都在 H 区域，终点则基本都在 M 区域，没有到达 L 区域。男声 5 调开商参数总体来说表现为 H。男声紧音 2、4 调平均开商参数全都集中在 L 和 M 区域，没有到达 H 区域，与松音参数区域相区别。女声松音 1、3、5 调开商参数都表现为降，起点开商参数都在 45%以上，而紧音 2、4 调开商参数起点都很低，且平均开商参数大多数数值都在 45%以下。体现出较好的松紧对立特性。

4.3.3. 速度商模式的区别性特征

下面分别给出 9 套例词的的男声和女声的速度商曲线表现，图中的每一个小图标（如* +）代表一个二维数值，分别表示一个例词速度商曲线的斜率和速度商曲线的起点值。其中横轴表示的是速度商曲线的斜率，纵轴表示的是速度商曲线的起点值大小，单位为%。

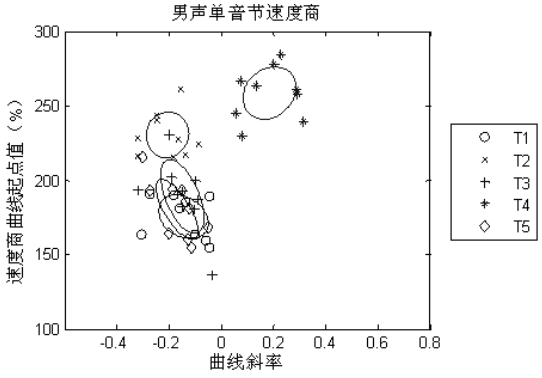


图 4.21　男声速度商曲线二维参数图

就斜率表现而言，男声 1-5 调速度商分为两组，1、2、3、5 调都为负；4 调为正，即表示男声 1、2、3、5 调速度商曲线都呈下降趋势；4 调速度商曲线呈上升趋势。从速度商参数起点值大小而言，1、3、5 调速度商曲线起点值大多集中在 200%以下，而紧音 2、4 调速度商曲线全在 200%以上，表现出很好的区分。从拟合椭圆位置分布看，松音 1、3、5 调拟合椭圆表现出了高度重合的情况，且位置都明显低于紧音 2、4 调拟合椭圆。故这里可以将男声速度商的起点参数值特征定义为两档：H 和 L，速度商起点数值在 200%以上的定义为 H，在 200%以下的定义为 L。

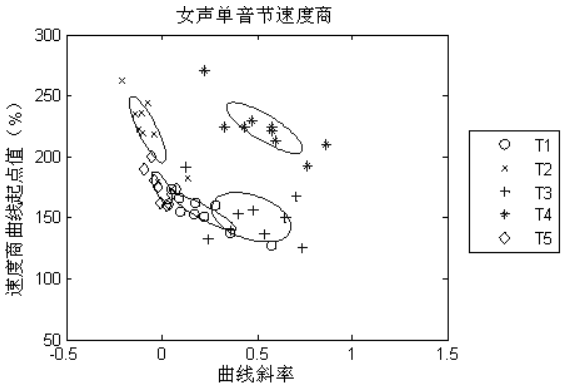


图 4.22　女声速度商曲线二维参数图

从拟合椭圆分布上看，1、3 调拟合椭圆有较大重合，都区别于松音 5 调情况，5 调速度商位置稍高于 1、3 调。紧音 2、4 调拟合椭圆位置明显高于松音 1、3、5 调。从曲线斜率表现看，1、3、4 调速度商曲线斜率为正，曲线呈上升趋势；2 调斜率为负，曲线呈下降趋势；5 调曲线斜率有正有负，但绝对值都较小，表现出较平的特性。从速度商曲线起点参数值而言，可分为高（H）和低（L）两档，基本情况是高于 200%的定义为 H，低于 200%的定义为 L。这样的话，2、4 紧音调起点参数值大多表现为 H，松音 1、3、5 调起点参数值绝大多数表现为 L。

根据这里给出的男女声速度商曲线表现及 4.1 节速度商平均参数描述，给出新平彝语单音节速度商模式的区别性特征表如下。

表 4.6 基于区别性特征的单音节速度商模式

调类	1 调	2 调	3 调	4 调	5 调
男声	LF	HF	LF	HR	LF
数值（%）	170-130	235-130	195-150	230-285	185-130
女声	LR	HF	LR	HR	L
数值（%）	165-190	230-180	150-235	210-400	180

男声松音 1、3、5 调速度商性质较接近，从区别性特征上看，都表现为 LF，其参数值全都集中在 L 区域，而紧音 2、4 调速度商参数则大多集中在 H 区域，只有 2 调末段有部分参数降到 L 区域，松紧对立表现出很好的参数差别。女声松音 1、3、5 调参数绝大多数集中在 L 区域，只有 3 调末段上升到 H 区域，而紧音 2、4 调速度商参数则基本集中在 H 区域，只有 2 调末段有部分参数下降到 L 区域。体现出很好的区分性质。

4.4. 新平彝语单音节发声模式讨论

下面先分男女声将新平彝语单音节发声模式列表如下：

表 4.7 新平彝语男声发声模式

调类	1 调	2 调	3 调	4 调	5 调
基频	M	M	LF	LF	H
开商	HF	LR	HF	L	H
速度商	LF	HF	LF	HR	LF

表 4.8 新平彝语女声发声模式

调类	1 调	2 调	3 调	4 调	5 调
基频	M	M	LF	LF	H
开商	MF	LR	HF	LF	MF
速度商	LR	HF	LR	HR	L

男女声的基频模式都表现为 M、LF 和 H 三种；男声开商表现为 HF、H、LR、L 四种，女声开商表现为 MF、HF、LR 和 LF 四种；男声速度商表现为 LF、HF 和 HR 三种；女声速度商表现为 LR、L、HF 和 HR 四种。男声 1、2 调，3、4 调基频模式相同，1、3 调开商模式相同，1、3、5 调速度商模式相同；女声 1、2 调，3、4 调基频模式相同，1、5 调开商模式相同，1、3 调速度商模式相同。

从嗓音参数关系看，基频相同，开商速度商表现可以不同，如男女 1、2 调基频都表现为 M，但开商和速度商表现则完全不同，而 1、2 调也正是通过开商和速度商参数相区别的。相反，基频不同，开商速度商参数也可以有相同表现，如男声 1、3 调基频分别表现为 M 和 LF，但开商都表现为 HF，速度商都表现为 LF。开商、速度商两个参数之间存在较为紧密的关系，一般地，开商带有 H 特征，速度商一般都带有 L 特征；而开商带有 L 特征的，速度商一般都带有 H 特征，二者表现出数值上的负相关关系，这在男女声数据中都得到了体现。另外，女声开商速度商曲线的变化趋势大多都呈负相关关系，除 5 调外，其他 4 个调类基本上是若开商为升 R，则速度商表现为降 F，若开商表现为 F，则速度商表现为 R。但这种负相关关系并不是绝对的，如男声发声模式很多都不体现这种负相关关系，女声 5 调开商表现为 F，但速度商并没有表现出明显的 R 趋势。

传统语言学研究多将松音归为元音所附带的性质，称为“松紧元音”。但从言语产生机制上说，“松紧”并不是元音的特征，而是发声模式的区别。针对这个问题，孔（2001）提出了“调质”（Tone Quality）的概念，从发声的角度将声带的不同发声方式定义为调质的不同。并将声调进一步分解为“调时”和“调声”，其中“调时”指的是肌肉对声带振动快慢的调制，在声学上对应于嗓音发声类型的时域特征，常用“基频”参数来反映；而“调声”指的是肌肉对声带振动方式的调制，声学上对应于嗓音发声类型的频率域特征，常用“开商”和“速度商”参数来表示。对新平彝语的嗓音参数的计算分析证明了“调质”概念的必要性，即在声调感知中，不仅反映时域特征的“基频”参数有着独立的意义，反映频域特征的“开商”、“速度商”参数也有着独立的意义。

