Dynamic Modeling for Chinese Shaanxi Xi'an Dialect Visual Speech

MISKOLCI EGYETEM

UNIVERSITY OF MISKOLC

Lu Zhao

A dissertation submitted for the degree of

Doctor of Philosophy

Scientific adviser:

Dr. László Czap

Faculty of Mechanical Engineering and Informatics

József Hatvany Doctoral School for Information Science

University of Miskolc

Miskolc, 2019

Declaration

The author hereby declares that this thesis has not been submitted, either in the same or in a different form, to this or to any other university for obtaining a PhD degree. The author confirms that the submitted work is her own and appropriate credit has been given where reference has been addressed to the work of others.

Signature of thesis author: Lu Zhao

Date: March 1, 2020

Acknowledgements

This research was completed with the careful guidance and help of the scientific adviser László Czap. Here, I would like to express my sincere gratitude! László Czap's profound professional knowledge, sincere attitude, rigorous academic attitude and professionalism have set a good example for me as a researcher and a model for me to learn. I would like to especially thank him for carefully selecting the direction of the doctoral thesis for me and carefully guiding my research process. I am thankful for József Hatvany Doctoral School, my doctoral school for providing instruments and a caring environment to my work.

I would like to thank Tamás Gábor Csapó from Eötvös Loránd University and speech processing lab for providing instruments to record the Chinese Shaanxi Xi'an dialect corpus, Robin Nagano from the University of Miskolc for helping me in English and Director Jie Liu from the Confucius Institute of University of Miskolc for giving me help with life in Hungary.

I would like to thank also my fellow PhD students and friends, especially Maen Alzubi, Mohammad Alsaudi, Mohammad Alsharif, Ahmed Bouzid, Tati Wadas, Rosa Alamian, Nikolina Mijic, and Thien Hoang for their valuable comments and discussions.

I am also thankful for my former work unit, Chairman Zhou Yanbo of Xi'an Siyuan University, and the dean of Liu Zhongxuan, the second-level college of engineering, and the two deputy directors, Yang Yan and Jia Xian.

I am also grateful to the experts and professors as reviewers of this paper, for their valuable advice and recommendations.

And last but not least I would like to thank the patient and loving support of my husband Sen Xiong , my daughter Zihan Xiong and my parents.

The described study was carried out as part of the EFOP-3.6.1-16-00011 "Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialization" project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

Abstract

A talking head is an animated articulation model that can be utilized, for instance in speech assistant systems for hard-of-hearing children or training for second language learners.

In this thesis, I create articulation features and a dynamic modeling system for visual representation of speech sounds for a Chinese Shaanxi Xi'an Dialect talking head. Firstly, I create a phonetic alphabet of the Chinese Shaanxi Xi'an Dialect and show the relationship of the phonemes of the dialect with Mandarin. I show the X-SAMPA code derived for the Chinese Shaanxi Xi'an Dialect based on the Hungarian X-SAMPA code, in addition to the correspondent regularities for whole syllable pronunciation. Secondly, I display the static lip viseme classification of Chinese Shaanxi Xi'an Dialect speech and analyze lip viseme parameters by processing images and videos of different lip visemes. I describe another experiment carried out to study both the timing and position properties of articulatory movements of the tongue in utterances recorded during dialect speech. The parameters derived from both of these two experiments are settled to create a dynamic viseme modeling system. Then I describe the interaction between phonemes and corresponding lip visemes. I explain the dynamic lip and tongue viseme classification that I created based on the dominance classification conception. Finally, I analyze the impact of the viseme parameters at different tempos. I define the uttered viseme and show the dynamic viseme modeling system of lip and tongue for a Chinese Shaanxi Xi'an Dialect talking head.

Content

List of Tabl

Y

List of Figur

Y

Introduction

Mandarin, the official language of the People's Republic of China, is widely studied in basic linguistics research and in speech synthesis and speech recognition technology. However, China is a multi-ethnic country and various minority languages and dialects are spoken in modern Chinese society. There are various studies relevant to the current topic. For instance, one study has looked at phonetic conversion from Mandarin to the Min dialect of Taiwanese, along with mixed speech synthesis in Chinese with English [1]. Another investigation of speech synthesis involves dialects focused on Tibetan, using a computer readable SAMPA scheme for conversion of text [2]. The Lanzhou, Liaocheng, Shenyang, and Tianjin dialects of Chinese have also been represented in speech synthesis [3]. However, there are no researchers in China working on the transfer from the phonemes to X-SAMPA code for Chinese Shaanxi Xi'an Dialect. This work supplies the basic foundation for speech synthesis of the Chinese Shaanxi Xi'an Dialect.

At present, there are no researchers in China focusing on 3D talking head modeling and animation of Chinese Shaanxi Xi'an Dialect. Therefore, work on a 3D talking head modeling and animation of this Dialect has great significance. This thesis provides a method that could be used in visual speech synthesis and prosody modelling for the Chinese Shaanxi Xi'an Dialect and more generally for speech conversion to other languages.

Xi'an was the capital of 13 dynasties in ancient China and it still occupies quite an important position in northwest China. The Shaanxi Xi'an Dialect (also known as the Qin language), with a history of three thousand years, is extensively used by 38 million people in the Chinese Shaanxi Xi'an area, with minimal articulation differences occurring in different regions of Shaanxi province. Chinese Shaanxi Xi'an Dialect is the direct successor of the ancient Chang'an dialect, but also the representative dialect of the Guanzhong area of the central plain Mandarin area [4], and thus has great important research value. It differs from Mandarin in vocabulary, grammar and especially in articulation. For further visual speech synthesis of this Dialect it is crucial to establish a transcription system labeling the phonetic information of the Chinese Shaanxi Xi'an Dialect.

The purpose of the whole thesis is creating the fundamentals of a talking head –an animated articulation model – for the Chinese Shaanxi Xi'an Dialect [5]. Figure 1 shows the structure of the Chinese Shaanxi Xi'an Dialect talking head system. The X-SAMPA code created in the thesis for the consonants and vowels of Chinese Shaanxi Xi'an Dialect is used to create visemes. The viseme library also contains a dominance model for the Chinese Shaanxi Xi'an Dialect talking head [6]. Viseme classifications was assisted by obtaining X-SAMPA codes of consonants (C) and vowels (V) and studying their regularities of C and V in the whole-syllable pronunciation of the Dialect. I began by carrying out a static viseme classification of Chinese Shaanxi Xi'an Dialect speech by the method for classifying Mandarin (Standard Chinese) static visemes. Then

we carried out an experiment to study both the timing and position properties of articulatory movements of the lip and tongue in VCV and CVC utterances of Chinese Shaanxi Xi'an Dialect speech at different tempos.

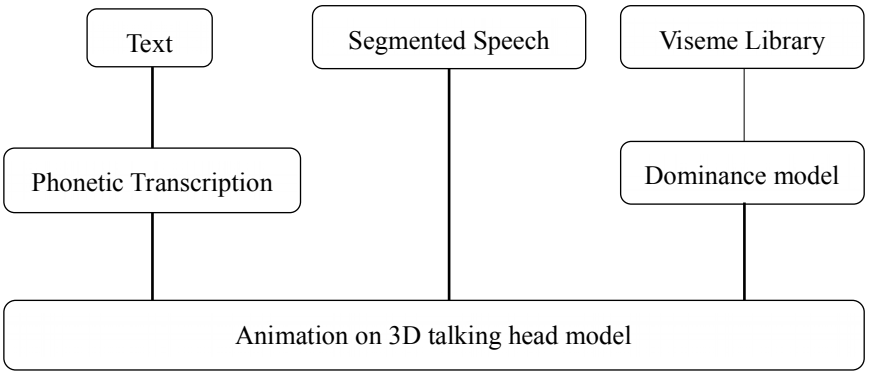


Figure 1. Structure of the Chinese Shaanxi Xi'an Dialect talking head system

The speech assistant (SA) system being developed will highly rely on visual modality, especially on the visual representation of tongue movement, which is hardly observable in real conversation. As human speech perception involves both visual and auditory modalities, it is clearly multimodal, and the conditions of speech determine which modality has more effect [7]. Various studies have examined the development of normally hearing children in comparison with deaf or blind children and have found that insufficient exposure to stimuli in both modes has a substantial effect on the ability to perceive and produce speech [8]. The audiovisual mode is more effective in transmitting articulatory features than any form of unimodal communication [9]. It has been proved by a number of clinical and laboratory investigations that combined auditory-visual perception yields better results than perception through one mode alone, and this has been found true for normal-hearing and hearing-impaired children and adults alike [10]. People understand speech better when they can see articulators like the lips, jaw, tongue tip and teeth, as well as the face. Thus, visual speech is an important aspect of speech perception, especially for deaf or hard-of-hearing people but also for normally hearing people in noisy surrounding [11]. Visual speech is studied and utilized in the fields of speech recognition [12], speech processing [13] [14], audio-visual speech synthesis [15], virtual talking head animation [16] and lip or tongue synchronization [17].

There has been a significant interest in the area of visual speech synthesis recently. It is essential that realistic face synthesis is more and more used for successful animation, film dubbing, computer avatar, video conferences, web-based automatic newsreaders, etc [18]. It is generally known that multimedia, particularly animation, plays an important role in language learning. It has made significant contribution to the language learning process among various age groups of learners, particularly a 3D animated talking head of virtual teachers in computer-assisted language learning applications. 3D animated talking head may be an essential instructional tool in supporting language learning through articulation modeling among non-native speakers [19]. The rapidly developing capabilities in computing, 3D modeling and animation have contributed to the visualization tools that can be utilized, as audiovisual talking heads can display a human-like face while also making internal articulators visible. A talking head labeled Baldi was used for

computer-assisted articulation training (CAPT) by Massaro and Cohen, who employed it as a tool inspeech therapy and second language learning [20]. They went on to compare the effectiveness of instruction in phonetic contrasts between languages through illustrations of the processes taking place in the oral cavity along with an external view of Baldi's face [21]. Badin et al. attempted to use MRI and CT data to configure 3D tongue positions and forms. Their corpus consisted of sustained articulations from a single subject speaking French. With this, they developed a linear articulatory tongue model [22] that was later built into an audiovisual talking head that was able to display the normally hidden articulators (tongue, velum) during articulation [23]. A 3D talking head was proposed by Fagel et al. as a tool for speech therapy; it was capable of making a large variety of synthesized utterances for visualizing articulatory movements inside the oral cavity [24]. Wik and Engwall described how intra-oral articulations displayed in animation were able to contribute to the perception of speech [25]. A synthetic talking head using computer animation to illustrate the facial motions of lips, the jaw and the tongue with was utilized in training in speech perception and production by Beskow et al. [26]. The talking head developed in Hungary, at the University of Miskolc provides the audio-visual representation of the speech process by visually displaying articulation, integrated in an educational framework that assists the speech teaching of hard of hearing children. The 3D transparent head model can simulate the motion of the tongue better than a human can show [27][28].

I applied the dynamic viseme modeling system of Chinese Shaanxi Xi'an Dialect to the speech assitant system of Hungarian developed in University of Miskolc to form the new speech assistant system of Chinese Shaanxi Xi'an Dialect. Figure 2 is the interface of speech assistant system of the Hungarian speech assistant system.

On the right-hand side of Figure 2 the transparent talking head is shown from two different views. In either or both windows the head can be displayed from 45° and 90° degree angle views, enabling comparison of the articulation in two separate phases of the same word or sentence. In the Hungarian speech assistant system, beside the visualization of the lips and tongue movements, an additional speech sound visualization technique helps in developing speech production. In the center of Figure 2 the bar chart represents the visualized reference sound (bottom) and the recorded sound of the trainee (top). The pointer of the bar chart and that of the Talking Head can be moved in parallel from one picture frame to another, thus the special educational needs (SEN) teacher can associate each sound graph with its articulation position. The Speech Assistant system proved to be a beneficial aid in individual speech therapy of hard of hearing pupils. The pedagogically planned methodology makes the speech therapy complete. I was able to directly apply the Chinese Shaanxi Xi'an Dialect in the prosody mode in Hungarian speech assistant system due to the prosody mode being language independent as showed in the upper bar chart of Figure 2.

Figure 2. Sample image of Hungarian speech assistant system with transparent face talking head and bar chart

In the SA the bar chart and the talking head visually represent the speech signal and articulation, realizing sensor bridging between the modalities [29][30]Error: Reference source not found. "The sensory information is transformed to an appropriate sensory modality in a way that the user can process it effectively." [32] In both hearing-impaired and normal- hearing people acoustic and visual signals are integrated by the brain. The degree of sensor sharing of modalities depends on the grade of hearing impairment. The more severe the hearing loss, the more the subjects rely on the visual modality. For profoundly deaf people the visual representation of speech can be considered as a sensory substitution Error: Reference source not found.

An expert system has been designed that aims at extending the Speech Assistant to ease the work of the SEN teacher for the hearing impaired as well as to assist those practicing on their own Error: Reference source not found. Through automatic assessment of articulation, the system recommends the next word for practicing that can be most easily uttered built upon the sounds and sound connections already pronounced correctly. Thus, a scheme for individual development can be planned that takes linguistic, acoustic and phonetic knowledge and regularities into consideration. In this way the SA can adapt the order of speech items to the cognitive abilities and the current speech production level of the trainee.

A detailed description of phonetic aspects of the Chinese Shaanxi Xi'an Dialect is illustrated in Chapter 1. Chapter 2 focuses on visemes of the Chinese Shaanxi Xi'an Dialect. The dynamic modeling system of the Chinese Shaanxi Xi'an Dialect is introduced in Chapter 3. I then summarize our progress towards achieving a talking head for the Chinese Shaanxi Xi'an Dialect based on these three chapters.

Chapter 1

Phonetic Aspects of the Chinese Shaanxi Xi'an Dialect

1.1. Phonetic alphabet of the Chinese Shaanxi Xi'an Dialect

The phonetic alphabet of the Chinese Shaanxi Xi'an Dialect has been created based on the Chinese Pinyin Scheme, which is the official Romanization system for Mandarin. It shares many phonemes with Mandarin and supplements some special phonemes and tone types.

1.1.1. The expression of the Chinese Pinyin scheme for Mandarin

Chinese Pinyin, often abbreviated to Pinyin, is the official romanization system for Standard Chinese in mainland China and to some extent in Taiwan. It is often used to teach Standard Mandarin Chinese, which is normally written using Chinese characters. The system includes four diacritics denoting tones. Pinyin without tone marks is used to spell Chinese names and words in languages written with the Latin alphabet, and also in certain computer input methods to enter Chinese characters [35]

Table 1. 1. Romanized phonetic alphabet of Mandarin

23 consonants					
Type	Unaspirated	Aspirated	Nasal	Voiceless fricative	Voiced fricative
Bilabial	b	p	m		
Labiodental				f	
Alveolar	d	t	n		l
Velar	g	k		h	
Palatal	j	q		x	
Dental sibilant	z	c		s	
Retroflex	zh	ch		sh	r
	w, y				
36 vowels					

23 consonants					
Type	Unaspirated	Aspirated	Nasal	Voiceless fricative	Voiced fricative
6 simple vowels		a, o, e, i, u, ü			
14 compound vowels		ai, ao, ei, ia, iao, ie, iou, ou, ua, uai, üe, uei, uo, er			
16 nasal vowels	8 front nasals	an, en, ian, in, uan, üan, uen, ün			
	8 back nasals	ang, eng, iang, ing, iong, ong, uang, ueng			

Note 1: the syllable 'wu' is pronounced as the Pinyin 'u' (the letter 'w' is in this case used to mark the beginning of a new syllable); The syllable 'yi' is pronounced as the Pinyin 'i' and the syllable 'yu' is pronounced as the Pinyin 'ü' (the letter "y" is in these cases used to mark the beginning of new syllables).

21consonants are seen in some classification without the phonemes 'w' and 'y'.

Note 2: The 'dialect' in all tables refers to the Chinese Shaanxi Xi'an dialect.

The Chinese Pinyin scheme for Mandarin consists of 56 basic phonemes, 23 consonants and 36 vowels. Combinations of consonants and vowels plus the special cases result in 413 possible combinations. Applying the four tones types of Mandarin Chinese to this, we get a total of around 1,600 unique syllables. Table 1. 1 presents the complete expression of this scheme [36] [37].

1.1.2. Phonetic alphabet of the Chinese Shaanxi Xi'an Dialect

The phonemes of the Chinese Shaanxi Xi'an Dialect are described by IPA, and there are 26 consonants and 40 vowels [38][39][40]. The 26 consonants include the 21 consonants (excluding w, y) of Mandarin and 5 unique consonants of Chinese Shaanxi Xi'an Dialect. The 40 vowels include 27 vowels of Mandarin and 13 unique vowels of the Chinese Shaanxi Xi'an Dialect. Its five unique consonants are presented in Table 1. 2 with gray background: 'pf' and 'pf^{h}' are both labiodental plosive fricative consonants, 'v' is a voiced labiodental fricative consonant, 'ŋ' is a velar nasal consonant, and 'ɲ' is a retroflex nasal consonant [41][42]. We compare the consonants of these two languages in Table 1. 2.