这里值得讨论的是基频、开商和速度商这 3 个参数在声调感知中到底各自起了多大作用。从前文对三者的讨论中可以看出，基频参数和开商、速度商参数是独立的关系，反映出“调时”和“调声”参数的区别；但开商和速度商参数之间存在着一定的负相关关系，这两个参数到底是独立在其作用还是作为一个整体在起作用，二者在整体起作用时，到底哪一个参数的作用更为明显，这些都是非常值得研究的感知课题。从新平彝语来看，整体而言，都表现为松调开商参数较大，速度商参数较小；紧调开商参数较小，速度商参数较大，二者很有可能是一起起作用的，想要研究这两个参数对于感知的贡献，需要设计合成感知实验进行论证，例如可以设计实验如下：用 1 调或 2 调的基频参数、1 调的开商参数和 2 调的速度商参数合成语音信号，让大量新平彝语母语者听辨，看被试者是听成 1 调更多还是 2 调更多，若听成 1 调更多，说明开商的感知贡献更大；若听成 2 调的更多，则说明速度商的感知贡献更大，若听成 1 调和 2 调的情况区别不大，则说明开商和速度商对于嗓音松紧的感知贡献区别不大。

纯粹从语言学的区分角度分析，完全可以将开商和速度商两个参数作为一个整体看待，将开商较大、速度商较小的情况归纳为“调声”的“松”；将开商较小、速度商较大的情况归纳为“调声”的“紧”，这样的话，“调时”参数基频形成了 3 个音位，而“调声”参数开商、速度商整体形成了 2 个音位，在与不同基频搭配时表现出来的差异可以理解为调声“松紧”音位在不同环境下表现出来的音位变体。

这样处理的好处是简单，将“调声”只处理为两种模式，配合“调时”基频参数的 M、LF 和 H 三种模式，最多可形成 2×3=6 种模式，新平彝语由于 H 调没有相应紧音与之搭配，一共出现了 5 种。在处理新平彝语音系时，可以将开商和速度商参数看作一个整体进行处理，只区分“松”和“紧”两种情况，这样处理会给音系带来非常简洁的形式，而开商和速度商参数之间的相关关系又给这种处理提供了一定的理论支持。而这种处理和传统语言学的“松紧元音”处理在音系学上可以说是殊途同归，传统语言学的“松紧元音”也是处理为两种形式，与这里的区别只是在于它将“松紧”概念附着于元音之上，未认识到“松紧”其实是嗓音发声模式的区别，而不是元音的区别。

5. 参考文献

[1].陈康（1988）彝语的紧调类，民族语文，第 1 期.

[2].戴庆厦（1979）我国藏缅语族松紧元音来源初探，民族语文，第 1 期.

[3].方特、高奋（1994）言语科学与言语技术，商务印书馆.

[4].胡坦、戴庆厦（1964）哈尼语元音的松紧，中国语文，第 1 期.

[5].孔江平（1996）哈尼语发声类型声学研究及音质概念的讨论，民族语文，第 1 期.

[6].孔江平（2001）论语言发声，2001 年 11 月，中央民族大学出版社.

[7].孔江平（待刊) 语言发声研究的基本方法.

[8].罗素（2003）人类的知识，商务印书馆.

[9].钱一凡（2009）言语发声与艺术发声嗓音特性比较研究，北京大学硕士毕业论文.

[10].石锋、周德才（2005）南部彝语松紧元音的声学表现，语言研究，第 1 期.

[11].汪锋、孔江平（2009）武定彝语松紧音研究，《中国语言学》2，山东教育出版社.

[12].王士元、彭刚（2006）语言、语音与技术，上海教育出版社.

[13].杨若晓（2009）基于发声的汉语普通话四声的范畴知觉研究，北京大学硕士毕业论文.

[14].Baer T., Lofquist A., McGarr NS., (1983) Laryngeal Vibrations: A Comparison between High-speed Filming and Glottographic Techniques, JASA, 73, 1304-1308.

[15].Baken.RJ（1992） Electroglottography, Journal of Voice，Vol.6, No.2, 98-110.

[16].Baken.RJ, Orlikoff RF（2000） Correlates of Vocal Fold Motion, Clinical Measurement of Speech and Voice, 2nd edition, Singular/Thompson Learning.

[17].Childers, D.G., Hicks, D.M., Moore, G.P., (1986) A Model for Vocal Fold Vibratory Motion, Contact Area, and the Electroglottogram, JASA, 80, 1309-1320.

[18].Christian Herbst, Sten Ternström (2006) A comparison of different methods to measure the EGG contact quotient, Logopedics Phoniatrics Vocology, 31, 126-138.

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

[20].Fourcin, A.J., Abberton, E., (1971) First Application of a New Laryngograph. Med Biol Illust, 21, 172-182.

[21].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.

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

[23].Kong, J, (2007) Laryngeal dynamics and physiological models: high speed imaging and acoustical techniques, Peking University Press.

[24].N.Henrich, C.d'Alessandro, B.Doval, and M.Castellengo (2004) On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. JASA 115, no.3, 1321-1332.

[25].Peng,G. (2006) Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese. Journal of Chinese Linguistics,34.1, 134-154.

[26].Peter Ladefoged, (2006) A course in phonetics, Thomson Wadsworth.

[27].Rothenberg, M (1981) Some relations between glottal air flow and vocal fold
contact area, ASHA Rep.11, 88–96.
[28].Rothenberg, M, and Mahshie, J. J. (1988) Monitoring vocal fold abduction
through vocal fold contact area, Journal of Speech and Hearing Research, 31,
338–351.
[29].Fabre P, (1957) Un procédé électrique percutané d'inscription de l'accolement
glottique au cours de la phonation: glottographie de haute fréquence: premiers
resultats. Bulletin de I'A cademie Nationale de Médécine,3–4, 66–69.

[i]图片取自 Glottal Enterprises 公司网站

武鸣壮语双音节声调空间分布研究

潘晓声 孔江平

摘要：本文以武鸣壮语为例，从音位负担和语音学的角度探讨了双音节声调分类的研究方法。传统的研究方法是从音节的声调具有最小对立的理论出发，研究双音节声调的分布和类型。本文认为以单音节声调的组合形式表示武鸣壮语双音节声调，其调类过多。由于双音节声调的音位功能负担量不同，其中功能负担很小的双音节声调，冗余性较大，已失去声调的功能，调值上也不稳定，它们会进行较大的合并。本文对武鸣壮语双音节声调的声学分布空间进行聚类分析，总结得出双音节声调的实际分布模式，并对这种方法进行了讨论。

关键词：双音节声调、音位功能负担量、自适应仿射传播聚类

1. 引言

壮语系属汉藏语系壮侗语族壮傣语支，其内部大体可以分为北部方言和南部方言，以武鸣壮语和龙州壮语为代表。武鸣壮语属邕北土语。1957 年 11 月，国务院批准的壮文方案，确定壮语以北壮方言为基础方言，以武鸣为标准音。武鸣县境位于广西壮族自治区中南部，是南宁市辖县，距南宁市区公路里程 37 公里。全县总面积 3366 平方公里，按 1990 年人口普查，壮族占全县总人口 86.5%。

赵元任创立了“五度值”用于描写声调调值，为声调的音位学和语言学研究奠定了基础。但这种方法是基于人们对声调的听觉感知和音位结构的处理，用这种方法《武鸣土语》(李方桂 2005)将声调分为 6 类，若短音节单独算则为 9 类。《壮语方言研究》(张均如 1999)将其归为 6 个舒声调和 2 个促声调，其中每个促声调又分为长元音和短元音两类，总共为 10 类。上述研究都采用了听辩记音的方法。由于记音人往往不是母语者，又没有做过调类的感知测试实验，很难做到类别准确和调值精准。因此，研究者常常会在文献中看到前人的调查记音的声调与今人调查有所出入。此时则无法确定记音的差别是由音变造成的还是由不同记音者听觉感知的差异造成。

对于此类问题，人们常采用实验语音学的方法，提取语音实际基频来验证记音的正确性。孔江平(孔江平，2005)的研究以实验语音学的方法提取音节的基频进行分析，将其分为舒声调和促声调。每个促声调分为长元音调、短元音调和变调，并指出武鸣壮语的声调具有稳定的调形和调值，实际的调值和韵母的长短也有密切的关系。促声调的长调和短调调值相同，最终将武鸣壮语分为 10 个调类,6 种调值。这种方法在实际的调值上达到准确，但基本方法还是基于传统语言学方法。目前国内用实验语音学研究声调大都以前人的调类为研究基础，如果研究者对调类的归纳出现了偏差，即便使用实验语音学的方法也无济于事。

音位功能负担（Function Load）的概念最初由布拉格学派提出，表示音位的信息量，指某个音位在一定条件下出现的频率。音位的单位可以是声母、韵母和声调。王士元首次实现了功能负担的计算(Wang 1967)，可以用之计算声调的音位功能负担量。《壮语简志》(韦庆稳、覃国生 1980)中指出壮语 1 调字数最多，4 调最少，6 调字也不多。每个调字数分布不均匀正好反应了声调的音位功能负担量并不同。

对于武鸣壮语双音节声调的研究，前人研究主要关注双音节的连读变调。但所有研究都是在双音节声调是单音节声调的线性组合这一假设前提之下进行。武鸣壮语单音节共 6 种调值，双音节声调的调值可以有 36 种。本文认为每个双音节声调的音位功能负担量不等。在双音节声调数量很多的情况之下，负担量很小的双音节声调，冗余性高，其作为最小对立的功能会逐渐失去，因此其调值变得不稳定，并导致双音节声调合并。因此，在壮语中，单音节声调的音位功能是显著的，以调类为单位进行声调的声学分析和归类是合理的。但双音节声调组合的音位功能已经很弱，调值也很不未定，此时再按调类进行声学处理，就很不合理了。

本文将从实际语音出发，用实验语音学的方法提取武鸣壮语单双音节声调的基频并进行相应的预处理。首先对单音节的基频数据进行聚类，以具有音位功能的单音节声调为基准，验证聚类方法得到

的调类与音位学归纳调类的一致性。在此基础之上对双音节基频进行聚类，研究双音节调类的实际分布。本文认为通过观察所有语音的基频分布空间，并对其进行聚类应有助于对于调类的归纳。

2. 语音材料

本项研究在孔江平(孔江平 2005)的研究基础之上进行。一共四位发音人，两男两女，每个词发两遍。语音数据分为单音节和双音节两类。单音节覆盖 6 个舒声调以及 2 个促声调。每个调基本上取 3 个词，共 32 个词，每人共 64 个单音节样本。为使双音节声调的样本覆盖面更可能广且均匀，双音节由单音节调类组合而得，其中促声调的长元音调和短元音调调值相同，故在实际组合时只取一个。双音节共 249 个词组，每人 498 个双音节样本。

语音材料在广西武鸣镇广播站录音室室采集，录音软件为 Audition1.5，采样频率为 22050Hz。使用北京大学语言学实验室语音处理平台 SpeechLab 进行语音标志，并使用自相关法提取基频。因为舒声调和促声调成互补分布，因此将舒声调和促声调都时长归一化为 30 个点。

3. 研究方法

研究声调可以通过观查声调拟合曲线的拟合系数实现，也可以直接观查声调的基频数据分布。因为前者数据更少，更简洁，所以人们常常通过分析拟合系数来研究基频。对于双音节声调，如果将前后音节视为一个整体进行拟合，使用低于 2 阶的多项式不易精确表示其调形。如果分别对前后音节进行 2 阶多项式拟合，最终得到 6 拟合系数用于描写一个双音节声调。在拟合系数多于 3 个时，其分布落于高维空间，就不易得到直观的分布图，因此无法通过观察声调曲线拟合系数的分布来了解双音节声调的分布情况。而双音节的基频虽然数据量较大，但不同音节的基频可以画在同一个二维图之中，可以直观的了解其分布情况。因此，本文不对双音节调形进行拟合操作，而是直接观察双音节基频分布。

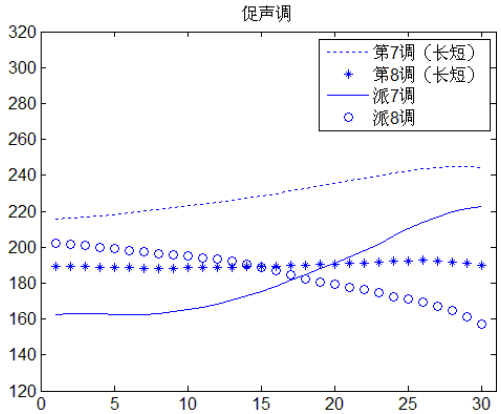
使用聚类算法对双音节声调进行聚类可以从统计上得到双音节声调的相似程度，并将相似程度高的双音节声调归为一类。传统的聚类算法需要人工指定将数据聚为几类，这增加了聚类结果的主观性。仿射传播聚类(Affinity propagation clustering, AP) (Frey and Dueck 2007)是2007年在《科学》（Science）上提出的一种新聚类算法,其思想是通过一个迭代循环计算吸引度，以产生m 个高质量的类代表和对应的聚类, 同时使聚类的能量函数最小化。AP算法有着如下优点：1、不需要指定类别数。2、求得的聚类中心是原始数据中真实存在的，而不是由多个数据点求均值而来。3、多次执行AP聚类算法，得到的结果是完全一致，不需要进行随机选取初值的过程。4、聚类结果误差小，正确率高。因此本文采用AP算法来进行聚类运算。

对于聚类算法而言，参与聚类的数据维数多少对聚类结果的无明显影响。双音节声调的拟合系数与原始基频相比，虽然参数的维数更少，且可以反映基频曲线的形状，但总是会丢失部分信息。因此本文直接针对双音节的原始基频数据进行聚类。

本文首先将 2 男 2 女的基频数据按音节作均值，得到消除发音人音域差异之后基频均值数据。然后使用聚类算法分别对每个发音人的单音节基频以及 4 个发音人单音节基频均值进行聚类处理，以验证对于调类较为稳定的单音节数据，使用聚类算法与传统归纳的方法所得的调类结果是否相同或相近。如果二者一致，则可以说明聚类算法可以较为有效的对声调进行提取调类的处理。本文在此基础之上对武鸣壮语双音节基频数据进行聚类处理，并认为聚类的结果反映了武鸣壮语双音调类的实际分布。

4. 结果及分析

武鸣壮语的舒声调有 6 种调值，促声调有 4 种调值。我们将 4 个发音人，每个单音节调类求均值，得出每个声调的调值，具体如图 1 所示。由



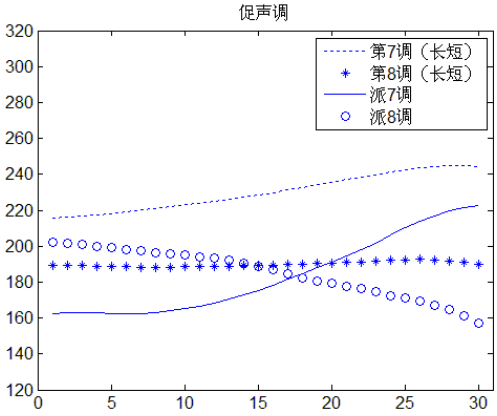


图 1 可以发现武鸣壮语的实际调值和传统方法的记音结果略有不同。本文对每个调类的具体调值同意杨峰(杨锋 2007)研究的结果，即第 1 调的实际调值为 34 而非 24，第 4 调和派 8 调的实际调值为 43 而非 42，第 6 调和第 8 调的实际调值为 44 而非 33。甚至从某种程度上说，5 度制并不足以描写实际调值的分布，比如第 6 调和第 8 调都记为 33，第 3 调和第 7 调都记为 55，但显然它们的调值还存在差异。此时，完全有可能出现人们把第 6 调的样本感知为 55，因为从其基频的分布上看，它更接近第 7 调（55）而非第 8 调（33）。