Table 1. 2. Consonants alphabet of Dialect and comparison with the Chinese Pinyin Scheme

Phonetic alphabet(IPA)		*Pinyin expression*	*character*	*Description*
Dialect	*Mandarin*			
p	p	b	玻	Voiceless bilabial plosive, unaspirated; It is articulated with airflow obstructed in the mouth, lips are closed then suddenly opened with the air released by light and short pronunciation.
p^{h}	p^{h}	p	坡	Voiceless bilabial plosive, aspirated; The airflow hindered with closed lip then suddenly a burst of

				air is released.
m	m	m	摸	Bilabial nasal; It is articulated with the air passing into the nasal cavity, lips are closed and the tongue is pulled backward, vocal cords are vibrated.
f	f	f	佛	Voiceless labiodental fricative; It is articulated with the upper teeth touching the lower lip to form a slit, the airflow to be squeezed out of the gap.
t	t	d	得	Voiceless alveolar plosive, unaspirated (blade-alveolar); It is articulated with the tip of the tongue against the upper alveolar ridge to block the air in the mouth, then air released in the mouth, the sound is formed of a burst of air.
t^h	t^h	t	特	Voiceless alveolar plosive, aspirated (blade-alveolar); It is articulated with the tip of the tongue against the upper alveolar ridge to block the air in the mouth, a burst of air is released to form the sound suddenly.
n	n	n	讷	Alveolar nasal(blade-alveolar);It is articulated with the tongue against the upper alveolar ridge, at the same time the air from the nasal cavity is rushed out of the tongue obstruction, vocal cords are vibrated.
l	l	l	勒	Alveolar lateral approximant (middle blade-alveolar); It is articulated with the lips slightly opened with the tongue against the upper alveolar ridge, vocal cords are vibrated with airflow out from both sides of the tongue.
ts	ts	z	资	Alveolar affricate, unaspirated (Supradental); It is articulated with the tongue against the upper back of the teeth ,at the same time the air flow is hindered, the weak air could be rushed out of the tongue obstruction, the sound is formed of friction .
ts^h	ts^h	c	雌	Alveolar affricate, aspirated (Supradental); It is articulated with tongue against the upper back of the teeth, the air flow is hindered in the mouth, the sound is formed of friction with strong air extrusion from the narrow channel.
s	s	s	思	Alveolar affricate, voiceless fricative (Supradental); It is articulated with the tongue close to the upper back of the teeth, leaving a

				narrow gap, the sound is formed of friction with the air out from the slit of tongue.
tʂ	tʂ	zh	知	Retroflex, unaspirated (blade-palatal); It is articulated with the tongue upturned, against the front of the hard palate, the sound is formed of friction with a weaker air breaking through the tongue obstruction.
tʂʰ	tʂʰ	ch	蚩	Retroflex, aspirated (blade-palatal); It is articulated with tongue upturned, against the front of the hard palate, the sound is formed of friction with a strong air breaking through the tongue obstruction from a narrow channel.
ʂ	ʂ	sh	诗	Voiceless retroflex fricative (blade-palatal); It is articulated with tongue upturned, close to the front of the hard palate, leaving a narrow gap with the air out from a narrow, the sound is formed of friction .
ʐ	ʐ	r	日	Voiced alveolo-palatal fricative (blade-palatal); It is articulated with tongue upturned, close to the front of the hard palate, leaving a narrow gap, then air passes through the throat with the airflow out from a narrow channel, vocal cords are vibrated, the sound is formed of friction.
tɕ	tɕ	j	基	Alveolo-palatal, unaspirated approximant; It is articulated with the tongue against the lower incisors and close to the front of the hard palate, the sound is formed with the airflow out forma narrow channel.
tɕʰ	tɕʰ	q	欺	Alveolo-palatal, aspirated affricate; It is articulated with the tongue against the front of the hard palate, the sound is formed of friction with the air breaking through the tongue.
ɕ	ɕ	x	希	Voiced alveolo-palatal fricative (front lingual surface); It is articulated with tongue against the lower incisors, then the tongue is raised near the front of hard palate, the sound is formed of friction with the airflow out from a narrow channel.
k	k	g	哥	Voiceless velar plosive, unaspirated (back lingual surface); It is articulated with anterior tongue against the soft palate to obstruct airflow, make the air break through the tongue of the obstruction, the

				sound is formed of bursting.
k^h	k^h	k	科	Voiceless velar plosive, aspirated (back lingual surface); It is articulated with front of the tongue against the soft palate to make the airflow obstructed, the air is broken through the tongue of the obstacles, the sound is formed of bursting.
x	x	h	喝	Voiceless velar fricative (back lingual surface); The root of the tongue is lifted up, close to the soft palate, the sound is formed of friction with the air out from a narrow channel.
ø	ø			Zero consonant, don't write it in the front of the vowels.
pf			普	Voiceless labiodental plosive fricative.
pf^h			吹	Voiced labiodental plosive fricative.
v			味	Voiced labiodental fricative; It is articulated with the lower lip and the upper teeth and produced by constricting air flow through a narrow channel at the place of articulation.
ŋ			爱	Velar nasal; It is articulated with the back of the tongue at the soft palate and produced by obstructing airflow in the vocal tract with the vocal cords are vibrated.
ɲ			女	Retroflex nasal; it is articulated palatal and produced by obstructing airflow in the vocal tract with the vocal cords are vibrated.

Table 1. 3 shows the vowels of the Chinese Shaanxi Xi'an Dialect and the corresponding relationship between the simple and compound vowels of the Dialect and of Mandarin. The Dialect has 13 unique vowels but also shares some of its phonemes with Mandarin. It is clear to see the correspondence between the Chinese Shaanxi Xi'an Dialect and Mandarin through the detailed description of phoneme category and articulation methods in Table 1. 3.

Table 1. 3. Vowels alphabet of Dialect and comparison with the Chinese Pinyin Scheme

phonetic alphabet		*Pinyin expression*	*character*	*Description*
Dialect	*Mandarin*			
a	a	a	啊	Open front unrounded vowel, single vowel.
ia	ia	ia	呀	Back ring compound finals; Treat it as 'i+a'.
ua	ua	ua	蛙	Back ring compound finals; Treat it as 'u+a'.
ɣ	ɣ	e	鹅	Voiced velar fricative, single vowel； Mouth position is middle back, tongue is back, both

				sides of the mouth are expanded into flat with vocal cords vibrating.
iɛ	iɛ	ie	耶	Back ring compound vowels; Treat it as 'i+e'.
yɛ	yɛ	üe	约	Back ring compound vowels, y is articulated first then slide to 'ɛ'. Mouth changes from round to flat.
ɿ	ɿ	(-i front)	是	-i front, only used with z, c, s.
ʃ	ʃ	(-i back)	失	-i back, only used with zh, ch, sh.
i	i	i	衣	Close front unrounded vowel, single vowel.
u	u	u	乌	Close back rounded vowel, single vowel.
y	y	ü	迂	Close front rounded vowel, single vowel.
o	o	o	喔	Close-mid back rounded vowel, single vowel.
uo	uo	uo	窝	Back ring compound vowels；Treat it as 'u+o'.
əɹ	əɹ	er	耳	Single vowel.
eɪ	eɪ	ei	诶	Front ring compound vowels ； ' e' is articulated first then slide to 'ɪ'.
ueɪ	ueɪ	uei	威	Middle ring compound vowels ； Treat it as 'u+ei'.
ɑʊ	ɑʊ	ao	熬	Front ring compound finals； 'ɑ' is articulated first, and then the tongue after the retraction, make the tongue root lift, mouth-shaped into a round, gently sliding 'ʊ'.
iɑʊ	iɑʊ	iao	腰	Middle ring compound vowels ； Treat it as 'i+ao'.
ɣʊ	ɣʊ	ou	欧	Front ring compound vowels；/ɣ/ is articulated first then slide to /ʊ/.
iɣʊ	iɣʊ	iou	忧	Middle ring compound vowels ； Treat it as 'i+ou'.
ɑŋ	ɑŋ	ang	昂	Front nasal vowels; /ɑ/ is articulated first then slide to /ŋ/.
iɑŋ	iɑŋ	iang	央	Back nasal vowels; Treat it as 'i+ang'.
uɑŋ	uɑŋ	uang	汪	Back nasal vowels; Treat it as 'u+ang'.
əŋ	əŋ	eng	赣	Front nasal vowels; /ə/ is articulated first then slide to /ŋ/.
iŋ	iŋ	ing	英	Front nasal vowels place tongue against the below gums and tongue is uplifted to the hard palate, nasal resonance sound.
ʊŋ	ʊŋ	ong	翁	Back nasal vowels; /ʊ/ is articulated first then slide to /ŋ/ which retract the tongue against the soft palate, tongue is uplifted, lips are rounded, nasal resonance into the sound.

iʊŋ	iʊŋ	yong	雍	Back nasal compound vowels; treat it as 'y+ong'.
æ			盖	Near-open front unrounded vowel； the tongue is positioned as far forward as possible in the mouth without creating a constriction， lips are not rounded.
iæ			岩	Compound vowels; 'i' is articulated first then slide to 'æ', mouth changing from near-close front to near-open front.
uæ			外	Compound vowels； 'u' is articulated first then slide to 'æ', mouth changing from near-close back to near-open front.
æ˜			安	Nasalized vowels, single vowel; When 'æ' articulated the soft palate descends and the mouth and nasal cavity open at the same time.
iæ˜			烟	Nasalized vowels, compound finals; 'i' is articulated first then slides to 'æ˜', when articulated, the soft palate descends and the mouth and nasal cavity open at the same time.
uæ˜			弯	Nasalized vowels, compound finals; 'u' is articulated first then slides to 'æ˜', when articulated, the soft palate descends and the mouth and nasal cavity open at the same time.
üæ˜			冤	Nasalized vowels, compound finals; 'ü' is articulated first then slide to 'æ˜', when articulated, the soft palate descends and the mouth and nasal cavity open at the same time.
ẽ			恩	Nasalized vowels single vowel; When 'e' articulated the soft palate descends and the mouth and nasal cavity open at the same time.
iẽ			因	Nasalized vowels, compound finals; 'i' is articulated first then slide to 'ẽ', when articulated, the soft palate descends and the mouth and nasal cavity open at the same time.
uẽ			温	Nasalized vowels, compound finals; 'u' is articulated first then slide to 'ẽ', when articulated, the soft palate descends and the mouth and nasal cavity open at the same time.
üẽ			晕	Nasalized vowels, compound finals; 'ü' is articulated first then slides to 'ẽ', The soft palate descends and the mouth and nasal cavity open at

				the same time when pronounced.
ɯ			核	Close back unrounded single vowel; The tongue is positioned as close as possible to the roof, as far back as possible, lips are not rounded.
yo			药	Compound vowels； /y/ is articulated first then slide to /o/, mouth changing from near-close, near-front to close-middle back.

1.2. Tone types intonation of the Chinese Shaanxi Xi’an Dialect

Tone is not a visible feature of speech. It is not essential for the Chinese Shaanxi Xi’an Dialect talking head but it is absolutely necessary for a Chinese Shaanxi Xi’an Dialect speech assistant system (SA) because it is an important part of assessment of intonation and prosody in the SA.

In the early 20th century, the tone and intonation research for Chinese entered into a new phase due to two phoneticians: Dr. Liu Fu[1] and Dr. Chao Yuan-ren. Chao pointed out that the syllabic tone patterns can be modified by the sentential attitudinal intonation, just like “the small ripples riding on top of large waves”. It makes clear the relation between syllabic (also phrasal) tone patterns and the sentential intonation contours, and gives a key solution to a number of problems in intonation analysis. He claims that register is used to describe English intonation, while contour is used in studying Chinese intonation.

It is common to describe Chinese tone by tone type and tone pitch. The five - degree mark method [43] is used to annotate the changes of tone pitch, which is dividing a vertical line into four quarters of five degrees to represent the Chinese tone pitch and drawing the corresponding lines or using two or three digital to describe the tone types as the method shown in Figure 1. 1. The Chinese Shaanxi Xi’an Dialect and Mandarin each have four kinds of tone types [44][45] which are different in the tone pitch.

Table 1. 4. Tone Types Comparison between Mandarin and the Chinese Shaanxi Xi'an Dialect

Tone type	*Mandarin*	*Dialect*	*Examples*
	Tone pitch	*Tone pitch*	
1st	55	21	山 shān——shan
2nd	35	24	文 wén ——wén
3rd	214	53	反 fǎn —— fàn
4th	51	55	动 dòng—— dōng

The Chinese Shaanxi Xi’an Dialect has 4 monosyllabic word tones. Yuan Jiahua classified the Chinese Shaanxi Xi'an Dialect as being in the northwest Mandarin area, and pointed out that the

[1] The order of Chinese name in the thesis is family name first and first name second.

tone pitches of the four monosyllabic words representing 1st tone, 2nd tone, 3rd tone and 4th tone were described by two or three numbers 21, 24, 53 and 45, respectively [46]. The Department of Chinese Language and Literature of Peking University introduced the phonological system of the Chinese Shaanxi Xi'an Dialect, giving the tone pitch levels for the dialect as 21,24,53 and 45 [47].What is more, the author pointed out that there is a slight rise trend in the 4th tone with the tone pitch 45. Wang Junhu divided the Chinese Shaanxi Xi'an Dialect into new and old school and describes the tone system, with the four tone pitches are 21, 24, 53 and 44 [48]. Ma Maopeng from the Department of Chinese of Nanjing University did the single-tone tonal acoustics experiments, finding that the tone pitches levels are 21, 24, 53 and 44 [49]. There is still some controversy regarding the pitch value of the 4th tone; this paper makes comparisons between Mandarin and the Shaanxi Xi'an Dialect in Table 1. 4 according to the tone pitch given by Peking University [47].

Based on the four kinds of tone pitches given in Table 1. 4, tone diagrams can be drawn according to the five - degree mark method. Figure 1. 2 shows the tone-contour patterns of Mandarin and the Chinese Shaanxi Xi'an Dialect. It is obvious to describe regular pattern of the four tone pitch changes and the corresponding relationship between Mandarin and Dialect.

Figure 1. 2. Tone-contour patterns of Mandarin (MT, above) and the Chinese Shaanxi Xi'an Dialect (DT, below)

In Figure 1. 2, the four tone-contour patterns and the changes of the pitch for Mandarin, in the upper diagram, and Dialect, in the lower diagram. The horizontal coordinate is used to present different tone styles while the vertical coordinate is used to present tone pitch level which is divided into five degrees, "Low", "medium low", "medium", "medium high" and "high"

represented by the Arabic numerals 1 to 5. The 1st tone of Mandarin 'MT' in Figure 1. 3 labeled '55' to illustrate that the tone pitch keeps on high level and unchanged in Mandarin while '21' illustrates that the tone pitch changes from medium low pitch level '2' to low pitch level '1' in the Chinese Shaanxi Xi'an Dialect. According to the regulation of the changes of each tone style for Mandarin and Dialect in Figure 1. 2, the meaning of the other three tone styles could be clearly expressed.

For Chinese Shaanxi Xi'an Dialect, a number of factors affect the pitch contour. For example, different stress patterns and sentence moods can change the pitch contour on the linguistic level; different consonants like voiceless fricatives and aspirated affricates can raise pitch contour while the sonorant and lip sounds can lower pitch contour on the segmental level. In addition, the mechanisms of tone perception and production also contribute to the complexity of tonal contours.

Generally speaking, the components of Chinese Shaanxi Xi'an Dialect intonation include three major parts [50]:

1. The first part is the pitch variation, which consists of the variation of pitch range, pitch register and the global trend of the pitch contour. Pitch range or tonal range is the pitch variable range of different tones. The variation of pitch range or tonal range refers to its compression or expansion. Pitch register or tonal register is the "key" of the pitch range used by Wu Zongji. Chao mentioned that there are two kinds of register change: entire pitch raise or entire pitch fall. The variation trend of the pitch contour is characterized by the variation of the top line and bottom line [44] [45].

2. The second part is the prosodic structure which is conventionally called rhythmic structure. The rhythmic group is a word, phrase or sentence, while the perceived pause is categorized as physiological pause, grammatical pause or psychological pause.

3. The third part is stress structure. Traditionally Chinese has word stress and sentence stress. The sentence stress consists of logical and attitudinal prominence. Unlike French, for instance, Chinese stress is variable. A syllable can be stressed, weak or medium. A weak syllable or neutralized syllable can change the words meaning.

1.3. Design of the Chinese Shaanxi Xi'an Dialect X-SAMPA

SAMPA (the Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987-1989 by an international group of phoneticians, and was applied in the first instance to the European Community languages such as Danish, Dutch, English, French, German, and Italian (by 1989); later to Norwegian and Swedish (by 1992); and subsequently to Greek, Portuguese and Spanish (1993). Under the aegis of COCOSDA it is hoped to extend it to cover many other languages and in principle all languages.

These X-SAMPA codes cover everything on the 1993 IPA Chart, including diacritics and tone types, and were put forward as a proposed standard way to transmit IPA- transcribed material by e-mail and for similar purposes. It is an extension of the SAMPA standard, with which colleagues may be familiar. The most frequently used symbols are mapped onto single keystrokes in the ASCII range 33,126. Less frequently used symbols are mapped onto a single keystroke plus the backslash, \. Diacritics (other than those already catered for in SAMPA) are mapped onto a keystroke with a preceding underscore, _.