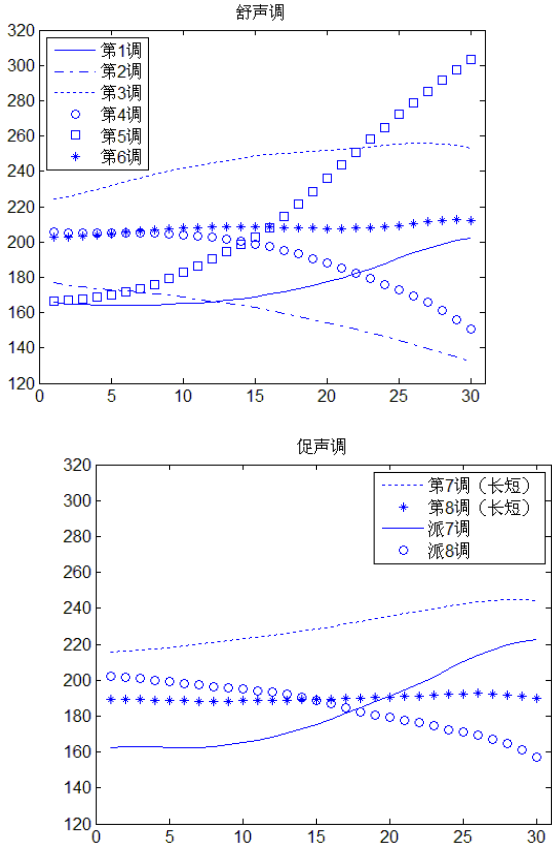


图 1 武鸣壮语单音节声调分布

针对 4 个发音人的单音节基频，其中 3 位聚类结果分为 6 类，每一类的基频分布与传统音位学所归纳的调值正好一致，由此可以说明聚类算法的结果正好与人脑对声调调值的感知结果一致。但其中的 2 号女发音人的基频数据被聚类算法分成了 7 类。经过分析，发现 55 调被分成了两类，这两类的分布具体如图 2 所示。

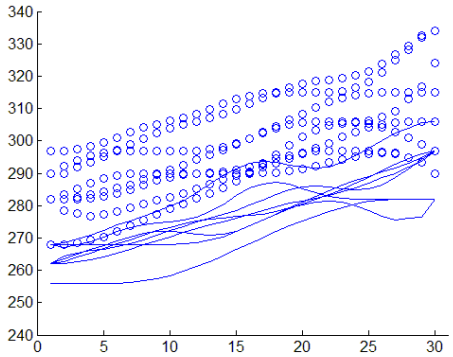


图 2 二号女发音人 55 调的聚类结果

调值为 55 的单音节主要是第 3 调和第 7 调。

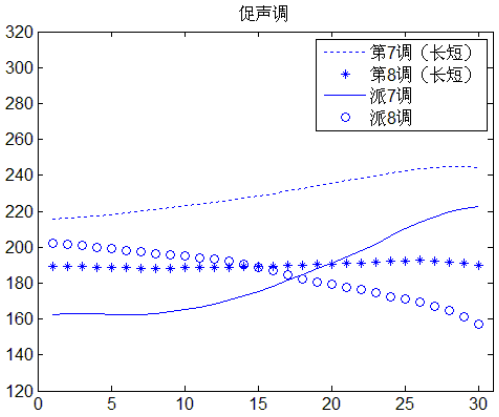


图 1 中，圆圈表示第 3 调，实线表示第 7 调。第 3 调和第 7 调虽然都记为 55 调，但第 7 调的基频要略低于第 3 调。因此，将调值为 55 的声调聚类为两类也有一定的道理。

对 4 位发音人的基频求均值，消除男女差异以及个人特征之后，对基频均值进行聚类，共分成 7 类。其中同样是调值为 55 的第 3 调和第 7 调被聚成了两类。这说明聚类结果受到了 2 号女发音人基频分布的影响。

武鸣壮语单音节共有 6 种比较稳定的调值，聚类算法对于单音节基频的聚类结果与 6 种调值的分布基本保持一致。舒声调与促声调是互补关系，我们在对基频数据进行聚类时将它们二合为一，如果从这个角度考虑，聚类算法甚至自动找出了调值相

同的第 3 调和第 7 调之间的细微差别，其聚类结果与人脑感知的调类一致，因此我们同样采用此算法对武鸣壮语双音节基频进行聚类，以此研究双音节调类的分布。

我们共有 498 个双音节样本，对四位发音人的双音节样本进行聚类，分别得到 24、25、28 和 29 种类别。其中不同主要由个人发音音域、调形等差异造成。对 4 位发音人的双音节基频求均值，消除个人特性之后，共聚为 28 类，具体如

表 1 所示。

表 1 双音节基频聚类结果

后字 / 前字	1 调	2 调	3 调	4 调	5 调	6 调	7 调	8 调	派 7	派 8
1 调	9|3	1	4|28	5	6|27	7	28|11|24	2|7	/	7
2 调	3	17	4	5	6	7|14	4	7	6|8	3
3 调	9	10	11|24	12	22|23	2|25	11|23	2	9	/
4 调	13	17	19	18	8|27	14	19	14|7	8|27	18
5 调	20	26	28	15|20	22	16	28	11|16	/	/
6 调	13|20	17	19	18	27	14|7	19	14|18|25	/	18
7 调	23|20	10|26	24	15|12	21|22|23	25	24|23	25|16	21	/
8 调	13	17|26	19	18	27	14|29	19|14	18|14|19	27	/
派 7	20	26|10	28	15	22|21	16	28	/	/	/
派 8	8|18	17	/	/	/	14	19	/	27	/

表 1 中数据表示聚类的结果，如果一种双音节声调被聚为多类，则各类之间用竖线分隔。斜杠部分表示没有此类样本。

以第 2 音节对齐，后音节与相应的前音节调值具体分布如

表 2 所示。

表 2 第二音节与第一音节声调配合表

第二音节调类	第二音节调值			第一音节调值						
第 1 调	24			22	25	32	45	55		
第 2 调	31	41		22	32	35	55			
第 3 调	55			23	32	42	44	55		
第 4 调	31	41	51	22	23	25	33	55		
第 5 调	25			22	32	34	45	55		
第 6 调	33	44		22	23	32	43	45	55	
第 7 调	35	44	55	23	32	33	44	55		
第 8 调	33	44	35	22	24	32	33	43	44	55

派 7 调	24	25		32	43	44	55			
派 8 调	32	44	53	33	55					

蔡培康(蔡培康 1987)的武鸣壮语的连读变调规律作了以下总结：标准壮语第 1 调音节在第 1、2 调音节和第 7、8 调(短元音)音节前，变读为第 3 调；第 2 调音节在第 1、2 调和第 7、8 调(短元音)音节前，变读为第 4 调；第 5 调、第 7 调(长元音)音节，

表 1 中，可以发现同一个调在作后字调时，其聚类的总数要远少于此调在作前字调时。这正是因为武鸣壮语的连读变调主要表现为前字变调。以 1 调为例，当其后字为 1 调时，前字有部分表现变为第 3 调，还有一部分保持 1 调不变，因为当前后调