With SAMPA and X-SAMPA, we considered approaches to getting the same system for the Shaanxi Xi'an Dialect, because it is a vital step in my research to label segments in speech corpora for the Chinese Shaanxi Xi'an Dialect. With the labeled material we can do many research projects. The Pinyin Scheme is an effective way to transcribe Mandarin (standard Chinese), but it does not entirely correspond to IPA. Using the international machine-readable symbol system SAMPA, Zhu Weibin and Zhang Jialu have transcribed a symbol system with SAMPA for labeling syllables [51] . They give Chinese SAMPA symbols including consonant, vowel and tone charts. They label isolated syllables in a database. This is an important work for transcribing Mandarin (standard Chinese), but it is not sufficient to describe for Standard Chinese in continuous speech. An application of SAMPA-C in Standard Chinese [52] presents a labeling system for Standard Chinese named SAMPA-C. They give some charts: consonant chart, vowel chart, tone chart, retroflex final chart, sound variation chart and non-speech symbol chart. The labeling system was used to label two corpora [52].

In Chapter 1.1.2 I have already created Chinese Shaanxi Xi'an Dialect alphabet, consisting of 26 consonants and 40 vowels. Its phonemes represented in IPA can be found in Table 1. 2 and 1.3. Its five unique consonants are presented in Table 1. 2: 'pf' and 'pfʰ' are both labiodental plosive fricative consonants, 'v' is a voiced labiodental fricative consonant, 'ɲ' is a velar nasal consonant, and 'ɳ' is a retroflex nasal consonant.

Based on these, firstly, we give some rules for designing the system and then we design an X-SAMPA labeling system for Chinese Shaanxi Xi'an Dialect. In order to establish the corresponding relationship between consonants and SAMPA ,the first step is to find out the same articulation as the international phonetic alphabet, then translate the phonemes of Chinese Shaanxi Xi'an dialect into X-SAMPA according to the IPA. The X-SAMPA code derived for the Chinese Shaanxi Xi'an Dialect is based on the Hungarian X-SAMPA code. Table 1. 5 shows the X-SAMPA symbols for consonants of Chinese Shaanxi Xi'an Dialect.

Table 1. 5. X-SAMPA for consonants of the Chinese Shaanxi Xi'an Dialect

Character	*Pinyin expression*	*IPA*	*X-SAMPA*
玻	b	p	b
坡	p	pʰ	p
摸	m	m	m
佛	f	f	f

Character	*Pinyin expression*	*IPA*	*X-SAMPA*
得	d	t	d
特	t	t^h	t
讷	n	n	n
勒	l	l	l
哥	g	k	g
科	k	k^h	k
喝	h	x	x
基	j	tɕ	tC
欺	q	$tɕ^h$	tS
希	x	ɕ	s\
知	zh	tʂ	dS
蚩	ch	$tʂ^h$	dC
诗	sh	ʂ	s`
日	r	ʐ	z`
资	z	ts	z
雌	c	ts^h	c
思	s	s	s
追		pf	pf
吹		pf^h	pv
味		v	v
爱		ŋ	N
女		ɲ	J

The same method was applied to vowels and tone type translation. In the previous section, I established phonetic alphabet for Chinese Shaanxi Xi’an Dialect including the same phonemes represented by Roman alphabet and some unique phonemes of the Dialect represented by IPA symbols. As shown in Table 1. 3, it also composed by simple vowels and compound vowels and 13 unique vowels but also shares some of the same phonemes with Mandarin. Based on this alphabet, the translation of vowels into X-SAMPA code [53][54]Error: Reference source not found is created and presented in Table 1. 6. The translation of the phonemes ‘v’ and ‘ɳ’ to X-SAMPA code is based on the Hungarian phoneme system.

Table 1. 6. X-SAMPA for vowels of the Chinese Shaanxi Xi'an Dialect

Character	*Pinyin expression*	*IPA*	*X-SAMPA*
啊	a	a	a
喔	o	o	o
鹅	e	ɤ	G
衣	i,yi,y	i	i
乌	u,wu,w	u	u

Character	*Pinyin expression*	*IPA*	*X-SAMPA*
迂	ü, yu	y	y
诶	ei	eI	eI
威	uei,ui,wei	ueI	ueI
熬	ao	ɑʊ	AU
欧	ou	ɣʊ	GU
忧	iou, iu, you	iɣʊ	iGU
耶	ie, ye	iɛ	iE
约	üe, yue	yɛ	yE
儿	er	əɹ	@ l\
昂	ang	ɑŋ	AN
赣	eng	əŋ	@N
英	ing, ying	iŋ	iN
翁	ong	ʊŋ	UN
呀	ia, ya	ia	ia
腰	iao, yao	iɑʊ	iAU
蛙	ua, wa	ua	ua
窝	uo, wo	uo	uo
央	iang, yang	iɑŋ	iAN
雍	iong, yong	iʊŋ	iUN
汪	uang, wang	uɑŋ	uAN
翁	ueng, weng	uəŋ	u@N
哀		æ	{
岩		iæ	i{
歪		uæ	u{
安		æ̃	{~
岩		iæ̃	i{~
弯		uæ̃	u{~
冤		yæ̃	y{~
恩		ẽ	e~
因		iẽ	ie~
温		uẽ	ue~
晕		yẽ	ye~
核		ɯ	M
药		yo	yo

In Chapter 1.2 I describe the tone type based on the five - degree mark method. It expresses clearly the tone pitch level of Chinese Shaanxi Xi'an dialect in Figure 1. 2. Based on the figure I found the corresponding IPA then translated these tone types into X-SAMPA code. Chinese Shaanxi Xi'an dialect tone translation expression is showed in Table 1. 7.

Table 1. 7. Chinese Shaanxi Xi'an Dialect tone translation expression in X-SAMPA

	1st tone	*2nd tone*	*3rd tone*	*4th tone*
Dialect	_M_B	_L_H	_T_M	_T
Mandarin	_T	_M_T	_L_B_H	<F>

The meanings of the symbols in Table 1. 7 are as follows:

- _M — mid tone
- _B — extra low tone
- _L — low tone
- _H — high tone
- _T —extra high tone
- <F> — global fall

The symbol '_L_B_H' means the tone changes from a low tone to an extra low tone then to high tone and so on.

1.4. Phonemic differences between Mandarin and the Chinese Shaanxi Xi'an Dialect

Pronunciation is the main difference between the Dialect and Mandarin, especially with consonants, although variation is also found in vowels. Though quite complex, the variation follows some rules Error: Reference source not found.

1.4.1. The correspondence of consonants 'v', 'ŋ', 'ɳ' in the Chinese Shaanxi Xi'an Dialect and Mandarin [57][58]

When a syllable starts from a consonant 'w', it is always read as 'v'. It is necessary to change all of the phonemes 'v' to 'w' when the dialect is spoken without consonants 'pf', 'pfʰ'. That means 'v' and 'w' have a one-to-one relationship between dialect and Mandarin. However, we should change only some of phonemes 'v' to 'w' and other instances of 'v' to 'r'. Table 1. 8 shows the correspondence relationship between 'v' in dialect and 'w' in Mandarin.

Table 1. 8. The relationship between 'w', 'r' and 'v' in Mandarin and the Dialect

Mandarin	*Dialect*	*Examples*
w	*v*	万 wàn-van; 问 wèn-ven; 望 wàng-van
r	*v*	若 ruì-vui; 软 ruǎn- vóng; 荣 róng-vong

Note: the left syllable of the horizontal line represents Mandarin expression while the right syllable of the horizontal line represents dialect expression in Table 1. 8*,* Table 1. 11 *and* Table 1. 15*.*

In Table 1. 8, it can be seen that '万' is the Chinese character and both sides of the horizontal line indicate the different consonants of pronunciation in the same word between Mandarin and the Dialect. It is important that 'v' is a unique consonant of the Chinese Shaanxi Xi'an dialect which does not occur in Mandarin.

It is common to add a consonant 'ŋ' before 'a', 'o', 'e' when syllable begin at these three phonemes 'ŋ' is a unique consonant in the Chinese Shaanxi Xi'an dialect. Table 1. 9 presents examples to show the corresponding phonemes between the two languages.

Table 1. 9. The relationship between [ŋ] and 'a', 'o', 'e' in Mandarin and the Dialect

character	爱	案	饿	恩	偶	奥
Mandarin	ài	àn	è	ēn	ǒu	ào
Dialect	ŋài	ŋàn	ŋè	ŋēn	ŋǒu	ŋào

In addition it is required to change 'ɳ' to 'n', 'y' in Chinese Shaanxi Xi'an dialect syllabic when the consonant 'n' is spelled with some special vowels or some particular vowels could be an independent syllabic following the phoneme 'y' in Mandarin. The correspondence of phonemes between two languages is presented in Table 1. 10.

Table 1. 10. Situations of changing to 'n', 'y' in the Chinese Shaanxi Xi'an Dialect

character	你	牛	鸟	娘	疑	压	页	严
Mandarin	nǐ	niú	niǎo	niāng	yí	yā	yè	yán
Dialect	ɳi	ɳiou	ɳiau	ɳiaŋ	ɳi	ɳia	ɳie	ɳian

1.4.2. The correspondence of some consonants between the Chinese Shaanxi Xi'an Dialect and Mandarin [59][60]

Different pronunciation can take place when we input the same character, which is mainly due to the different pronunciation in the corresponding syllable of the character. Table 1. 11 shows the relationship of some consonants between Dialect and Mandarin. The syllables beginning with the two consonants 'n' or 'l' basically have the same articulation as Mandarin, but there also exists a situation with such a character as '农(nóng)' , where 'n' is articulated 'l' in this situation and 'ch', 't', 'd', 'k' also have the corresponding phones in Dialect. It also can be seen that some syllables starting with one of the consonants 'j', 'q', or 'x' in Mandarin should be articulated consonants 'g', 'k', or 'h' in Dialect.

Table 1. 11. The relationship of some consonants between the Dialect and Mandarin

Mandarin	*Dialect*	*Examples*

n	l	拿 ná-la; 奈 nài-lai; 弄 nòng-lòng; 暖 nuān-luan
ch	sh	尝 chāng-shǎng; 盛 chéng-sheng; 晨 chén-shen; 蝉 chán-shan
t	q	踢 tī- qi; 调 tiáo-qiao; 田 tián-qian; 贴 tiē-qie
d	j	滴 dī-ji; 跌 diē-jie; 掉 diào-jiao; 丢 diū-jiu
k	f	哭 kū-fu; 苦 kǔ-fu; 酷 kù-fu
j	z	俊 jùn-zun; 炯 jiǒng-jiong； 精 jīng-zing
q	c	全 quān-cuan; 群 qún-qun；晴 qíng-cing
x	s	选 xuān-suan; 讯 xùn-sun; 削 xūe-suo; 行 xíng-sing

A considerable number of non-aspirated consonants 'b', 'd', 'g', 'j', and 'z' in Mandarin are replaced by aspirated consonants 'p', 't', 'k', 'q', and 'c' in Dialect. A series of examples is given in Table 1. 12 to explain this phenomenon.

Table 1. 12. The relationship between non-aspirated and aspirated consonants

character	鼻	柜	旧	知	早	国
Mandarin	bí	guì	jiù	zhī	zǎo	guó
Dialect	pi	kui	qiu	chi	cao	gui

In most parts of the Xi'an area, when 'zh', 'ch', and 'sh' appear at the beginning of the syllables they should be articulated as two groups 'zh', 'ch', ' sh' or 'z', 'c', 's'. The basic rule is as follows: when the spelling 'zh', 'ch', 'sh' occur with vowels such as 'a', 'ai', 'an' , and 'en' ,they are articulated 'z', 'c', 's'. Otherwise, they are usually pronounced 'zh', 'ch', 'sh'. In addition, some pronunciation does not observe this rule, for instance, '傻(shǎ)'is articulated as 'sha' and '师(shī)' is articulated as 'shi'. Some examples are listed to express this rule in Table 1. 13

Table 1. 13. The relationship between 'zh', 'ch', 'sh' and 'z', 'c', 's' phonemes

character	暂	知	产	潮	省	陕
Mandarin	zhǎn	zhī	chǎn	chǎo	shěng	shǎn
Dialect	zan	zi	can	cao	seng	san

In addition, there is a special pronunciation for syllables such as 'zhu', 'chu', 'shu', and 'ru', which most speakers in the Chinese Shaanxi Xi'an Dialect area articulate as if they were 'z', 'c', 's', and 'z' with a certain kind of vowel or syllable in Chinese Shaanxi Xi'an. 'pf', 'pfʰ' and 'f' are fricative, voiceless and unaspirated, while 'v' is voiced fricative. Table 1. 14 sets some examples to explain this articulation.

Table 1. 14. Complex reading of syllables 'zhu', 'chu', 'shu', 'ru'

character	猪	追	入	吹
Mandarin	zhū	zhuī	rù	chuī

dialect	pfu	pfui	vu	pf^{h}ei

This section has introduced and analyzed the phonetic features of consonants for the Chinese Shaanxi Xi'an Dialect and expressed the phonemic changes by comparison with Mandarin. We now turn to vowels.

1.4.3. Vowels feature of the Chinese Shaanxi Xi'an Dialect comparing with Mandarin

When consonants written 'd', 't', 'n', 'l' ,'z', 'c' ,'s' and 'zh' ,'ch' ,'sh' appear in the front of the vowel 'u', 'u', its articulation will be changed to 'ou'. This phenomenon is extremely common, and is widely recognized an accent feature of Xi'an even for the Shaanxi people. There are also some other vowel changes. Table 1. 15 presents examples to show the corresponding phonemes between the two languages.

Table 1. 15. The relationship between parts of vowels

Mandarin	*Dialect*	*Examples*
u	ou	读 dú-dou; 路 lù-loù; 足 zú-zoú; 醋 cù-cou; 数 shù-soù
e	i	液 yè-yi
ie	i	携 xié- xī
u	i	婿 xù-xi
uo	u	措 cuò-cù

It is very common to articulate 'an', 'ian', 'uan', 'üan' as 'a' or 'ai' ('æ') which is a nasalization tone in the Chinese Shaanxi Xi'an Dialect. Table 1. 16 presents examples to express these changes with heavy nasal sound.

Table 1. 16. Special vowel changes

character	三	端	捐	电
Mandarin	sān	duān	jüān	diàn
dialect	sain	duain	jüai	die

The situation of vowels division is very complicated in the Chinese Shaanxi Xi'an Dialect. The most typical rule is that the three vowels 'e', 'ai', 'o' in Mandarin are pronounced 'ei' in the Chinese Shaanxi Xi'an Dialect. Some examples are given in Table 1. 17.

Table 1. 17. The relationship between parts of vowels

character	择	色	墨	拍
Mandarin	zé	sè	mò	paī
dialect	zei	sei	mei	pei

In 'White read' called vernacular reading, also known as oral reading, in opposition to text read, classical reading, which is the concept of written language reading. It is common to articulate the phoneme 'xi' in Mandarin as 'h' in the Chinese Shaanxi Xi'an Dialect. Some examples are listed to express this change in Table 1. 18.

Table 1. 18. Changes of 'White read' in the Dialect

character	下	鞋	咸	项	杏
Mandarin	xià	xié	xiá	xiàng	xìn
dialect	ha	hai	han	hang	heng

1.5. Theoretical method of transcription from the Chinese character to X-SAMPA

I have already created a labelling system in the X-SAMPA code as follows: consonant chart, vowel chart, tone type chart in Table 1. 5, Table 1. 6 and Table 1. 7. Based on these three tables I supply a method to develop the labelling system. This allows me to segment and label the corpora with X-SAMPA for the Chinese Shaanxi Xi'an Dialect.

As we know, the Chinese Shaanxi Xi'an Dialect shares the same characters with Mandarin (the official language of China) but the pronunciation is different. We propose methods to translate the Chinese character to X-SAMPA code for Chinese Shaanxi Xi'an Dialect. For example, the Chinese character 赵璐, can first be translated in to Pinyin (Latin expression) or IPA (because there exist some phonemes in Shaanxi Xi'an Dialect that have no Latin expression), then labeled in the X-SAMPA code I created in this thesis for the Chinese Shaanxi Xi'an Dialect. Based on this we can translate all the Chinese sentences and characters into the machine readable code X-SAMPA of the Shaanxi Xi'an Dialect. This is the basis for other applications such as TTS for the Chinese Shaanxi Xi'an Dialect.

This process is like this: 赵璐(Chinese characters)——zhāo lū (Pinyin or IPA)——zhao1 lu1 (number 1 represents the tone type for dialect) ——dS AU _M_B l u _M_B (X-SAMPA code for phoneme and tone type)

1.6. Thesis

I created a phonetic alphabet of the Chinese Shaanxi Xi'an Dialect based on the Chinese Pinyin Schedule. I developed an X-SAMPA code for phonemes and tone types of the Chinese Shaanxi Xi'an Dialect which was derived based on the Hungarian X-SAMPA code. I presented a method for the phonetic transcription of the Chinese Shaanxi Xi'an Dialect. [61] [109].

1.6.1. Novelty

At present, there are few scientists in China who focus on 3D talking head modeling and animation of the Chinese Shaanxi Xi'an Dialect. So it has great significance to do 3D talking head modeling and animation of this dialect. My work provides a method that can be used in visual speech synthesis and prosody modelling for the Chinese Shaanxi Xi'an Dialect and speech conversion to other dialect and languages.

This thesis has been built on a collection of a large number of audio books and papers related to the Chinese Shaanxi Xi'an Dialect. Therefore, the articulation properties and classification of the phonetic alphabet and the phonemes of the Chinese Shaanxi Xi'an Dialect are quite precise. Therefore, the X-SAMPA code of the Chinese Shaanxi Xi'an Dialect formed on this basis is very accurate.

1.6.2. Measurements

Firstly, I develop a phonetic alphabet for Chinese Shaanxi Xi'an dialect based on the Chinese Pinyin Scheme through the comparison with Mandarin on the aspects of different articulation of consonants, vowels, and tone types. Detailed description of the tone styles of Chinese Shaanxi Xi'an dialect is given and contrasted with that of Mandarin. A 5-degree tone figure of dialect is made. Secondly, an X-SAMPA analysis of dialect phonemes and comparison with Pinyin are set up. Finally, a transcription method for a machine readable phonetic transcription system for the Chinese Shaanxi Xi'an Dialect has been developed in this thesis.

1.6.3. Limits of validity

The focus of this thesis is on the conversion of individual phonemes and tone types of Chinese Shaanxi Xi'an dialect into X-SAMPA codes. It is impossible to calibrate the actual phonetic sentences and paragraphs in a single speech phone. In the process, there is a single sub-combination. Some changes in the way words are pronounced require further revision of the calibration rules based on the particular linguistic phenomenon of the pronunciation of the words.

Before recording the visual speech database, we have to consider the speaker, speech material and some techniques to capture data conveniently. Using only one speaker can simplify the recording and analyzing of data. Considering the feasibility, we chose a speaker (the author) who is a native speaker of the Chinese Shaanxi Xi'an Dialect.

1.6.4. Consequence

The purpose of this thesis is creating the fundamentals of a talking head – and transcription model – for the Chinese Shaanxi Xi'an Dialect. The automatic transcription system for the Chinese Shaanxi Xi'an Dialect could be developed based on the X-SAMPA code developed in this thesis.

It can be used in visual speech synthesis and prosody modelling for the Chinese Shaanxi Xi’an Dialect and speech conversion to other languages.

1.6.5. Related published paper

Czap, László, and Lu Zhao: Phonetic aspects of Chinese Shaanxi Xi'an dialect [C]. 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, 2017.

Zhao L, Czap L.: Visemes of Chinese Shaanxi Xi’an Dialect Talking Head[J]. Acta Polytechnica Hungarica, 2019, 16(5): 173-193.

Chapter 2

Visemes of the Chinese Shaanxi Xi'an Dialect

Visual speech plays a significant role in speech perception especially for deaf or hard of hearing people, and even for hearing people in a noisy environment. Lip reading depends on visible articulators to improve speech perception. However, not only the movement of lip and face provide part of phonetic information, The motion of the tongue which is generally not entirely seen carries an important part of the articulatory information not accessible through lip reading.

Human speech perception is clearly multimodal involving both auditory and visual modalities, but we cannot distinguish which modality has more effect, as it depends on the conditions of speech Error: Reference source not found. The divergent research results shows that speech perception and production abilities are largely affected by a lack of exposure to multimodal stimuli through comparing the development of normally hearing children with deaf or blind children Error: Reference source not found. All articulatory features are better transmitted in the audiovisual mode than in any unimodal communication Error: Reference source not found. Numerous clinical and laboratory studies on the auditory-visual performance of normal-hearing and hearing-impaired children and adults demonstrate that combined auditory-visual perception is superior to perception through either audition or vision alone Error: Reference source not found. The perception of visible articulators including the lips, jaw, face, tongue tip, and teeth helps humans to understand speech. Thus visual speech plays a significant role in speech perception especially for deaf or hard of hearing people or hearing people in noisy environment. Visual speech is well studied and used in speech recognition [66], speech processing [67], audio-visual speech synthesis [68], virtual talking head animation and lip or tongue synchronization [69].

2.1. Viseme definition

In visual speech research, it is necessary to put forward the term viseme which is an essential unit to build on in research on the visual features of speech. The term viseme refers to a group of phonemes that is visually closer to each other than to other phonemes. This is not to say that members of the same viseme class are thus similar in many features Error: Reference source not found.

A viseme is one of several speech sounds that look the same, for example when lip reading (Fisher 1968). Visemes and phonemes do not share a one-to-one correspondence. Often several phonemes correspond to a single viseme, as several phonemes look the same on the face when

produced, such as /k, g, ŋ/, (viseme: /k/). However, there may be differences in timing and duration during actual speech in terms of visual 'signature' of a given gesture that cannot be captured with a single photograph [70].

With the development of visual speech, the MPEG-4 standard defines viseme, which is used to describe the physical state of the corresponding visual vocal organ when a certain phoneme is sent. It is often not enough to describe it only on a static image. We refer to the complete change process of the lip shape from generation to disappearance as a dynamic viseme, which can provide more information for people to understand the speech Error: Reference source not found. Although MPEG-4 classifies the articulation of international phonetic symbols into 15 static visemes, considering the articulation characteristics of various languages and the different phoneme components, scholars from various countries have done a great deal of research on the pronunciation of different languages. For example, Bothe classifies the German articulation lip into 12 static visemes [71], Le Goff classifies the French articulation into 19 static visemes [72], Ezzat divides the English articulation into 16 static visemes [73], Lande divides Italian articulation into 23 static visemes [74]. László Czap make the viseme classes based on the lip shapes, the invisible tongue position can be different. He divides 10 consonants visemes and 7 vowels visemes [5], and so on. These viseme classes are defined for Hungarian language.

The visual vocal organs of the talking head are mainly tongue position and the lip shape. Below I will present my study of the visemes of the Chinese Shaanxi Xi'an Dialect in these two aspects.

2.2. Static lip viseme classification of the Chinese Shaanxi Xi'an Dialect

The human vocal organs consist of a respiratory organ, the throat, the vocal folds, the oral cavity, and the nasal cavity. In the process of articulation, it is very complicated to consider the movement of the lower half of the face at each moment due to the pulling of the muscle from a physiological anatomical point of view.

2.2.1. Research on static lip viseme of Mandarin

In Mandarin, each speech unit that can be naturally distinguished in a language is a syllable. Usually a Chinese character is a syllable. Generally, a syllable is composed of consonants and vowels. The articulations of consonants in a syllable are short, and then quickly slide to the vowels. In the Chinese Pinyin schedule introduced in the first chapter, there are 23 consonants and 36 vowels. The vowels are classified into simple vowels and compound vowels. When the simple vowel is articulated, the lip shape and the tongue position are unchanged throughout the articulation process, so it can be regarded as a viseme. The compound vowels and nasal vowels have different changes in the lip shape and tongue position during the articulation process. For example, when the compound vowel 'ai' is articulated, the lip shape is opened and gradually closed, and the lip shape is gradually flattened until it is close to the lip of the 'i' phoneme. The real-time nature of the system requires only one or two visemes per Chinese character, so the

articulation of 'ai' can be described by a viseme, which can be regarded as the viseme that is a linear combination of the viseme 'a' and 'i'.

There are relatively few studies on Chinese viseme. Qi Jie classifies Mandarin into six basic types and implements text-driven lip synthesis, but the classification is overly simplified [75]. Table 2. 1 lists the six basic lip viseme types. Zhong Xiao et al. classified Mandarin into 12 categories (including 10 basic lip visemes and two transitional lip visemes) to study the identification of basic lip types [76]. Table 2. 2 is a description of the lip visemes.

Table 2. 1. Six basic viseme types for Mandarin according to Qi Jie [75]

Viseme		*Description*
Simple vowel viseme	'a' viseme	The tongue is centered, the mouth is wide open, and the lips are in a natural state.
Simple vowel viseme	'o' viseme	The tongue is retracted, the tip of the tongue is drooping, and the back of the tongue is raised, and the mouth is rounded.
Simple vowel viseme	'e' viseme	The mouth is slightly open, the lips are closed, slightly rounded, the corners of the mouth are spread out on both sides, and the lips are unfolded.
Simple vowel viseme	'i' viseme	The mouth opens very small, the lip is flat, and the corner of the mouth is pulled back.
Simple vowel viseme	'u' viseme	The mouth opens very small, the lips are rounded, slightly protruding forward.
Bilabial viseme	'b', 'p', 'm'	The lips are closed and hinder the airflow.

Table 2. 2. Classification of 12 lip visemes for Mandarin according to Zhong Xiao [76]

Viseme code	*Viseme*	*Number of visemes*
A	a	1
B	o	1
C	e	1
D	j q x i n ng l z c s	10
E	u	1
F	ü	1
G	er r	2
H	d t g k h	5
I	b p	2
G	m	1
K	f	1
L	zh ch sh	3
Total		29

After analyzing the characteristics and Mandarin articulation phonetic structure, Wang Zhiming et al. classified the Mandarin articulation into 28 basic static lip visemes Error: Reference source not found. In view of the pronunciation habits of Mandarin, they study the classification of mouth shape from the perspective of the vowel. Each consonant corresponds to a consonant phoneme, but there is a many-to-one relationship in the phoneme-to-lip viseme mapping. For example, the articulations of the 'b', 'p' and 'm' phonemes are very similar. On the other hand, due to the influence of coordinated articulation, the same consonants may change in different Pinyin combinations. For example, the patterns in 'du' and 'da' are different, so the mapping from phoneme to lip shape is not the same one. According to the articulation method, the consonants can be classified into bilabial sounds, labiodental sounds, alveolar sounds, palatal sounds and velar sounds. They studied the relationship between the articulation of the consonants and articulation part, then analyzed a large number of related pronunciation data. Their classification of the consonants is shown in Table 2. 3. The consonant viseme groups 'd, t, n' and 'z, c, s' could be unified to one consonant viseme group but the timing is different during the articulation.

Table 2. 3. Lip viseme of consonants for Mandarin Error: Reference source not found

Consonant viseme classification							
b, p, m	f	d, t, n	l	g, k, h	j, q, x	zh, ch, sh, r	z, c, s

Wang Zhiming et al. [62] also mentioned that the articulation of the simple vowels is relatively stable and there is great difference between them. Therefore, the articulation of each simple vowel can be used as a static lip viseme, including 'a', 'o', 'e', 'i', 'u', 'ü', 'er'. But 'o' is a special case, after the 'o' is issued, the lip shape is stable, forming a circle, and the opening is moderate. The actual articulation process is that the mouth shape is first that of the lip shape of phoneme 'u', then transitions to the lip shape of phoneme 'o'. Generally a compound vowel is composed of multiple phonemes, and the lip shape transitions from one lip shape to another. However, some compound vowel combinations are more compact, and the lip shape is difficult to split into multiple lip types. For example, the lip shape of the Mandarin two compound vowels can be considered as a single lip shape, such as 'ai', 'ei', 'ao', 'ou'. In addition, the Mandarin syllables endings '–n', '-ng' have less influence on the lip shape after some specific phonemes, and the tongue position of '–n' is slightly lower than the '–ng' tongue position. Thus, 'an' and 'ai' are similar, 'en' and 'ei' are similar, but they can also be considered as a single lip shape. Therefore, the basic lip viseme of the vowel is classified into 13 static visemes and the lip viseme of other compound vowels is composed of multiple simple vowels. Table 2. 4 shows this classification of lip visemes of vowels for Mandarin.

The analysis by Wang Zhiming et al. for the static lip visemes of mandarin is the most detailed and comprehensive by far. So we will analyze the static lip visemes for Chinese Shaanxi Xi'an Dialect based on this classification.

Table 2. 4. Lip viseme of vowels for Mandarin [62]

	a, ang	ou	er	ü
Simple Vowel	ai, an	e,eng	i	-i （-i front）

	ao	ei, en	u	-i（-i back）
	o			
Compound Vowel	ia, ie, in, ing, iao, iou			
	ian, iang, ua, uo, uai			
	uei, uæ̃, uan, un, uang, ueng,ong			
	üan, üe, ün, iong			

2.2.2. Static lip visemes of the Chinese Shaanxi Xi'an Dialect

We have established a set of static visemes based on the characteristics of the Chinese Shaanxi Xi'an Dialect and phonetic composition referring to the classification of Mandarin static viseme by Wang Zhiming et al., as the Dialect shares many of the same phonemes with Mandarin and supplement these with some special phonemes. The phonemes of Chinese Shaanxi Xi'an Dialect are described using IPA, and there are 26 consonants and 40 vowels. The statistical methods of factor analysis and classification are applied based on the results of description statistics. The viseme system in Standard Chinese is determined by means of statistics analysis. Fifteen basic static vowel visemes are classified and other compound vowels are composed of a multiple of single vowel visemes, as illustrated in Table 2. 5. The static consonant viseme classification of Chinese Shaanxi Xi'an dialect speech is shown in Table 2. 6, where 11 static consonant viseme groups are classified.

Table 2. 5. Static vowel visemes classification for the Dialect

Simple Vowel	a, ang	啊，昂	er	儿
	æ, ɛ̃	哀，安	i	衣
	ao	奥	u	乌
	o	喔	ü	迂
	ou	欧	-i（-i front）	是
	e,eng	鹅，赣	-i（-i back）	失
	ei, ẽ	诶，恩	ɯ	核
	yo	药		
Compound Vowel	ia=i+a; ie=i+e; iẽ =i+ẽ; ing=i+eng; iao=i+ao; iou=i+ou			
	iæ=i+æ; iæ̃ =i+ æ̃; iang=i+ang; ua=u+a; uo=u+o; uæ =u+ æ			
	uei=u+ei; uæ̃, uæ̃ =u+ æ̃; uẽ =u+ ẽ; uang=u+ang; ueng,ong=u+eng			
	yæ̃=y+ æ̃; üe=ü+e; yẽ=y+ẽ; iong=i+ong			

Note: 'æ', 'iæ', 'uæ', 'ɛ̃', 'iæ̃', 'uæ̃', 'yæ̃', 'ẽ', 'iẽ', 'uẽ', 'yẽ', 'ɯ', 'yo' are phoneme illustrate by IPA and others are romanized expression of phonemes.

Table 2. 6. Static consonant visemes classification for the Dialect

Consonant	b,p,m	d,t,n	l	g,k,h	j,q,x	zh,ch,sh,r	z,c,s	f,v	pf, pf^{h}	ɳ	ŋ
开口呼	爸	大	拉	哈	机	沙	杂	发	追	女	爱

合口呼		毒	路	姑	句	书	组		吹		

Note: 'v', 'pf', 'pfʰ', 'ɲ', 'ŋ' are phoneme illustrate by IPA and others are Pinyin expression of phonemes.

2.3. Quantitative description of lip static visemes

In the previous section, the lip visemes of Chinese Shaanxi Xi'an Dialect were classified. In order to better measure the viseme, we need a quantitative description of the Chinese Shaanxi Xi'an Dialect lip visemes. Many application fields need to have objective quantitative measures of lip visemes, such as virtual talking head synthesis, speech assistant system, machine automatic lip reading and so on.

2.3.1. Experimental design

The number of speakers depends on the purpose of the study. If used for visual speech synthesis, the average data of multiple speakers is not required, and the personal characteristics of the speaker are necessary in the synthesis. If a lip reading system of a language is studied, it is necessary to find the commonality among multiple speakers. However, beginning with only one speaker can simplify the whole process of data collection and analysis.

The aim of our research is to find suitable parameters and data to reflect the static lip visemes of Chinese Shaanxi Xi'an Dialect. The quality of the visual image is important in this endeavor, so we have created a controlled lighting environment to eliminate the reflection and shadow on the speaker's face, and we repeatedly adjust the contrast of the camera equipment to achieve the best effect.

A Pentax K-30 camera is used to record the video corpus of Chinese Shaanxi Xi'an dialect syllables. The Pentax K-30 is a 16.3-megapixel Pentax digital single-lens reflex camera. It has a stainless steel chassis. It can shoot continuously at up to 6 frames per second with a maximum shutter speed of 1/6000th of a second. It can capture video at 1080p at either 30, 25, or 24 fps. Like all current and recent Pentax DSLRs it features in-body shake reduction, removing the need for each lens to have image stabilisation.

The subject was one adult female – the author of this dissertation – who is a native speaker of Chinese Shaanxi Xi'an Dialect. In this study 26 consonants and 40 vowels as well some prescribed syllables were recorded by this camera for the quantitative analysis of visemes for Chinese Shaanxi Xi'an Dialect.

2.3.2. Analytical method

After recording the videos of phonemes and some prescribed syllables, I can obtain lip visemes images by processing these videos via the software ViedeoPad Video Editor.

First, sub-videos for each phoneme are split from the whole videos I recorded. Then the sub-videos of each phoneme are saved as series frames so that I can get images of lip shape sequence

for the whole articulation process. Finally, I take the snapshot for the sequence to get the central frame from each sub-videos of each phoneme as shown in Figure 2. 1 Images of each lip viseme will be obtained through this processing. Figure 2. 2 shows the image of the central frame for 'ɲ'.

Useful lip visemes parameters can be obtained by processing these series of images we get from the software VideoPad Video Editor. Then the most important things to decide what kinds of parameters are needed for the whole system of the Chinese Shaanxi Xi'an Dialect talking head. In my whole project we chose a 3D model of a talking head similar to the appearance of the Chinese female for our system from Poser pro. Poser pro is a 3D rendering software package for posing, animating and rendering of 3D polymesh human and animal figures. I want to create a phonetically correct, dynamic speech production model and within a parametric model, facial movements such as movements of lip, tongue, chin, eyes eyelids, eyebrows and the whole face are formed.