表 1 中如 1 调在 2 调前没有被聚为两类，是因为语音样本覆盖面不够，正好没有选中前字变调的样本所造成的。

对于所有后音节为 1 调的双音节，我们查看其基频分布，具体如图 3 所示。

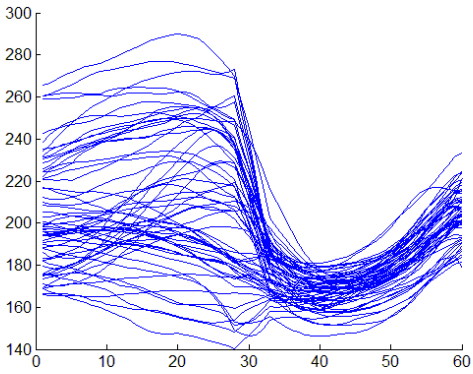


图 3 后音为 1 调的双音节基频分布图

从图 3 可以发现，后音节没有发生变调情况，这与连读变调规则相一致。因此，后音节为 1 调的双音节声调被聚类为几类最终由前字调决定。从单音节的基频分布可知，理论上 1 调与派 7 调为同一调值，应归为同一类；3 调和 7 调为同一调值，4 调与派 8 同一调值，6 调和 8 调为同一调值。但从双音节的聚类结果中可知当后音节调值相同时，这些前音节理论上调值相同的双音节并没有被聚为相同的类。在我们已经验证了聚类算法的有效性的前提之下，只可能理论上前音节调值相同的双音节，其前音节的基频调值发生了变化，最终导致理论上调值相同的双音节被聚类为不同的类。通过观察实验的基频，发现很多双音节的前字基频的值发生了变化，所以导致聚类结果与理论值不同。我们此处以

在第 5 调和第 7 调音节前，变读为第 3 调；第 6 调、第 8 调音节除在第 4 调音节前不变调外，其他变读为第 4 调；第 3、4 调音节为第 2 音节时，前一音节不变调。

从

都是 1 调时就被分成了两类。同理，1 调在 2 调、7 调和 8 调（短元音）音节前也会分为高低两类。由此，当前音节是 1 调时，其聚类总数会远多于后音节是 1 调的情况，聚类结果正好反映了连读变调现象。而

后音节为 1 调，前音节分别为 1 调和派 7 调的双音节样本为例，具体见图 4。图 4 中实线为前音节为 1 调，圆圈为前音节为派 7 调的样本。显然前音节的 1 调和派 7 调都发生了变调，故聚类算法最终没有将它们聚为一类。其它双音节声调聚类结果也同样受前音节基频变化的影响，此处不在逐一做图证明。

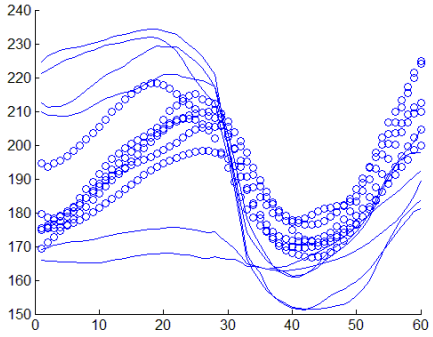


图 4 后音节为 1 调，前音节分别为 1 调和派 7 调的双音节基频分布

5. 结论

武鸣壮语的双音节调类数量较多，很难找到完全使用双音节声调为最小对立的词组。因此有些双音节调已经失去了调位的功能。按音位功能负担理论，这些失去最小对立功能的双音节调位功能负担量小，冗余性高，基频不稳定，易发生变化。以此，本文从音位负担的角度指出直接使用调位的概念来处理武鸣壮语的双音节声调的不合理性，并探讨了一种聚类分析的可行性。

若以单音节声调的组合表示双音节声调，则有 36 种双音节声调。加上连读变调形式的话，双音节声调的总类别数会更多。本文使用聚类算法，最终武鸣壮语双音节声调被分为 28 类。聚类结果远少于

理论类别数。双音节调类数量的减少的原因是某些双音节调类的实际基频与其它调类相似度很高，在聚类时被归并成一类所造成的。因此本文认为这些被归并为一类的双音节声调中，有些声调发生了变调。

虽然不能说双音节的聚类结果与调类在人脑感知的结果完全一致，但聚类结果反映的是基频的相似程度，相似度高的基频，往往在听觉感知上会比较一致。这一点从单音节的聚类结果上已经得到了证明。因此本文认为聚类结果与人脑感知的结果应比较接近，以感知实验来验证聚类结果将是我们下一步的工作。

参考文献

Frey, B. and D. Dueck (2007). "Clustering by passing messages between data points." Science **315**(5814): 972.

Wang, W. S. Y. (1967). "The measurement of functional load." Phonetica **16**(1): 36-54.

孔江平 (2005). 武鸣壮语声调的声学研究. 第 38 届国际汉藏语会议.

李方桂 (2005). 武鸣土语, 清华大学出版社.

蔡培康 (1987). "武鸣壮话的连读变调." 民族语文 (01): 20-26.

张均如 (1999). 壮语方言研究, 四川民族出版社.

杨锋 (2007). "武鸣壮语声调声学空间分布的统计分析." 河池学院学报 **27**(003): 74-78.

韦庆稳、覃国生 (1980). 壮语简志, 民族出版社.

再论圆唇的定义

潘晓声 孔江平

提要 本文目的在于讨论语音学中圆唇特征的定义问题。研究从汉语普通话录像中提取出元音发音时关键帧的唇形内外轮廓唇圆度、宽度和开口度等参数，用于分析圆唇元音、非圆唇元音与上述参数之间的关系。实验结果表明，唇圆度和唇开口度并不具备区分圆唇元音和非圆唇元音的功能，而唇宽度具有区别意义的功能。完整发音过程可以分为张嘴与闭嘴两个阶段，二者有着不同的唇形运动模式。综合此二阶段唇形的变化规律，本文主张使用外唇和内唇宽度之和特征，其变化可以表示圆展唇运动的动态过程。实验发现普通话发音的唇部生理运动幅度具有一定的随意性，但总能保持同一发音人非展唇元音比展唇元音的唇宽度更窄这一特性。

关键词 汉语普通话 圆唇 唇宽度 唇开口度 唇圆度 唇突度

1. 引言

关于圆唇的定义是语音学和音系学的一个基本问题，在各种教学参考书中都可以见到，但各家的定义并不统一，本文希望通过实验语音学的方法可以最优描述圆展唇运动的特征。

音系学研究中，有学者(Chomsky & Halle, 1968; Ladefoged, 1975)将其作为一种区别性特征，用于区分语音类别。语音学研究中，人们认为元音的音质是由舌位的前后、高低和是否圆唇所决定，因为这三者都通过对共鸣腔的形状的影响来改变元音的共振峰结构。另外，语音学家们在研究唇的协同发音时(Hardcastle & Hewlett, 1999)，也着重关注圆唇元音和非圆唇元音之间过渡时唇形的变化规则。因此对圆唇定义的混乱显然会影响研究的结果。