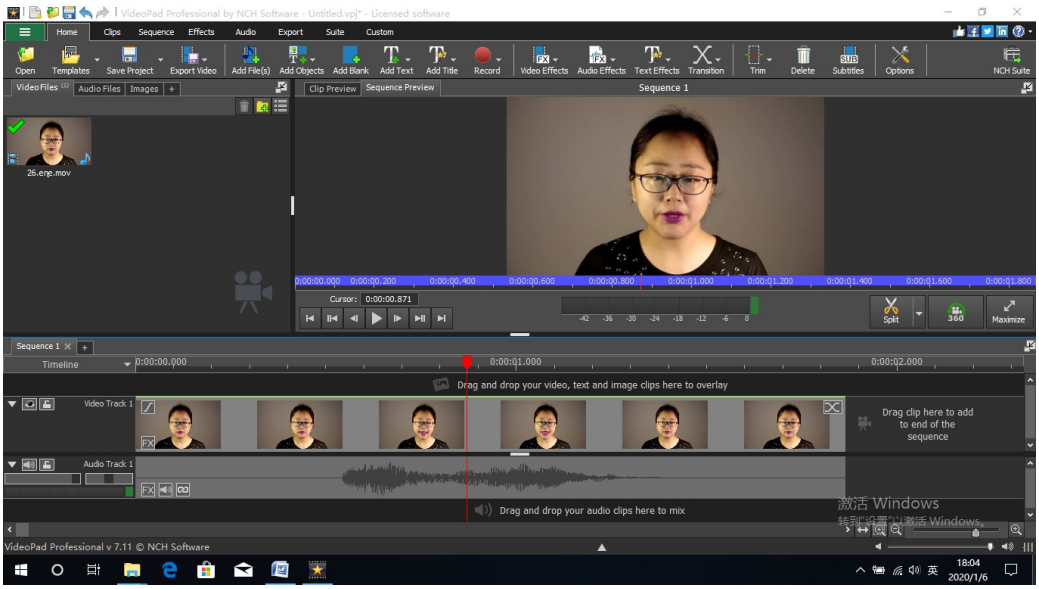


Figure 2. 1. Take the snapshot for the sequence in VideoPad Video Editor

Figure 2. 2. Image of central frame 'ɲ'

This process allows control of a wide range of motions using a set of parameters associated with different articulation functions. These features can be directly matched to particular movements of the lip, tongue, chin, eyes, eyelids, eyebrows and the whole face, and can vary from -1 to 1. Table 2. 7 contains the 3D parameters used for speech animation according to the Poser phrasing [77].

Table 2. 7 Mouth and tongue parameters of 3D speech animation[77]

MOUTH	*TONGUE*
width	length
opening	width
pucker	thickness
pout	tip down
smile	tip up
labiodental	raise
bilabial	retraction
rounding	back up
corner out	rotation up
kiss	
cry	

So, we need to analyze the images obtained of the 26 static lip visemes based on the parameters in Table 2. 7. We can establish the facial animation on lip with these parameters for lip visemes.

2.3.3. Results

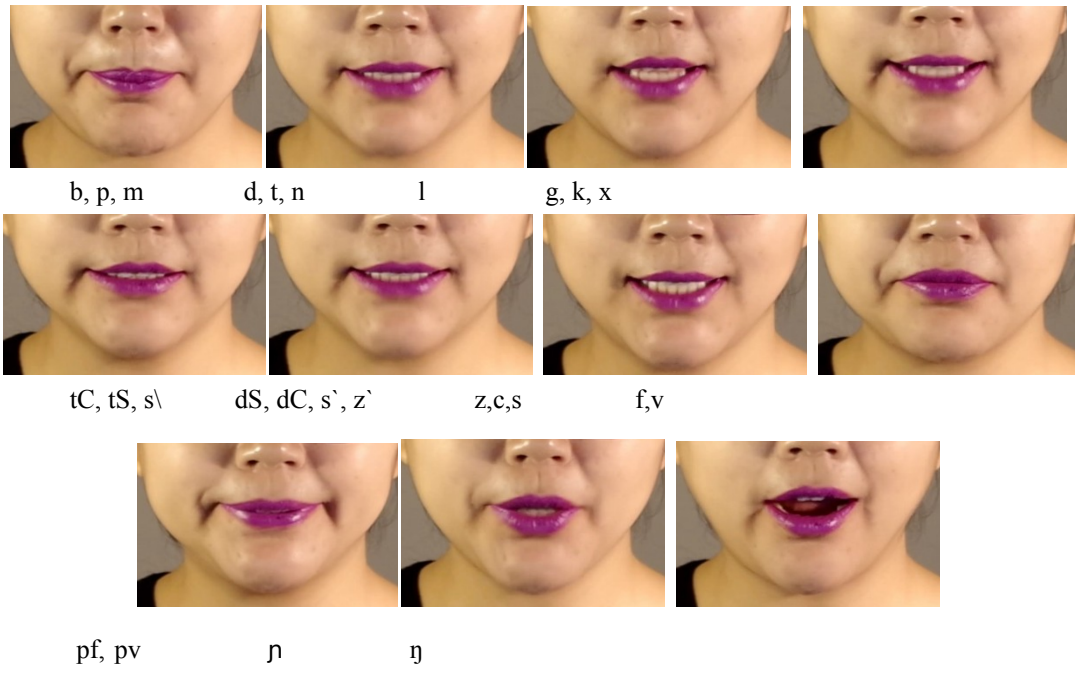


Figure 2. 3. Consonants lip visemes classification results of the Chinese Shaanxi Xi'an Dialect

The series of images of central frame for each phoneme obtained through the analytical method I introduced in Chapter 2.3.2. Consonant and vowel lip viseme classification results of Chinese Shaanxi Xi'an Dialect are obtained based on the set of static viseme based on the characteristics of the Chinese Shaanxi Xi'an Dialect and phonetic composition referring to the classification result of Mandarin static viseme described in Chapter 2.2.2. Figure 2. 3 shows the classification results for consonant lip visemes of Chinese Shaanxi Xi'an dialect, while Figure 2. 4 shows the classification results for vowel lip visemes.

Vowels lip viseme classification results of the Chinese Shaanxi Xi'an dialect correspond to the static viseme classification in Table 2. 5. The compound vowel can be treated as a combination with two or three simple vowels. So the whole vowel lip visemes are obtained since I get the basic vowel lip visemes shown in Figure 2. 4 that the X-SAMPA symbols are used to describe these vowel lip visemes.

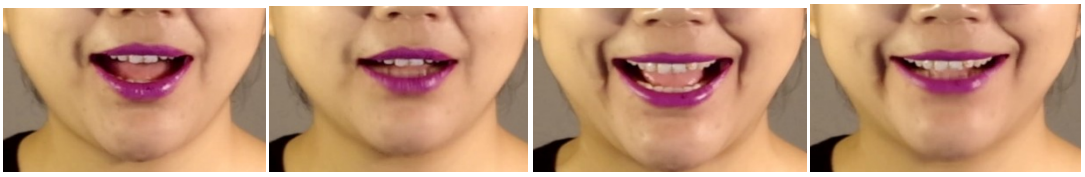


Figure 2. 4. Vowels lip visemes classification results of the Chinese Shaanxi Xi'an Dialect

As I mentioned before the visual counterpart of the shortest acoustic unit, the phoneme, is called a viseme. The set of visemes has fewer elements than that of phonemes as utterances of several phonemes are visually the same. This is because the voiced quality is invisible and the voices of the same place of articulation that are different only in duration or intensity belong to the same viseme class. Figure 2. 5 shows the similarity of the same viseme in the speaker’s photograph and the 3D model.

Figure 2. 5. 3D model of the Chinese Shaanxi Xi'an Dialect (above) Pronunciation of 't' and 'd' by human and virtual speaker (below)

Features controlling the lips and tongue are crucial. Basic lip properties are opening and width, their rate is related to lip roundness. The lip opening and the visibility of teeth are referred to the jaw movement. The tongue is described by its horizontal and vertical position, its bend and shape of the tongue tip.

2.4. Analysis of static tongue visemes of the Chinese Shaanxi Xi'an Dialect

Viseme classes are based on the lip shapes, the invisible tongue position can be different. The motion of the tongue, which is generally not entirely seen, carries an important part of the articulatory information not accessible through lip reading. A speech teaching application should primarily show the motion of the tongue which is hard to see in human-to-human communication. So we need to do tongue movement contour tracking to research both static and dynamic tongue visemes in order to create dynamic modeling systems for the Chinese Shaanxi Xi'an Dialect talking head.

2.4.1. Evolution of tongue movement measurement technologies

The evolution of tongue movement measurement technologies has gone through several stages. Adams et al. introduced an x-ray microbeam system to examine the effects of speaking rate on the velocity profiles of movements of the lower lip and tongue tip during the production of stop consonants Error: Reference source not found. Napadow et al. used a special non-invasive tagging technique that represents tissue as discrete deforming elements in order to quantify local deformation in the human tongue Error: Reference source not found. Electromagnetic tracking systems have been developed by several research groups to study the physiology of speech production. This instrumentation, initially introduced as "electromagnetic medial articulography", uses transmitters surrounding the head and sensor coils attached to multiple points on the tongue, lips and jaw in the mid-sagittal plane Error: Reference source not found Error: Reference source not found Error: Reference source not found Error: Reference source not found. Ultrasound imaging has been used to represent tongue positions for over 15 years. Like other imaging systems, it provides a 2D measurement of the tongue surface contour in a single plane Error: Reference source not found. Ostry et al. used a computerized pulsed-ultrasound system to monitor tongue dorsum movements during the production of consonant-vowel sequences in which speech rate, vowel, and consonant were varied Error: Reference source not found. Stone gave an introduction to and general reference for ultrasound imaging for new and moderately experienced users of the instrument and introduced methods of extracting contours from ultrasound images, displaying and analyzing them Error: Reference source not found. Although there are a number of drawbacks, ultrasound technology is the most attractive method of producing a sequence of images of the tongue in motion because it can provide real-time capture rates, it is non-invasive, convenient for experimentation, and significantly less expensive than other technologies. Alternative methods are too slow to record movement very expensive such as MRI, or they expose subjects to radiation like in X-rays Error: Reference source not found.

2.4.2. Subjects and speech material

In this thesis I record a small-scale visual speech database using combinations of the consonants and vowels of the Chinese Shaanxi Xi'an Dialect. The tongue movement contour is tracked through processing of the ultrasound image in the speech database, while the viseme system for the Chinese Shaanxi Xi'an Dialect determined through dynamic analysis. In this thesis I focus on how to describe the static visemes of tongue.

The subject was one adult female – the author of this thesis – who is a native speaker of Chinese Shaanxi Xi'an dialect. I have two-speech corpora, covering all phonemes involved in our experiment (26 consonants and 40 vowels of Shaanxi Xi'an dialect).

2.4.3. Tongue movement recording method

The method used in my present study is recording a small-scale visual speech database such as the special structure of the combination of the consonants and vowels of Chinese Shaanxi Xi'an Dialect, tracking the tongue feature point by processing the ultrasound image of the speech recorded as the speech database and investigating the viseme system for the Chinese Shaanxi Xi'an Dialect through dynamic analysis.

More precisely, our approach is composed of three steps: first, localizing the tongue and tracking its motion; secondly, extracting precise and pertinent visual features from the speaker's face; finally, using the extracted features for viseme classification and recognition.

I use the 'Micro' ultrasound system (Articulate Instruments Ltd.). a speech recording instrument with a 2–4 MHz/64 element 20 mm radius convex ultrasound transducer at roughly 82 fps. The angle of view was 90°; there were 842 pixels in each of the 64 scanlines in the raw data. An ultrasound stabilization headset (Articulate Instruments Ltd.) fixed the transducer during recording. An Audio-Technica ATR 3350 omnidirectional condenser microphone was set approximately 20 cm from the lips when recording the speech samples.

A photograph of this instrumentation is presented on the left side of Figure 2. 6, while the Probe Stabilization Headset is shown in place on the right side. The headset was individually fitted with the main goal of obtaining an image appropriate for phonetic analysis [88]. The headset fixes the subject's head while capturing the images and also fixes the ultrasound transducer under the chin [89].

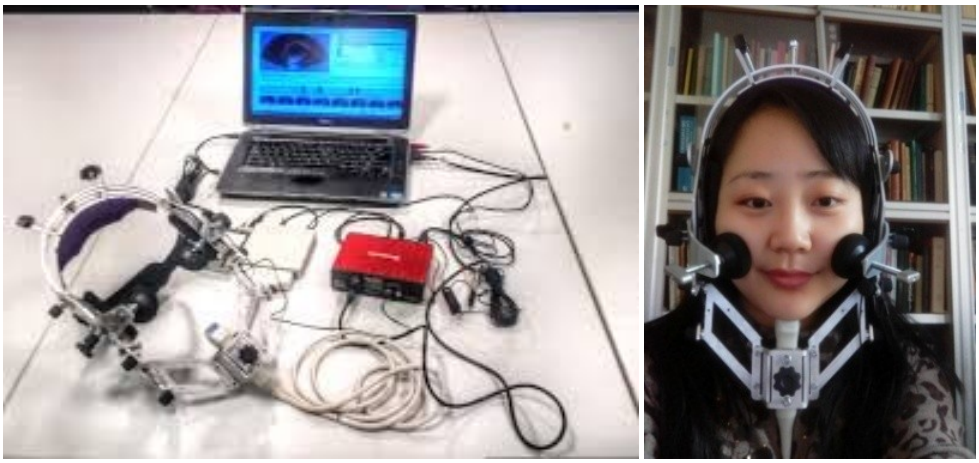
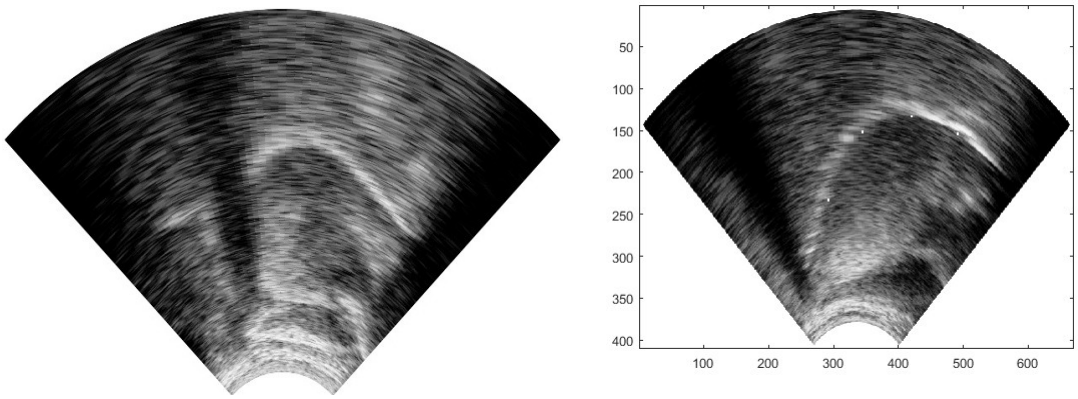


Figure 2. 6. Left: 'Micro' Ultrasound system. Right: Probe stabilization headset installation

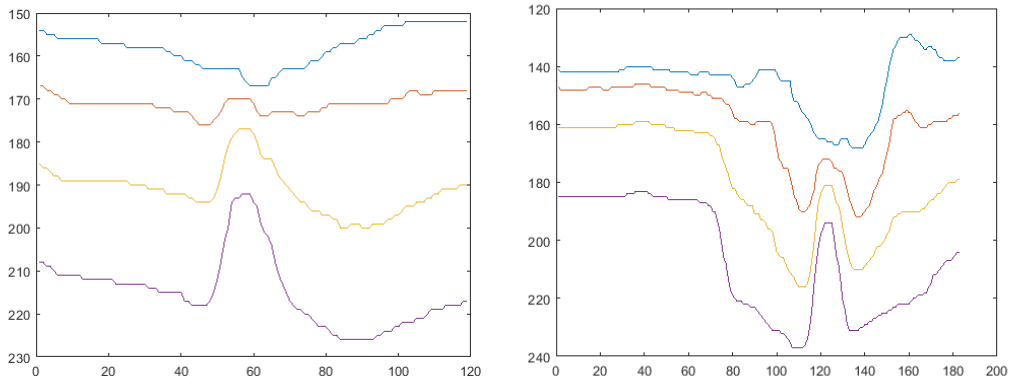
2.4.4. Tongue movement contour tracking

After recording the ultrasound images of phonemes and sentences, I can obtain tongue feature points by processing these images. The approach for tongue contour tracking is that the tongue pixels have a different feature compared to those of other parts of the mouth pixels. I created an algorithm to get the tongue feature points based on this concept. In fact, this type of methods identify the rapid changes in locations of the zones of interest and make some very simple measures of it, such as horizontal and vertical position of the tongue contour.

All of the tongue feature points tracking for each frame of the speech are gained based on this algorithm. Figure 2. 7 presents an example of tongue contour tracing in Matlab for 'm' in the syllable structure 'eme'. This is a basic task to build a dynamic visual system and dominance classification for Chinese Shaanxi Xi'an Dialect. Figure 2. 7(a) shows an original image of the static single frame 'm' in the syllable structure 'eme' of the Chinese Shaanxi Xi'an Dialect speech corpora while Figure 2. 7(b) shows the tongue feature point tracking image for the same situation. The left image describes the pixel map of the tongue contour by dealing with the ultrasound images in Figure 2. 7(a) recorded by the test instrument from -45° to 45° (the full measuring range of the measuring instrument). The right-hand image describes the pixel map of the tongue contour gained by the transformation from the left-hand image of Figure 2. 7(b), converting the tongue contour map based on rays coordinates in the radial-shaped area from -45° to 45° into 90° Cartesian coordinates. Figure 2. 7(c) shows the tongue contour changes on five certain parts of the tone represented from tongue tip to tongue root according to the whole articulation of the syllable structure 'VCV' and 'CVC'. The figure shows the ordinate of the pixel.



(a) Original image (b) Four feature points of the tongue image



(c) The trajectory of tongue feature points in the phrase 'ede' and 'ada'

Figure 2. 7. Tongue contour tracing in Matlab

2.4.5. Analysis of static tongue visemes

The contour of static tongue visemes was obtained by dealing with the ultrasound images recorded in the experiment based on the algorithm I developed.

First, the frames of the consonant or vowel in spectrum audio speech being focused on are chosen manually in the software Praat (doing phonetics by computer) and the corresponding order of the frame image structure is found in the collection of JPG images of the ultrasound frames recorded by the 'Micro' ultrasound system. Then the tongue feature point of the same phoneme in paired structure is traced and the tongue contour of the phoneme is made in Matlab based on the tongue contour tracing algorithm. Figure 2. 8 is the central frame of 'm' in the syllable of the Dialect speech corpora chosen in Praat. Figure 2. 9 show the tongue contour in the central frame of 'm' in the syllable 'eme'. The blue curve shows the tongue contour of central frame of 'm' in syllable structure 'eme' and the real shape is mirrored along the x-axis.

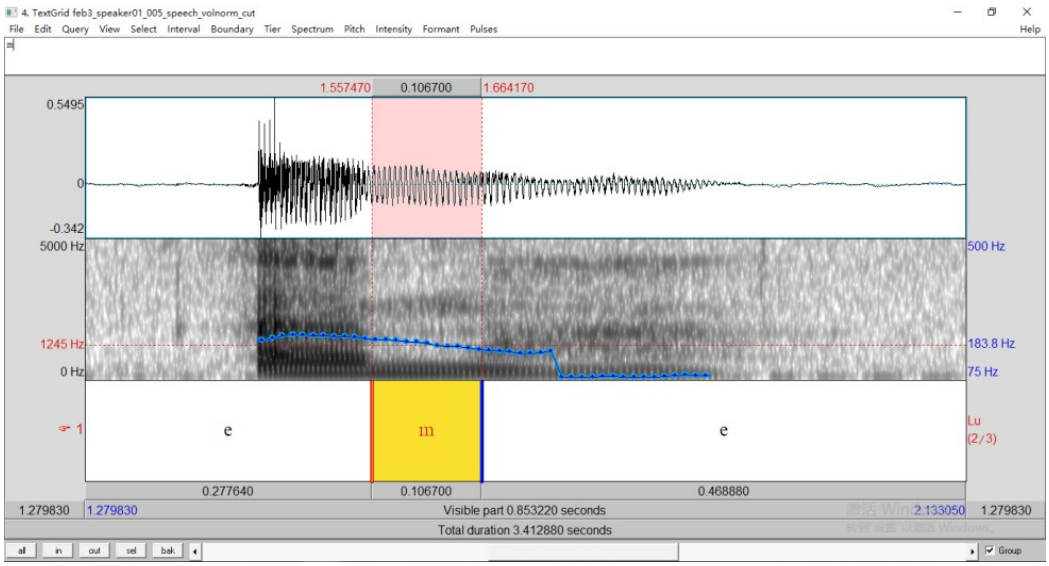


Figure 2. 8. Central frame of 'm' in the phrase 'eme' chosen in Praat

The process involves the following steps: Open the documents of the syllable structure 'eme' in the Chinese Shaanxi Xi'an dialect speech corpora in the software Praat and find the time position of the central frame 'm' manually. Then calculate the frame order of consonant 'm' in this syllable. By processing with the algorithm I designed in the Matlab, I can gain the tongue contour of the central frame of 'm' in the syllable structure 'eme'.

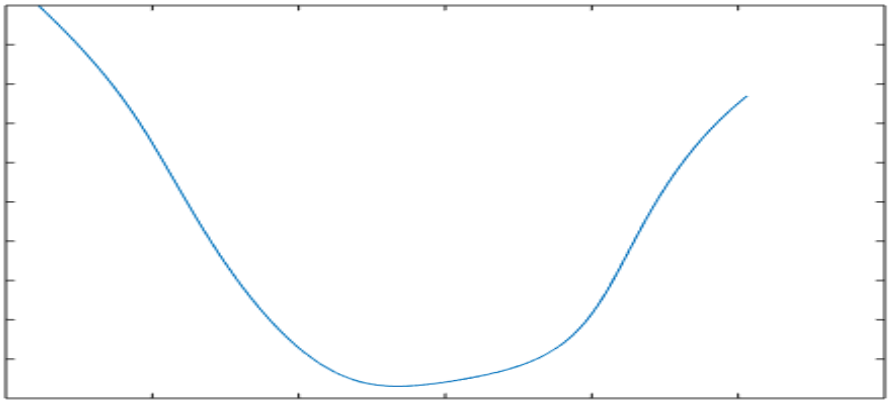


Figure 2. 9. Tongue contour of central frame of 'm' in the phrase 'eme'

As we can see, the tongue contour of the central frame is showed in Figure 2. 9. We can get the static viseme of the tongue for each phoneme at each frame though this processing method.

2.5. Thesis

In this thesis, a static lip viseme classification of Chinese Shaanxi Xi'an Dialect speech was created. I give a quantitative description of the dialect lip visemes. I carried out an experiment with the purpose to study both the static lip and tongue visemes features of the Chinese Shaanxi Xi'an Dialect. I developed an algorithm to automatically track spatial-temporal tongue movement contours from the ultrasound images [90] [109].

2.5.1. Novelty

I gained a quantitative description of Chinese Shaanxi Xi'an Dialect lip visemes by processing the videos of lip visemes recorded by camera. In addition I describe the method to do research using tongue contour tracking for the Chinese Shaanxi Xi'an Dialect speech and set an example to show how to analyze static tongue visemes. I propose a system for preparation to create a dynamic viseme model for visual speech synthesis. This thesis gives a very comprehensive analysis of the static lip and tongue visemes of Chinese Shaanxi Xi'an Dialect. It provides great practical value for Chinese Shaanxi Xi'an Dialect application research, especially for dynamic modeling of talking head for this Dialect.

2.5.2. Measurements

For static lip viseme analysis, the method is processing the images of lip visemes taken by camera classified by categories to gain the set of 3D parameters used for speech animation according to the Poser phrasing.

For static tongue viseme analysis, the method is processing the ultrasound images of tongue taken by the 'Micro' ultrasound system (Articulate Instruments Ltd.) which is a speech recording instrument. Then the data obtained in the Matlab by the algorithm I developed help us to gain the set of 3D parameters used for speech animation according to the Poser phrasing.

All of the data we obtain by processing the images of lip and ultrasound images of tongue are the basis to create for a dynamic viseme modeling system for Chinese Shaanxi Xi’an Dialect. This part will have a detailed description in Chapter 3.

2.5.3. Limits of validity

Co-articulation of visemes in context has not yet been taken into account in the static viseme analysis. This will be detailed illustrated in the following chapter.

2.5.4. Consequence

The long-term objective is to create a dynamic viseme model that is applied to animate articulation for Chinese Shaanxi Xi’an dialect speech within a 3D virtual talking head. This is intended for use in a speech assistant system for hard-of-hearing children and second language learners. So this thesis supplies a very strong foundation and makes a full preparation for dynamic viseme modeling system.

2.5.5. Related published paper

Lu Z, Czap L.: Modelling the tongue movement of Chinese Shaanxi Xi'an dialect speech[C]. 2018 19th International Carpathian Control Conference (ICCC). IEEE, 2018: 98-103.

Zhao L, Czap L.: Visemes of Chinese Shaanxi Xi’an Dialect Talking Head[J]. Acta Polytechnica Hungarica, 2019, 16(5): 173-193.

Chapter 3

Dynamic Modeling of the Chinese Shaanxi Xi'an Dialect Speech

3.1. Introduction

With the development of computer technology and talking head modeling technology, the research on talking head animation has made some progress. Realistic talking head animation has become a hot spot in the field of audiovisual information processing and multimedia research. It can be widely used in human-computer interaction, video conferencing, hearing aids, videophones, game scenes and special effects in movies. The visual dynamic articulation model – the interface between speech and animation – is a vital link in talking head animation. It is the model for mapping continuous speech to corresponding visemes, and determines the acceptability and comprehensibility of a talking head animation synthesis.

3.1.1. The main methods and problems of current visual speech research

Voice-synchronized talking head animation is classified into three main modes: text-driven [91], voice-driven [92] and hybrid driven [93]. Text-driven talking head animation converts the text information pre-processed by segmentation into speech and synchronously generates synthesized speech and animated articulation through text input, but the relative lack of rhythm, tone and other information of speech affects, the naturalness of speech synthesis and there is a certain distance compared to natural speech. Voice-driven talking head animation directly extracts phonetic information such as phoneme information, duration, and rhythm from the input of natural speech, and establishes the association map of phoneme to animation, driving the generation of synchronized talking head animation. Its disadvantage is in continuous language. In the stream, the recognition rate of phonemes is not always satisfactory, which hinders the accurate extraction of phonemes, thus affecting the accuracy of animation generation. Therefore, with current speech recognition technology it is difficult to achieve a high level of accuracy. Hybrid driven input has both text information and speech signals, through the fusion of multiple information source it can take advantage of various technical methods. For the construction of a talking head animation needs a variety of information to enable a more natural and reliable talking head animation.

At present, there are still many technical difficulties to be solved in the realistic sense of speech animation in continuous stream state. Modeling of visual collaborative articulation is one of the

main problems. Massaro et al., through years of research on phonetics, proposed a dynamic articulation model based on weight fusion [94], which models the visual phenomenon of dynamic visemes and visual collaborative methods in a continuous stream. Bregler et al. proposed Video Rewrite technology, and use context-sensitive phoneme model to find the best matching action with a massive database and audio information [95]. In this approach an automatic animation frame is generated, but the disadvantage is that it requires a large database to be constructed, which requires a large synthesis cost. Pelacaud et al. model the cooperative articulation phenomenon by establishing a corresponding rule set [96], and obtain a relatively simple model. It is helpful to establish real-time speech-visual mapping. However, due to the limitations of the scope of the rule set, it is difficult to find a satisfactory rule set to model all collaborative articulation processes and collaborative articulation phenomena.

In the development of speech animation modeling, under the joint efforts of phonetics researchers and face synthesis researchers, the visual dynamic articulation modeling is changing with each passing day. In order to achieve practical goals, the current visual dynamic articulation modeling is moving forward. The simplification of the model is developing in combination with the accuracy of the phoneme-visual mapping.

3.1.2. Research results in China

Research on talking head animation started late in China, and the development is not extensive. The current research results are still very limited, but some research institutions have also achieved some good results, such as Tsinghua University's data-driven video TTV system [97], Beijing University of Technology's Chinese Face Animation System and Zhejiang's imaginary mixed-drive face-up speech animation system [98]. Some publications do not involve the study of the Chinese visualized dynamic articulation model, some systems involve it in terms of research, but do not conduct in-depth research, the applicability of the obtained models are still very limited. Therefore, solving the problem of Chinese visualized collaborative articulation modeling is promising and challenging.

3.2. Taking coarticulation into account

There is a phenomenon of coordinated articulation in the continuous speech stream, which is expressed by the fact that the current phoneme articulation is influenced by the phoneme before and after it. The visual representation of the lips and tongue is not only dependent on the current articulation, but also depends on the previous and subsequent articulation. When a person is speaking, he does not articulate viseme by viseme, rather during the articulation of the current viseme, the organ is already preparing for the articulation of the next viseme, gradually transitioning to the phoneme of the next viseme. The coarticulation has a certain influence on the current viseme. Therefore, in the continuous speech stream, a complete dynamic aritculation involves the current phone, pre-phones and post-phones holding the original corresponding lip shape and tongue position, but also variations on this: some phonemes deviate from the original

lip and tongue position while some phonemes deviate with the extent of deviation before and after different contact with different phonemes. Visual coarticulation visually makes reaction of the relevant ministry of the lip and tongue position change mainly in continuous language stream. For example, the viseme lip and tongue position are different for the consonants 'b', 'h', and 'ch' in different articulation syllables. Therefore, in the process of articulation, the same phonemes may have different lip shapes in different syllables due to the effect of dynamic articulation [99].

3.3. Research method and processes

In this chapter, I present my research on the dynamic articulation in a continuous speech stream of Chinese Shaanxi Xi'an Dialect. I will synthesize a realistic speech animation based on the rule set and the visual dynamic articulation model based on the dominance model. First, according to their visual similarity static visemes, are classified to reduce the number of key lips and tongues to be synthesized. Secondly, a set of rules related to the visual representation of the context-related synergy is established. The rule set describes the current state of knowledge of the phonemes in the Chinese Shaanxi Xi'an Dialect, which is influenced by the phonological factors. Oral tongue mapping is, combined with a 3D talking head model to determine the key tongue shape of each key frame in the dynamic visual speech animation. Finally, through the construction of the transition lip tongue database sample, the transition between the two phonemes takes place.

3.3.1. Dominance classification concept

In connected speech, some parameters take on their characteristic value, while others do not reach their target value during pronunciation, especially in fast speech. All features of the talking head determining the lip shape and tongue position have been grouped according to their dominance, with every articulation feature of each speech sound being assigned to a dominance class [100]. This is different from the general approach where only the phonemes are classified by their dominance. Features of the parametric model can be divided into four dominant grades:

• stable — co-articulation has no effect on them (e.g. tongue position of alveolar plosives, lip shapes of bilabials),
• dominant — co-articulation hardly affects them (e.g. lip shapes of vowels),
• flexible — the neighboring sounds affect them (e.g. tongue positions of vowels),
• uncertain — the neighborhood defines the feature (e.g. tongue position of bilabials, lip shapes of 'h' and 'r').

The speech-synchronized visual animation is synthesized from the candidate synthesized transitional tongue database and combined with the phoneme sequence obtained by the annotation and the track time information obtained from the speech. When handling the dominance classes:

- Stable features do not change, neither smoothing nor pre-articulation filters are applied.
- Dominant features will slightly adapt to their neighbors by a pre-articulation filter.
- Flexible features will adapt to previous and next stages by an averaging smoothing filter.

•Uncertain features do not have their own values; they are interpolated by neighboring features.

After the interpolation process a pre-articulation filter is applied except for the stable features. The pre-articulation filter is an averaging filter applied for the previous and next frames. Previous frames show preparation for the next viseme before it is said.

3.3.2. Research method

In this thesis, since the motion of the tongue – which is generally not entirely seen – carries an important part of the articulatory information not accessible through lip reading, I do some research work concentrating on the visual configuration of the tongue. First, the static viseme classification of Chinese Shaanxi Xi'an Dialect is illustrated according to the method that carried out to classify Mandarin static visemes in the previous chapterError: Reference source not found. Then we use the 'Micro' ultrasound system to record the speech materials of Chinese Shaanxi Xi'an Dialect speech with VCV and CVC sequences in different tempos which eventually form the JPG images and MP4 videos after processing with the Assistant Advanced software (Articulate Instruments Ltd.). Furthermore, I have developed an algorithm that can automatically carry out a spatial-temporal tracking of tongue movement contour from the ultrasound images we recorded. The extracted visual information gained by this method is then classified in order to define the uttered viseme and is used to create dynamic viseme system of tongue motion for the Chinese Shaanxi Xi'an Dialect. This dynamic viseme system is used to create the fundamentals of a talking head – and animated articulation model for Chinese Shaanxi Xi'an Dialect speech Error: Reference source not found [103]. We conclude that this talking head is intended for utilization in a speech assistant system for speech retarded children, in the perception and production rehabilitation of hearing impaired children, and in pronunciation training for second language learners.

Two structures were investigated, VCV and CVC, covering all phonemes involved in our experiment (the 26 consonants and 40 vowels of Shaanxi Xi'an dialect), following the method introduced in Chapter 2.3.1. In the VCV structure, 'e' and 'u' (eCe and uCu) are used to compare the different dominance features of the same consonant. Similar tongue positions mean high dominance, while different values mean low dominance. In the CVC structure, 'sh' and 's' are the two phonemes used (shVsh and sVs) to compare the different dominance features of the same vowel. These phonemes were chosen for the database because of the rear tongue articulating the phonemes 'u' and 's' and front position when articulating 'e' and 'sh'.

3.4. The interaction between phonemes and the corresponding lip shape [99]

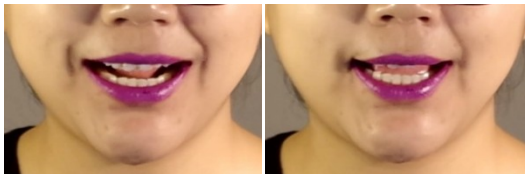
Studies have shown that the phenomenon of Chinese Shaanxi Xi'an Dialect coarticulation is very obvious in the change of lip shape. Among them, the consonants are affected by the vowels followed by them, and the influence between the vowels is particularly obvious. Different consonants are affected by the same vowel. For example, 'h' in 'ha' is more affected by 'a' than

'b' in 'ba', and the characteristic of 'h' has largely disappeared from the lip shape and turned to the feature of viseme 'a'. The same vowel is affected by different vowels, and even has trans-syllable effects. For example, 'e' in the next syllable of 'da de' and 'wu de' is affected by the vowel in the previous syllable, and the expression of the lip viseme is not the same.