常用的网络工具维基百科和百度百科都定义圆唇元音为嘴唇呈圆形时所发的元音。而专业工具书中，《普通语言学纲要》(罗常培，王均, 2002)关于圆唇元音的定义是双唇撮敛成圆形，不圆唇元音是两唇舒展，成扁平形或保持自然状态；《语音学教程》(林焘，王理嘉, 1992)中指出嘴唇圆的是圆唇元音，嘴唇不圆的是不圆唇元音；而《现代汉语》(北京大学中文系现代汉语教研室, 2009)中的定义则是嘴唇向前伸，呈圆形,是圆唇元音。然而某些非圆唇元音，比如[ɳ]的唇形状也很圆，因此唇形状上呈现圆形并不能用于区分圆唇元音和非圆唇元音。此外，上述教材都明确指出圆唇元音的定义主要考虑的是唇形呈圆形，但都没有进一步说明圆唇指的是唇形外轮廓还是内轮廓呈现圆形。除了在《现代汉语》中指出圆唇元音的发音与突唇动作相关之外，另几种教材都仅仅是以嘴唇的形状是否是圆形作为圆唇元音的标准。我们知道唇突度会改变声道的长度，从语音学原理上来说，唇突度对元音性质的影响是不可忽视的，因此在定义圆唇元音时抛开突唇动作也是不合理的。另外，乔姆斯基和哈利(Chomsky & Halle, 1968)在《英语音型》一书中，专列了一项关于唇的区别性特征：圆唇和非圆唇，并指出圆唇由唇孔变窄所形成，非圆唇则否。赖福吉(Ladefoged, 1975)的区别性特征系统中也包括了圆唇化这一特征，但他用两嘴角水平宽度的倒数来定义圆唇程度。乔姆斯基和赖福吉定义圆唇特征的区别在于，前者使用唇形内轮廓的宽度，而后者使用唇形外轮廓宽度。艾伯瑞等(Abry & Boe, 1986)在研究中指出，所有语言中区分圆唇元音和非圆唇元音的参数是嘴唇宽度，即开口时的水平宽度，它和所有关于嘴唇突度的参数成反比。综上所述，前人更多从个人观察的主观经验出发对圆唇给出定义，观察的角度不同，则使用的参数也不同，唇圆度、唇突度、圆宽度和唇开口度都有人使用，因此各家的观点并不一致。

《实验语音学概要》(吴宗济，林茂灿, 1989) 中鲍怀翘提取了五位被试的唇形参数，其实验结果表明展唇元音的开口度大于圆唇元音的开口度，舌高度相同的元音，展唇元音的唇形面积大于圆唇元音。另外，作者认为唇突度特征对于区分元音圆展作用不大。但由于当时的实验条件，无法提取到发音时嘴唇的动态变化，所采集的唇形参数的精度也受实验器材和算法的限制。在科技水平有了很大进步的今天，我们可以利用数字化技术提取到精度更高的唇形参数，因此本文对唇形参数的提取和分析做进一步细化处理。

鉴于圆唇的定义对于语音学、音系学都是一个重要的基础问题，各家对于圆唇特征定义不同，会造成后学在概念上的混乱，对于语言学研究也会造成干扰。前人在圆唇的定义上各有不同的主要原因可以归为两点：1、圆唇只是一个概念，在舌位相同时人们发现圆唇的变化对元音的声学特征有影响，

因此用它作区别性特征。究竟用唇圆度还是其它参数来定义并不重要，因此没有必要精确的提取参数作分析。2、受实验条件的限制，无法提取某些参数，或提取的参数精确度不够。本文认为随着研究的深入，人们越来越关注唇形生理上的变化对声学造成的影响，而科技的发展使的我们有更好的设备和技术对唇形进行分析。

因为发圆唇元音时，唇形呈现撮拢状态，而非圆唇元音则否，因此本文假设圆唇元音比非圆唇元音的唇宽更窄，唇形更圆，唇开口度更小。通过对 8 个发音人汉语普通话圆唇元音和非圆唇元音样本的唇形参数进行统计，希望找出最符合假设的参数作为定义圆唇的特征。

2. 实验方案

2.1. 实验材料

汉语普通话的单元音中，/ə/、/ʌ/和/ɑ/都不能单念，一定要和辅音相配合才能读，且配合不同发音部位的辅音时，这几个元音所表现的唇形不同，即元音唇形受到辅音发音动作的影响，因此将其排除。而/i/的两个变体/ɿ/和/ʅ/也必须和辅音相结合才能发音，但/ɿ/只能与舌尖前辅音/ts/、/ts^h/和/s/相配合，/ʅ/必须要与舌尖后辅音/tʂ/、/tʂ^h/、/ʂ/和/ʐ/相配合，即它们受辅音的影响相同，且不存在不受影响的形式。因此我们最终选择 8 个汉语普通话零声母元音和两个非零声母单元音/ɿ/和/ʅ/进行唇形研究。其中元音[ɻ]和[ɳ]分别配合声母[ts]和[tʂ]发音。具体见表 1。

表 1 汉语普通话元音的国际音标及对应汉字和拼音标注

国际音标	汉字	汉语拼音
ʌ	阿	a
ɤ	饿	e
ɛ	诶	e
ɚ	儿	er
i	衣	i
o	哦	o
u	屋	u
y	淤	ü
ʅ	之	zhi
ɿ	资	zi

为了使样本来源具有更广泛的代表性，而不仅仅是代表受过专业语言学训练的发音人，我们的 8 位发音人中并没有全都受过语言学专业训练，但所有被测都受过高等教育。

此外，不是所有发音人都认识国际音标，因此实验不要求发音人读这些元音所对应的国际音标，而是给出这些音节所对应的汉字，并求发音人按拼音方案中的标音朗读。发音人具体朗读的汉字以及拼音标音见表 1。

8 位发音人。每人朗读所提示的汉字各三次，共得到 270 段录像。为更精确的提取唇形参数，我们要求发音人正面朝向录像机，在发音时头部基本保持静止不动，以免得到的参数受点头运动和发音人身体晃动的影响。发音从闭唇状态开始，发音完成后最终恢复到闭唇状态，在两个音节之间停顿 1 到 2 秒，由此可以尽量减少前后音节的协同发音影响。发音持续一段时间，以保证可以从中提取至代表此元音发音生理动作的唇形关键帧。

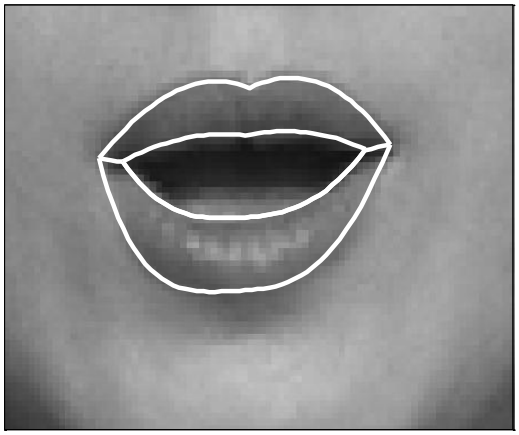
实验数据的采集在上海师范大学语言研究所的专业隔音室录像室中进行。实验使用录像机采集发音人的唇形数据，录像速率为 30 帧每秒。

2.2. 唇形参数提取

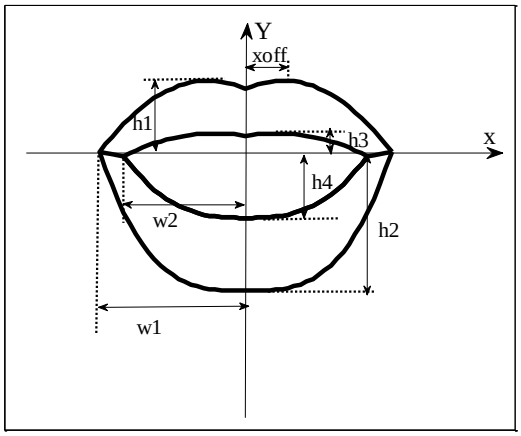
本实验采用的是北京大学中文系语音乐律实验室二维唇形处理平台，该平台在 Windows 平台下用 Matlab 语言编程实现。

由于嘴唇和舌头的颜色相近，目前国际上流行自动提取图像中物体轮廓线的活动轮廓模型法（Active Shape Model）无法精确的自动提取唇形的内轮廓，因此我们采用了半自动方法手工提取唇形轮廓。

我们在刘伟聪的算法(Liew, Leung, & Lau, 2000)的基础之上，加入了描述唇形内轮廓的参数，建立了一个二维唇形几何模型，具体见图 2，本文称所有这些用于重构唇形内外轮廓的参数称之重构参数。