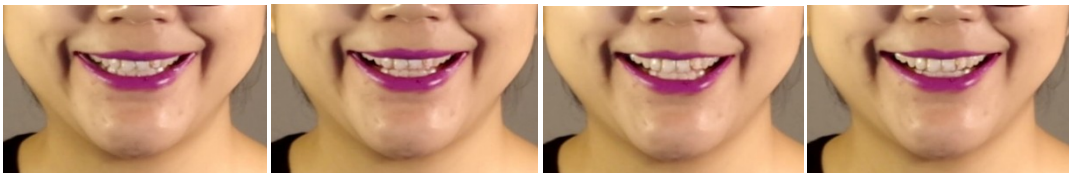
3.4.1. Method towards interaction between lip visemes

Videos include the two structure 'VCV' and 'CVC' covering all phonemes – 26 consonants and 40 vowels of the Chinese Shaanxi Xi'an Dialect involved in our experiment taken by camera. In structure 'VCV', 'e' and 'u' are two phonemes used(eCe and uCu) to compare the different affect levels of the same consonant while in the structure 'CVC', while 'sh' and 's' are two phonemes used(shVsh and sVs) to compare the different affect levels of the same vowel. These we chosen because the tongue position is rear when articulating the phonemes 'e' and 'sh' while the tongue position is in front when articulating the phonemes 'u' and 's'. So it is easy to find the affect level of the frame we focus on.

Central frame in these two structures could be extracted to compare by the lip parameters width and openess to make sure whether the middle phoneme in this structure of syllable is affected, and if so, how much. Two major types of examples for vowels and consonants will be illustrated there.

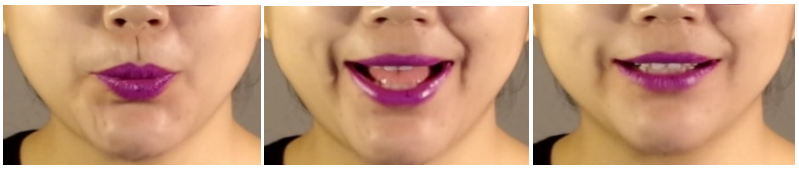


(a) Lip visemes of central frame of 'a' in the phrase 'shash'(left) and 'sas'(right)

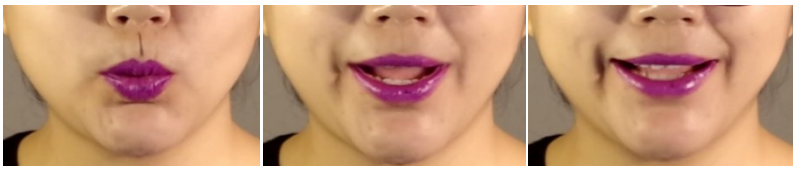


shẽsh (e,˜) sẽs (e,˜)

(b) Lip visemes of central frame of 'ẽ' in the phrase ' shẽsh'(left) and ' sẽs '(right)



shuæ˜sh(u,i,æ,˜)

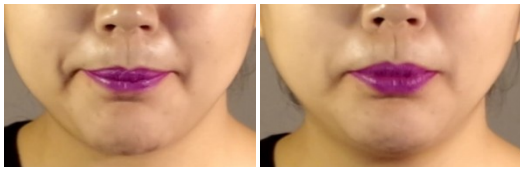


suæ˜s (u,i,æ,˜)

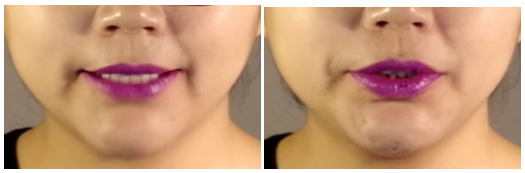
(c) Lip visemes of central frame of ' æ˜' in the phrase ' shẽsh'(left) and ' sẽs '(right)

Figure 3. 1. Dominance grade images for vowels of lip visemes

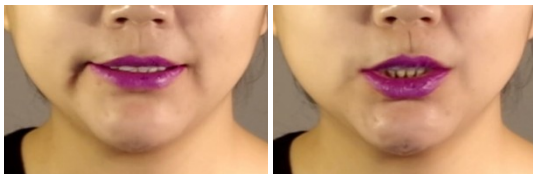
Figure 3. 1 shows the lip visemes of central frame of different types of vowel visemes. As can be seen in Figure 3. 1(a) (b), the length and openess of the lip viseme 'a', ẽ are similar in different syllable structure 'shVsh' and 'sVs'. I treat the viseme 'uæ˜' as the combination of three simple phonemes and make the comparison of each corresponding phoneme in syllable structure 'shVsh' and 'sVs'. According to Figure 3. 1 (c), the length and openess of corresponding phoneme are also similar.



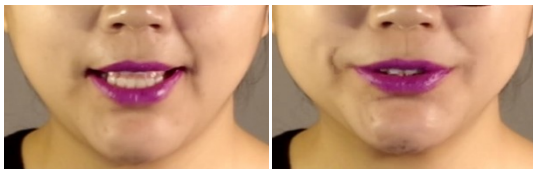
(d) Lip visemes of central frame of 'b' in the phrase ' ebe'(left) and ' ubu '(right)



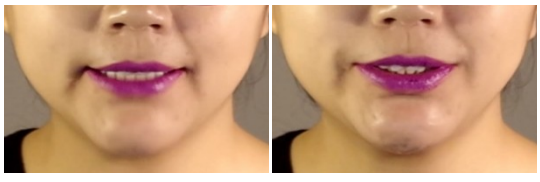
(e) Lip visemes of central frame of 'd' in the phrase ' ede'(left) and ' udu '(right)



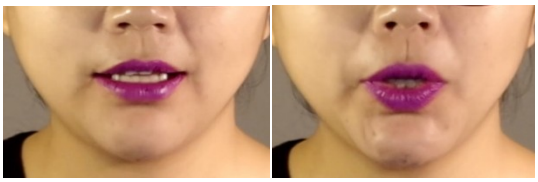
(f) Lip visemes of central frame of 'r' in the phrase ' ere'(left) and ' uru '(right)



(f) Lip visemes of central frame of 's' in the phrase ' ese'(left) and ' usu '(right)



(g) Lip visemes of central frame of 'zh' in the phrase ' ezhe'(left) and ' uzhu '(right)



(h) Lip visemes of central frame of 'h' in the phrase ' ehe'(left) and ' uhu '(right)

Figure 3. 2 Dominance grade images for selected consonants of lip visemes

Figure 3. 2 shows the lip visemes of central frame of different types of consonant visemes. As you can see, the length and width of different consonant lip visemes are not same in the different syllable structures 'eCe' and 'uCu'.

3.4.2. Results of dominance grade for lip visemes

From the difference in the degree of influence of the vowels on the corresponding oral visual elements, I reached the conclusion that all the features of vowel visemes are dominant.

The consonants are divided into four different dominance grades based on the degree of influence by the vowels before and after it. Table 3. 1 shows the different dominance grades of lip visemes of the Chinese Shaanxi Xi'an Dialect.

Table 3. 1. Different dominance grade of lip visemes

	lipswidth	*lipsopen*
b, p, m	uncertain	stable
f, v	uncertain	stable
t, d	uncertain	dominant
g, k, n, l	uncertain	flexible
s, z, c	uncertain	dominant
sh, zh, ch	uncertain	dominant
r, h	uncertain	uncertain
j, q, x	flexible	dominant
pf, pf^h	uncertain	stable
ŋ, n̲	flexible	dominant
vowels	dominant	dominant

3.5. Dynamic tongue viseme classification

A static viseme can be show by a still human face picture. But when we pronounce a phoneme, the movement of our tongue is more like a dynamic process than a static state. So we introduce here the concept of dynamic tongue model, which represents the whole process of the physical organ movement during the pronounciation of a given phoneme. Like Cohen's co-articulation model, the dynamic viseme model is a composite of dominance and parameter value[104].

Analyzing the phonetic aspects of Shaanxi Xi'an dialect is the basic research for creating its dynamic speech production model. Facial movements are produced within a parametric model, i.e. a collection of polygons is manipulated using a set of parameters. This process allows control of a wide range of motions using a set of parameters associated with different articulation

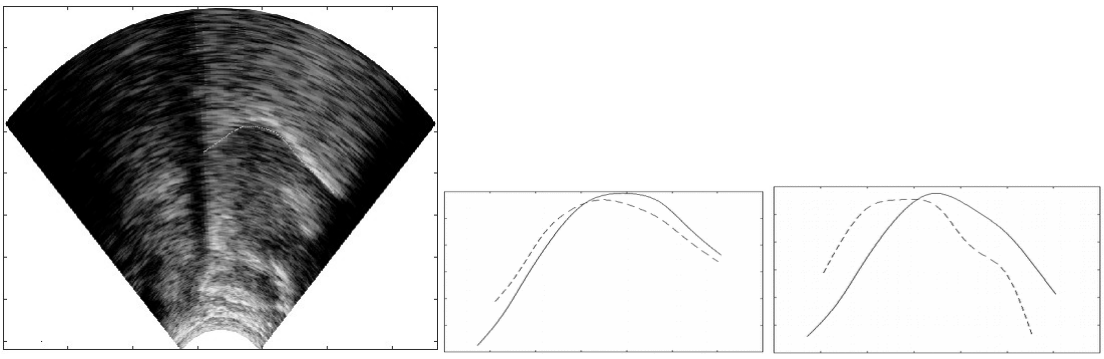
functions. A full description of the co-articulation effects, determination of the nature of certain dominant characteristics, and the refinement of interpolation rules between the parameters were defined.

Problems revealed in biomedical image analysis such as user fatigue, user bias, and difficulty in reproducing results may also occur when manually tracking tongue contours [105]. We thus developed an algorithm to extract and track 2D tongue surface contours from ultrasound sequences in the 'Micro' ultrasound system. Traditionally, visemes are defined as a set of static mouth and tongue shapes that represent clusters of contrastive phonemes [106]. However, the movement of phoneme pronunciation is less a static state and more a dynamic process. Here we present the concept of the dynamic viseme, representing the entire process of organ motion during articulation of a given phoneme. Similarly to the co-articulation model of Cohen [107], our dynamic viseme model blends dominance and parameter values.

3.5.1. Tongue dominance classification for the Chinese Shaanxi Xi'an Dialect

The phonemes of Chinese Shaanxi Xi'an Dialect are described in Chapter 1, A visual speech database of the Chinese Shaanxi Xi'an Dialect was set up containing data such as the special structure of the combination of the consonants and vowels and sentences in different tempos recorded by the 'Micro' ultrasound system introduced in the previous section and a collection of JPG images of the ultrasound frames gained from this system.

Two structures 'VCV' and 'CVC', are recorded covering all phonemes involved in our experiment – the 26 consonants and 40 vowels of the Chinese Shaanxi Xi'an Dialect using the recording method introduced in Chapter 2.4.3. In the ultrasound recording structure 'VCV', 'e' and 'a' are the two vowels used(eCe and aCa) to compare the different dominance features of the same consonant, while in the structure 'CVC', 'k' and 't' are the two consonants used(kVk and tVt) to compare the different dominance features of the same vowel because the tongue position is rear when articulating the phoneme 'a' and 'k' while the tongue position is in front when articulating the phoneme 'e' and 't'. Thus it is easy to find the dominance features of the frame we focus on.



a b c

Figure 3. 3.

a: Sample ultrasound image with tongue contour tracking; b: Tongue contours of 't' in 'ete' (—) and 'ata' (- -); c: Tongue contours of 'p' in 'epe' (—) and 'apa' (- -)

Two central frames of the same consonant or vowel in the audio speech spectrum of a paired structure are selected after manual segmentation in Praat. JPG images of the ultrasound frames are analyzed. Then I trace the tongue feature points of the targeted phoneme in the paired structure and compare the tongue contour of the same phonemes in both structures in MATLAB using the algorithm to trace tongue contour. In Figure 3. 3, (b) and (c) show the tongue contour comparison in the frame belonging to the burst of 't' and 'p' in the two structures.

The continuous curve shows the tongue contour in that frame for 't' in the structure 'ete' and 'p' in 'epe', while the dashed line shows the tongue contour of 't' in the structure 'ata' and 'p' in 'apa'. The dominance feature of the invisible tongue tip of 't' is classified as stable, while the tongue position of 'p' is uncertain, approaching that of the neighboring sounds. In future research, I plan to focus on a sequence of frames for more accurate classification of the dominance grade of viseme features.

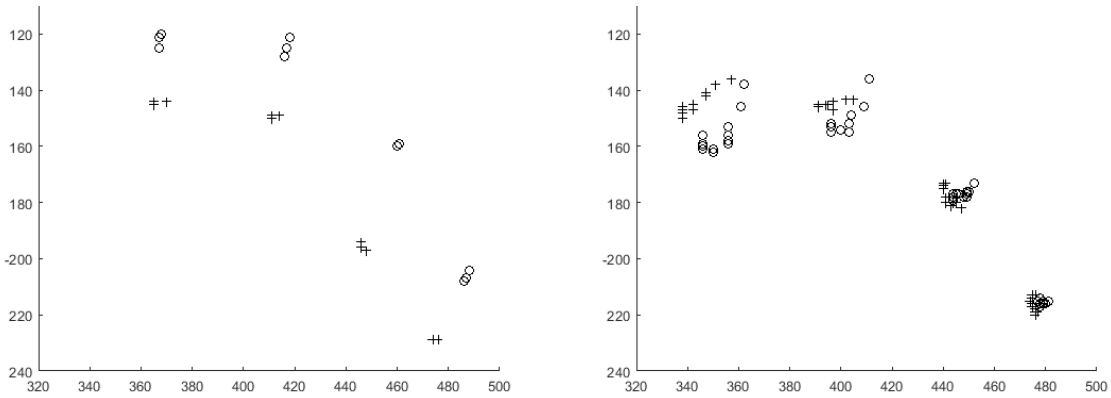
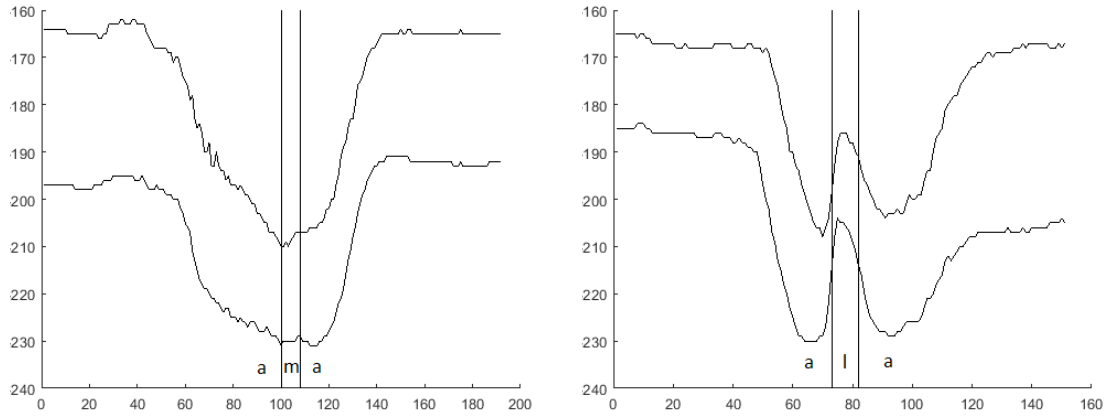


Figure 3. 4.

a: positions of the four feature point of sounds 'p', b: 'sh' in the environment of 'e' (o) and 'a' (+)

In Figure 3. 3, the complete tongue contour that can be seen in the ultrasound image is shown after automatic contour tracking, the uneven curve has been smoothed with discrete cosine transformation filtering. The description of the smoothed tongue contour makes it possible to draw further conclusions on the basis of the selected feature points of the curve. Four feature points were selected at 20, 40, 60 and 80% of the arc of the smoothed curve. In Figure 3. 4 (a), the positions of the feature points of the sound 'p' in VCV words 'apa' and 'epe' can be seen for the three image frames before the burst (altogether 36 ms). Figure 3. 4 (b) shows the position of the feature points of the sound 'sh' in words 'asha' and 'eshe' for the whole range of the sound. The uncertain character of 'p' and the dominant character of 'sh' can be seen very well. (Similarly to Figure 3. 3, on the left hand side, the back and on right hand side, the front of the

tongue can be seen. The numbers of rows increase from top to bottom, as it is usual in the representation of images.)



a b

Figure 3. 5.

Vertical position of the first two feature points of the tongue when a word is uttered a: 'ama' and b: 'ala' are being uttered

The movement of the tongue during speech can be described with the changes in the coordinates of the feature points. Figure 3. 5 shows the vertical positions of the two front feature points while the VCV words 'ama' and 'ala' are being uttered. This representation not only shows the uncertain character of 'm' and the dominant character of the vertical position of the front part of the tongue in case of 'l' but also makes it possible to investigate of the interpolation between key frames.