(a)使用唇模型重构的唇形轮廓



(b)去随伴随动作后重构的唇形

图 1 唇轮廓重构图和唇模型

图 1（a）是通过手工改变唇模型的重构参数所重构的唇形轮廓。可以发现，用唇模型能精确描述唇形内外轮廓。

人们发音时总是不自觉的会带有歪嘴和头部倾斜等伴随动作，因此图 1（a）中的两嘴角高度并不处在同一水平线上。头部倾斜幅度越大，两嘴角坐标的垂直距离就相差越大，两嘴角坐标的水平距离也会相应的变小，此时两嘴角之间的水平宽度并不能真正代表唇宽度。为避免发音时伴随动作幅度的不同而产生唇形参数之间的不可比性，我们在提取唇形参数时，将表示伴随动作的参数值归零，即去除了表示歪嘴和头部倾斜的伴随动作。图 1（a）去除伴随动作后重构的唇形如图 1（b）所示。

重构参数共有 11 个，具体包括：左侧外唇宽度（w1）、左侧内唇宽度（w2）、上唇外轮廓开口度（h1）、下唇外轮廓开口度（h2）、上唇内轮廓开口度（h3）、下唇内轮廓开口度（h4）、人中凹陷程度（xoff）、下唇圆弧度（□）、唇闭合处曲率（ql）、头部倾斜程度（qx）和歪嘴程度（s）等。

从图 1（b）中可以提取去除伴随动作之后的唇形轮廓几何特征，本文称之为唇形几何特征参数。实验所提取的唇形几何特征参数共有 10 个，分别为：1、外唇水平宽度：外唇两嘴角之间的水平距离，宽度为两倍的左侧外唇宽度（w1）。2、内唇水平宽度：内唇两嘴角之间的水平距离，宽度为两倍的左侧内唇宽度（w2）。3、外唇开口度：唇形外轮廓在垂直方向上的最大距离，高度为唇外轮廓开口度（h1）和下唇外轮廓开口度（h2）之和。4、内唇开口度：嘴唇内轮廓在垂直方向上的最大距离，上唇内轮廓开口度（h3）和下唇内轮廓开口度（h4）之和。5、外唇面积：嘴唇外轮廓所包含的区域内像素点的总数。6、内唇面积：嘴唇内轮廓所包含的区域内像素点的总数。7、外唇周长：嘴唇外轮廓线上的像素点总数。8、内唇周长：嘴唇内轮廓线上的像素点总数。9、外唇圆度：嘴唇外轮廓接近圆形的程度。10、内唇圆度：嘴唇内轮廓接近圆形的程度。

一些文献在假设唇形是一个椭圆的前提下，用椭圆的长短轴之比来表示唇圆度。但严格来说，唇形并不是椭圆，而是一个更加复杂的几何形状。如果仍然使用长短轴之比表示唇圆度的话就不够精确。比如正方形 A 和圆 B，如果 A 的边长等于 B 的直径，则二者的宽高比相同，但它们的圆度完全不同。因此，本文采用数学上定义圆形度的方法来表示唇圆度(Baddeley, et al., 2009)：

圆形度 = 4≺*面积/周长的平方

形状越圆，圆形度的取值越大，当唇形为绝对圆形时，圆形度达到最大值 1。

具体提取唇形参数的步聚如下：首先由手工从中录像中提取可以代表元音音节发音时的嘴唇轮廓形状的关键帧；然后手工调整重构参数，通过目测使重构的唇形轮廓和关键帧的嘴唇内外轮廓线之间的误差最小；最后将代表伴随动作的重构参数 s 与 qx 置零后再次重构唇形轮廓，并从中提取唇形几何特征参数。

2.3. 实验方案

每一个人唇形宽窄厚薄都由于生理条件的差异而各有不相同，另外发音的动作习惯也都有所区别。因此在发不同元音时，其唇形轮廓参数具有个人生理特征，不同发音人的唇形等参数不能相互比较。

实验 1：本文假设圆唇元音比非圆唇元音的唇宽更窄，唇形更圆，唇开口度更小，分别对 8 个发音人录像样本进行统计分析。通过比较同一发音人的圆唇元音和非圆唇元音发音特征的唇形参数，以找出其中的一个或多个作为圆唇的区别特征。

实验 2：本文通过统计张嘴与合嘴阶段内外唇参数的变化范围，来说明嘴唇发音动作在不同发音阶段有着不同的运动模式。并综合两个发音阶段的规律，提取出适合描述圆展唇动作动态过程的特征。

实验 3：本文通过计算得到各个发音人的轮廓参数的最小值和最大值，二者之差就是此发音人发音时轮廓参数的变化范围。将此变化范围做归一化处理，可以消去不同发音人嘴唇轮廓参数的个人生理特征，但保留了个人的发音动作习惯。从归一化之后的唇形参数分布可以说明发音人的发音动作习惯是否具有共性。

3. 实验结果及分析

以外唇宽度为例，每个发音人的 9 个圆唇元音样本大于 21 个非圆唇元音样本的情况总共可能出现 189 次。本文对每个发音人所读的圆展唇元音样本的 6 个唇形参数分别进行圆唇元音和非圆唇元音间的比较，统计出每个参数实际满足实验假设的次数以及其占最大满足假设次数的百分比，结果如表 2 所示。

表 2 中 W1、W3、R1、R3、H1 和 H3 分别表示非圆唇元音的外唇水平宽度、内唇水平宽度、外唇圆度、内唇圆度、外唇开口度和内唇开口度；W2、W4、R2、R4、H2 和 H4 分别表示圆唇元音的外唇水平宽度、内唇水平宽度、外唇圆度、内唇圆度、外唇开口度和内唇开口度。括号外的是每个参数实际满足实验假设的次数，括号内的为每个参数实际满足实验假设的次数占最大满足假设次数的百分比，结果四舍五入取整数。

表 2 圆唇元音和非圆唇元音唇形参数比较

	W2<W1	W4<W3	R2>R1	R4>R3	H2<H1	H4<H3
发音人 1	189/100	186/98	77/41	61/32	160/85	182/96
发音人 2	189/100	188/99	154/81	88/47	139/74	167/88
发音人 3	189/100	180/95	142/75	152/80	115/61	115/61
发音人 4	188/99	185/98	102/54	149/79	120/63	144/76
发音人 5	180/95	185/98	74/39	125/66	127/67	171/90
发音人 6	187/99	188/99	169/89	132/70	87/46	137/72
发音人 7	189/100	188/99	164/87	144/76	99/52	157/83
发音人 8	189/100	184/97	83/44	146/77	153/81	154/81
平均	188/99	186/98	121/64	125/66	125/66	153/81

从表 2 可得，平均 99%的样本，外唇宽度满足实验假设。平均 98 的样本，内唇宽度满足假设，其它唇形参数最多只有 81%的样本满足实验假设。实验 1 结果表明外唇宽度和内唇宽都适合作为圆唇的区别特征，而内外唇圆度和开口度无法有效区分圆唇元音和非圆唇元音。

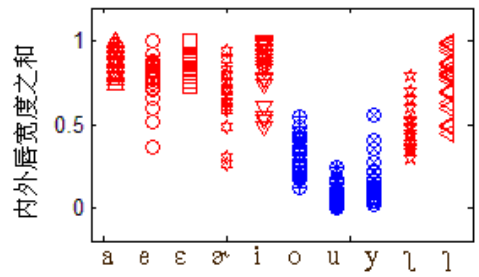
为描述圆展唇运动的动态过程，完整发音过程可以分为合嘴与张嘴两个阶段。合嘴阶段包括发音的准备动作阶段和发音完成后的复位阶段，此时内唇宽度、圆度和开口度值都恒为 0，外唇宽度则随着圆展唇运动相应变化。外唇宽度的变化在合嘴阶段能有效的描述嘴唇发音动作。张嘴阶段主要是有声段，所有唇形样本的关键帧均来自于有声段。实验 2 对每个发音人圆唇元音和非圆唇元音的内外唇宽度的变化范围进行统计平均，结果如表 3 所示。由表可知圆展唇元音内唇宽度的平均变化范围远大于外唇宽度的平均变化范围。另外，有声段语音的声学特性和内唇的开口形状相关。因此，在张嘴阶段内唇宽度的变化更适合表示圆唇展运动。综合考虑发音的两个阶段，本文提出使用内外唇宽度之和为特征，其变化可以良好的描写圆展唇运动的动态过程。

表 3 发音人圆唇元音及非圆唇元音内外唇宽度的平均变化范围（单位：像素）

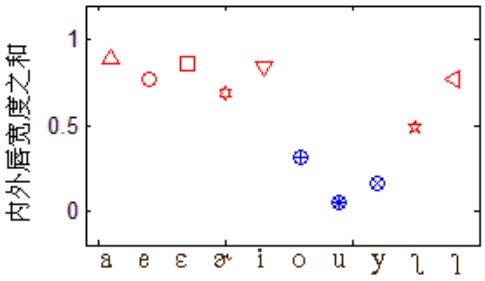
发音人	内唇宽度平均变化范围	外唇宽度平均变化范围
1 号	37	11
2 号	52	17
3 号	40	13
4 号	52	17
5 号	41	8
6 号	43	17
7 号	49	23
8 号	39	9

个人生理特征和发音动作习惯的不同造成了不同发音人的唇形参数不能直接相互比较。实验 3 使用内唇宽度和外唇宽度之和为特征，对其变化范围进行归一化处理，以消除不同发音人嘴唇宽窄各有不同的个人生理特征，并保留了个人的发音动作习惯。归一化之后，内外唇宽度之和的分布如图 2（a）所示，对所有发音人的元音样本唇宽参数取均值，其分布如图 2（b）所示。

从图 2（a）可知每一个元音的内外唇宽度之和变化范围大，说明不同发音人的发音习惯有唇部的发音动作具有一定的随意性，有人发音动作幅度较大，有人发音动作幅度较小。而从图 2（b）可知，虽然发音动作具有随意性，但总体上保持非圆唇元音的内外唇宽度之和大于圆唇元音。



（a）所有发音人元音样本



（b）所有发音人元音样本均值

图 2 内外唇宽度之和进行归一化处理之后的分布(单位：像素)

由图 2（a），圆唇元音/o/、/y/和非圆唇元音/ɭ/的许多样本不满足圆唇元音唇宽度小于非圆唇元音的假设。每人每个元音读 3 次，因此圆唇元音/o/和/y/对于非圆唇元音/ɭ/最大满足假设 9 次。本文对每个发音人分别统计/o/、/y/的内外唇宽度之和小于/ɭ/实际所出现的次数有占最大满足假设次数的百分比，结果如表 4 所示。

表 4 每个发音人/o/、/y/和/ɭ/样本的唇宽关系

发音人	/o/</ɭ/	/y/</ɭ/
1 号	9（100）	8（89）
2 号	9（100）	9（100）
3 号	2（22）	9（100）
4 号	9（100）	9（100）
5 号	9（100）	9（100）
6 号	9（100）	9（100）
7 号	9（100）	9（100）
8 号	8（89）	9（100）

由图 2（a）及表 4 可知，虽然由于发音动作习惯不同，不同人发的圆唇元音的内外唇宽度之和可能大于非圆唇元音，但同一发音人的发音动作有着自我调节功能，基本可以保证相同发音人的圆唇元音内外唇宽度之和一定小于非圆唇元音。本实验中，只有 4 号发音人发/o/时唇形较宽，其多数圆唇元音样本的内外唇宽度之和大于非圆唇元音。

4. 讨论

本文提出圆唇的本质是唇宽变窄，使用内外唇宽度之和为特征可以描述圆展唇的运动过程，发音动作幅度有着较大的随意性，但同一发音人总会保持圆唇元音的唇宽小于非圆唇元音的唇宽。

一些文献使用唇突度为圆唇的特征，本文不对唇突度进行讨论有如下原因：1、使用录像机采集侧面人脸提取唇突度的方法有着较高的技术难度。其它三维数据采集设备虽然可以实时采集唇突度，但价格昂贵，数据预处理复杂，距大范围应用还有一定距离。2、本文提出以内外唇宽度之和为特征已经可以良好的描述圆展唇运动。3、鲍怀翘已经研究证明唇突度对于区分元音圆展作用不大。

由于提取唇形关键帧没有一定的标准，因此本文所有的数据及推论都是建立在我们对关键帧的提取标准之上，即元音共振峰稳定段中，发音动作基本不变时的某一帧。在唇形录像样本中提取关键帧的位置不同，会影响采集到的唇形几何特征，导致实验结果发生变化。而事实上，如果没有事先作保持发音动作一段时约定，我们会发现在元音共振峰的稳定段中，发音动作仍然处于变化之中，即发音的生理数据并不如所想象的到发音动作目标后会保持一定的稳定时间。

王志明(王志明，蔡莲红, 2002)使用算法来自动提取唇形关键帧，可以实现大量数据的快速处理，我们的方法则是对语音的共振峰和录像数据进行人工判断，由此决定关键帧的位置。他们的方法好处在于提取的所有关键帧都是按相同的标准所得。我们的方法优点在于更多的利用了语音学的知识，缺点是同时也具有更多的主观性，在唇形相似的相邻帧中，选择关键帧往往具有随意性。另外，手工提取关键帧速度较慢，无法做到大量快速提取。

由于圆唇特征定义的混乱，许多关于唇的研究都没有一个统一的标准，这造成研究成果之间不具备相互比较的基础。本文对 6 种唇形参数进行统计分析后，分别指出其单独做为圆唇特征的不足之处，并提出使用内外唇宽度之和作为圆唇特征，可以区分圆展唇元音，其变化能描述动态的圆展唇发音动作。

5. 总结

语音学还是一门不很成熟的学科，很多术语的概念也没有形成多数人接受的共识。对于圆唇的定义也是如此。本文通过实验采集各种唇形参数，利用统计的方法，提出圆唇的本质是唇宽度的变窄，而非嘴唇的形状变圆。因此，本文提出（非）展唇

元音比（非）圆唇元音更合适描写圆唇这一特征。在生理上，发音是一个动态过程，本文提出内外唇宽度之和的变化可以描写圆展唇发音动作的变化。此外，虽然普通话发音时，唇部发音动作具有较强的随意性，但总能保持同一发音人非展唇元音比展唇元音的唇宽度更窄这一特性。

参考文献

Abry, C., & Boe, L.-J. (1986). "Laws" for lips. *Speech Communication, 5*(1), 97-104.

Baddeley, D., Jayasinghe, I., Lam, L., Rossberger, S., Cannell, M., & Soeller, C. (2009). Optical single-channel resolution imaging of the ryanodine receptor distribution in rat cardiac myocytes. *Proceedings of the National Academy of Sciences, 106*(52), 22275.

Chomsky, N., & Halle, M. (1968). The sound pattern of English.

Hardcastle, W., & Hewlett, N. (1999). *Coarticulation: Theory, Data and Techniques*: Cambridge Univ Pr.

Ladefoged, P. (1975). *A course in phonetics*: Harcourt Brace Jovanovich New York.

Liew, A. W. C., Leung, S. H., & Lau, W. H. (2000). *Lip contour extraction using a deformable model.* Paper presented at the Image Processing, 2000. Proceedings. 2000 International Conference on.

王志明，蔡莲红 (2002). 汉语语音视位的研究. *应用声学, 21*(003), 29-34.

北京大学中文系现代汉语教研室 (2009). 现代汉语(重排本)

林焘，王理嘉 (1992). *语音学教程*: 北京大学出版社, 北京.

吴宗济，林茂灿 (1989). 实验语音学概要. *北京: 高等教育出版社.*

罗常培，王均 (2002). *普通语音学纲要*: 商务印书馆.