Dominance is analyzed for all the features involved in the animation. The bilabial sounds in the previous examples are, e.g. stable as regards the openess of the lips but are of uncertain character as regards the position of the tongue.

The standard deviation of the feature examined combines the essence of the analyses shown in thesis: the greater the deviation, the lower the dominance. Our animation process, elaborated for the Hungarian language, accomplishes the screening of the features according to dominance class.

This method will be used to determine the dominance grade for all viseme features of the dialect, so that we can create a dynamic viseme system using the tongue contour with a dominance model.

3.5.2. Results of dominance grade classification for tongue visemes

From the difference in the degree of influence of the vowels on the corresponding oral visual elements, I conclude that all the features of vowel visemes are dominant.

The consonants are divided into four different dominance grades based on the degree of influence by the vowels before and after it. Table 3. 2 shows the different dominance grades of tongue visemes of the Chinese Shaanxi Xi'an Dialect.

Table 3. 2. Different dominance grade of tongue visemes

	Tongue position
b, p, m	uncertain
f, v	uncertain
t, d	stable
g, k, n, l, r	dominant
s, z, c	dominant
sh, zh, sh	dominant
j, q, x	dominant
pf, pf^{h}	dominant
ŋ, n̲, h	dominant
vowels	dominant

3.6. Results of face animation on the Chinese Shaanxi Xi'an Dialect talking head

I gained the basic face animation based on the Chinese Shaanxi Xi'an Dialect talking head created in the 3D rending software Poser Pro and the parameters of the basic lip and tongue visemes obtained in dynamic modeling system of Chinese Shaanxi Xi'an Dialect visual speech in my thesis. Figure 3. 6. shows the face animation of Chinese Shaanxi Xi'an Dialect talking head in software Poser Pro (left) and Hungarian speech assistant system (right). I created face animation in continuous speech involves not only lip movement, but also face mimics and movement of chin/eyebrows and head is showed on the left of Figure 3.6. Dynamic modeling of corresponding visemes focus on lip and tongue of Chinese Shaanxi Xi'an Dialect talking head applied in Hungarian speech assistant system is showed on the right of Figure 3.6.

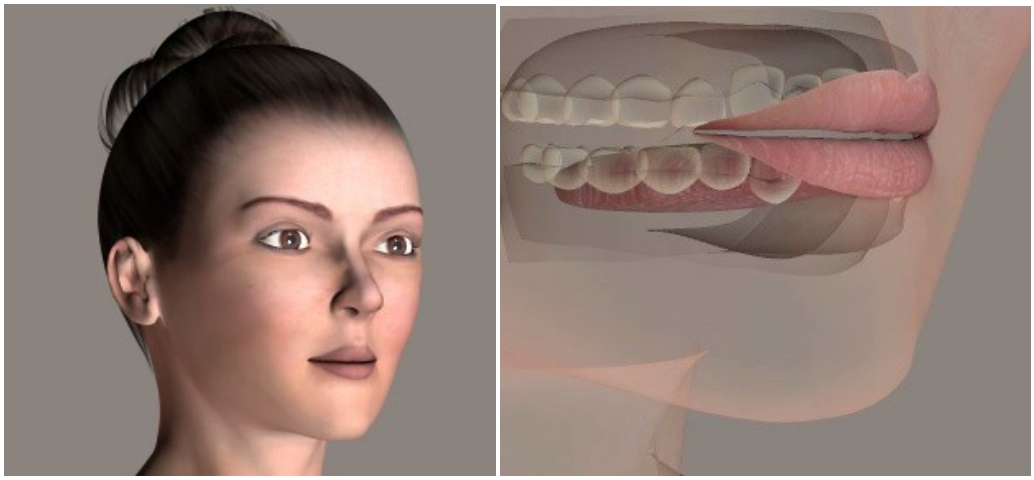


Figure 3. 6. Face animation of Chinese Shaanxi Xi'an Dialect talking head in software Poser Pro (left) and Hungarian speech assistant system (right)

Table 3. 3. shows a couple of characteristic values for mouth and tongue visemes in the software Poser Pro when I created dynamic face animation for the Chinese Shaanxi Xi'an Dialect talking head.

Table 3. 3. A couple of characteristic values for mouth and tongue visemes

MOUTH				
width	i	0.7	u	-0.4
opening	a	0.8	i	-1.3
pucker	u	1.1	a	0.3
labiodental	f, v	1		
bilabial	b, p, m	1		
corner out	a	-0.2	e	0.5
TONGUE				
length	d	1.8	g	0.2
width	l	-0.2	j	1.1
thickness	d	-0.7	g	2
tip down	j	1.8	l	-0.8
tip up	n	0.9	k	0
raise	k	0.6	t	0
retraction	k	1	r	-0.6
back up	o	1.7	i	0
rotation up	r	0.9	k	0

For verification of the animation of our talking head, we have committed a subjective test. The subjective test was organised in China. There are 425 native speakers of Chinese Shaanxi Xi'an dialect assessed the 33 recordings: 10 sentences uttered by the author of theses recorded at the

University of Miskolc, 10 textured animations of these sentences by the talking head and 10 animations with a transparent face of the same sentences had to be rated by naive students involved in the subjective test. The scores could be selected from 1 to 5. The instruction for the subjects was: The three kind of videos of the same sentence are one group (uttered by the author, textured animations by the talking head and animations with a transparent face) respectively marked Sen1.1, Sen1.2, Sen1.3, ? Sen11.1, Sen11.2, Sen11.3. Please give three comparison subjective scores with grades from 1 to 5 and fill the corresponding score in the Excel document made by the author.。

There were three contradicting samples (one with each appearance): The video of a sentence was dubbed by the voice of the other sentence. Answers scoring the contradicting sentences above the 80% of overall average were excluded. The results of the subjective test can be seen in Table 3. 4.

Table 3. 4. Results of the subjective test.

Appearance	Original	Textured	Transparent
Score	4.33	4.27	4.30

The scores of different appearances are just slightly different. This verifies the correctness of our visual speech synthesis. I consider the higher scores of the transparent face animation to the visibility of the tongue.
The ability of untrained participants to perceive aspects of the speech signal has been explored for some visual representations of the vocal tract (e.g. talking heads), suggesting that these images can be interpreted intuitively to some degree. Speakers possess a natural capacity for lip-reading; analogous to this, there may be an intuitive ability to "tongue-read"[108].

3.7. Thesis

This thesis proposes a visual dynamic articulation lip and tongue model. This model establishes the rule set for Chinese Shaanxi Xi'an Dialect dynamic articulation, realizes the mapping of each phonetic features to its corresponding lip shape in a continuous flow of Chinese Shaanxi Xi'an dialect, and synthesizes the key lip shapes corresponding to each sound in the continuous stream. The method selects the appropriate transition shape from the candidate transition lip library and finally generates Chinese Shaanxi Xi'an Dialect articulation.

I developed an algorithm to automatically track spatial-temporal tongue movement contours from the ultrasound images. The visual information is classified by dominance and other features to define the uttered viseme and will form the basis of a dynamic viseme system of tongue motion for the Shaanxi Xi'an dialect. Similar classification of lip shape features is completed through analysis of the video recordings by camera. The interpolation between articulation features is refined with the analysis of the ultrasound image (position of the tongue) and video (shape of the lips) made during the continuous reading of a long text. The standard deviation of the feature examined well combines the essence of the analyses shown in this thesis. [109]

3.7.1. Novelty

The lip model proposed in this thesis combines phonetic linguistics with face-and-lip animation and obtains a dialect continuous stream that is simple and practical and has a certain sense of reality animation sequence.

We propose a method to create dynamic viseme model for visual speech synthesis that can deal with co-articulation problem and various pauses in continuous speech. This dynamic viseme system is used for creating the fundamentals of a talking head – an animated articulation model for the Chinese Shaanxi Xi'an dialect.

3.7.2. Measurements

For dynamic lip modeling, the measurements are divided into two processes. One is to capture and analyze the oral video of the real person in the continuous stream state to construct an image library for visualizing collaborative pronunciation modeling. The other is to synthesize lip animation in the continuous stream state according to the visual articulation rule based on the dynamic model.

For dynamic tongue modeling, a series of visual speech database of structures 'VCV' and 'CVC' covering all phonemes – 26 consonants and 40 vowels – of the Chinese Shaanxi Xi'an dialect in different tempo is recorded by the 'Micro' ultrasound system in order to trace the tongue contour feature point and classify the dominance grade. The standard deviation of the feature examined combines the essence of the analyses shown in thesis: the greater the deviation, the lower the dominance. Our animation process, elaborated for the Hungarian language, accomplishes the screening of the features according to dominance class. We propose a method to create dynamic tongue viseme model for Chinese Shaanxi Xi'an Dialect visual speech synthesis.

3.7.3. Limits of validity

The model proposed in this thesis can be applied to the synthesis of speech animation with constant speech rate, but it is not suitable for synthesizing speech animation with variable speech rate. Due to the influence of speech rate, the speaker's lip shape and tongue changes with the change of speech rate when the same viseme is emitted. This thesis does not take into account the impact of changes in speech rate. I will focus on the lip-moving model and tongue-moving model under different speech rate conditions.

3.7.4. Consequence

I studied both the timing and position properties of articulatory movements of the tongue in Chinese Shaanxi Xi'an dialect speech utterances spoken at different tempos by one native speaker of the dialect. She read randomized lists of VCV utterances containing the vowels /e/ or /a/ and CVC utterances containing the consonants /k/ or /t/ in all possible combinations of the

dialect's 26 consonants and 40 vowels. The 'Micro' ultrasound system recorded the utterances and the Assistant Advanced software formed JPG images and MP4 videos. I developed an algorithm to automatically track spatial-temporal tongue movement contours from the ultrasound images. The visual information is classified by dominance and other features to define the uttered viseme and will form the basis of a dynamic viseme system of tongue motion for the Chinese Shaanxi Xi'an dialect. Similar classification of lip shape features is through analysis of the video recordings. The interpolation between articulation features is being refined with the analysis of the ultrasound image (position of the tongue) and video (shape of the lips) made during the continuous reading of a long text. The standard deviation of the feature examined well combines the essence of the analyses shown in Figures 3.3-3.5: the greater the deviation, the lower the dominance. Our animation process, elaborated for the Hungarian language, accomplishes the screening of the features according to dominance class.

3.7.5. Related published paper

Zhao L, Czap L.: Visemes of Chinese Shaanxi Xi'an Dialect Talking Head[J]. Acta Polytechnica Hungarica, 2019, 16(5): 173-193.

Summary

The first thesis presents a method for the phonetic transcription of the Chinese Shaanxi Xi'an Dialect and the conversion of its basic phonemes into a computer readable phonetic alphabet. Transcription was based on the phonetic alphabet of the dialect, mapping the phonemes shared with Mandarin supplemented by several phonemes unique to Shaanxi Xi'an. I show the relationship of phonemes of the dialect with Mandarin and the X-SAMPA code derived for the Chinese Shaanxi Xi'an Dialect based on Hungarian X-SAMPA code, in addition to the correspondent regularities for whole syllable pronunciation. I presented a method for the phonetic transcription of the Chinese Shaanxi Xi'an Dialect. The purpose is to obtain the fundamental data needed to create a talking head for the Chinese Shaanxi Xi'an Dialect.

The second thesis applied the classification method for Mandarin static visemes to static viseme classification of Chinese Shaanxi Xi'an Dialect speech. I display the static lip viseme classification of Chinese Shaanxi Xi'an Dialect speech and analyze lip viseme parameters by processing images and videos of different lip visemes recorded by camera. I describe another experiment carried out to study both the timing and position properties of articulatory movements of the tongue in utterances recorded during dialect speech based on the algorithm I developed. Static viseme features both of lip and tongue also shown. The parameters derived from both of these two experiments are combined to create a dynamic viseme modeling system.

In the third thesis, I give a detailed description of co-articulation phenomena in speech stream and introduce the dominance concept, which is a rule to determine the dominance grade for lip and tongue visemes. The interpolation between articulation features is refined with the analysis of the ultrasound image (position of the tongue) and video (shape of the lips) made during the continuous reading of a long text. The standard deviation of the feature examined well combines the results of the analyses shown. Finally I give the results of the dominance grade for lip and tongue visemes. The purpose is to define the uttered viseme and create a dynamic viseme modeling system of lip and tongue for a Chinese Shaanxi Xi'an Dialect talking head.

List of Publications

International journals

Zhao L, Czap L.: Visemes of Chinese Shaanxi Xi’an Dialect Talking Head[J]. Acta Polytechnica Hungarica, 2019, 16(5): 173-193.

International conferences

Czap L, Zhao L.: Phonetic aspects of Chinese Shaanxi Xi'an dialect[C]. 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, 2017: 000051-000056.

Z Lu, Czap L.: Modelling the tongue movement of Chinese Shaanxi Xi'an dialect speech[C]. 2018 19th International Carpathian Control Conference (ICCC). IEEE, 2018: 98-103.

Hungarian publications

Zhao L., Czap L: A nyelvkontúr automatikus követése ultrahangos felvételeken (Automatic tongue contour tracking on ultrasound images, In Hungarian) [J]. Beszédkutatás 2019 Vol. 27 : 1 pp. 331-343

Chinese publications

Zhao Lu, Jia Xian. Low power tipping-hopper flowmeter based on MSP430F5418[J].Yibiao Jishu.(11):37-40.2017. (In Chinese, 赵璐, 贾先. 基于 MSP430F5418 的低功耗翻斗流量计[J]. 仪表技术(11):37-40.2017.)

Zhao Lu, Jia Xian, Xiong Sen. Retrofit of Electro-Mechanical Heating System Based on LTUY900 Full Hydraulic Paver [J]. Electronics World, No.543 (9): 117-119.2018.(In Chinese, 赵璐, 贾先, 熊森. 基于 LTUY900 型全液压摊铺机电加热系统的改造[J]. 电子世界, No.543(9):117-119.2018.)

References

[1] Lyu R. Y: A bi-lingual Mandarin/Taiwanese (Min-nan), Large Vocabulary, Continuous speech recognition system based on the Tong-yong phonetic alphabet (TYPA)[C] Sixth International Conference on Spoken Language Processing. 2000.

[2] Liu B, Yang H, Gan Z: Grapheme-to-phoneme conversion of Tibetan with SAMPA[J]. Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications), 2011, 47(35): 117-121. (In Chinese)

[3] Guo W, Yang H, Song J: Research on Text Analysis for Dialect Speech Synthesis[J]. Computer Engineering, 2015. (In Chinese)

[4] Wurm S A, Li R, Baumann T: Language Atlas of China[M]. Australian Academy of the Humanities; Longman Group (Far East), 1987.

[5] Czap L, Mátyás J: Hungarian talking head[C]. Proceedings of Forum Acusticum 4th European Congress on Acoustics. Budapest, Hungary, 2005. pp.

[6] Czap L, Mátyás J: Virtual speaker[J]. Infocommunications Journal Selected Papers, 2005, Vol. 60, 6, pp. 2-5

[7] Wang A, Bao H, Chen J: Primary research on the viseme system in standard Chinese[J]. Proceedings of the International Symposium of Chinese spoken language Processing, 2000.

[8] Bailly G, Badin P: Seeing tongue movements from outside[C].Seventh International Conference on Spoken Language Processing. 2002.

[9] Robert-Ribes J, Schwartz J L, Lallouache T: Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise[J]. The Journal of the Acoustical Society of America, 1998, 103(6): 3677-3689.

[10] Erber N P: Auditory-visual perception of speech[J]. Journal of Speech and Hearing Disorders, 1975, 40(4): 481-492.

[11] Czap L, Pinter J M: Multimodality in a Speech Aid System[J]. Journal on Human Machine Interaction, 2014, 1: 64-71.

[12] Werda S, Mahdi W, Hamadou A B: Lip localization and viseme classification for visual speech recognition[J]. arXiv preprint arXiv:1301.4558, 2013.

[13] Massaro D W, Beskow J, Cohen M M: Picture my voice: Audio to visual speech synthesis using artificial neural networks[C]. AVSP'99-International Conference on Auditory-Visual Speech Processing. 1999.

[14] Czap L: On the Audiovisual Asynchrony of Speech[J]. Proceedings of Auditory-Visual Speech Processing (AVSP) 2011. Volterra, Italy, International Speech Communication Association (ISCA), pp. 137-140.

[15] Železný M, Krňoul Z, Císař P: Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis[J]. Signal Processing, 2006, 86(12): 3657-3673.

[16] Pintér J M, Czap L: Improving Performance of Talking Heads by Expressing Emotions[C]. 3nd CogInfoCom Confrence. Košice, Slovakia, IEEE, pp. 523-526.

[17] Zorić G, Pandžić I S: Real-time language independent lip synchronization method using a genetic algorithm[J]. Signal Processing, 2006, 86(12): 3644-3656.

[18] Liu WenTao, Yin BaoCai, Jia XiBin, Kong DeHui: A Realistic Chinese Talking Face[C]. Proceedings of the 1st Indian International Conference on Artificial Intelligence, IICAI 2003, Hyderabad, India, December 18-20, 2003

[19] Segaran, Kogilathah, Ahmad Zamzuri Mohamad Ali and Tan Wee Hoe: Talking Head Animation as Pedagogical Agent in Language Learning: A Review on Instructional Strategy and Media[J]. 2014.

[20] Massaro D W, Cohen M M: Visible speech and its potential value for speech training for hearing-impaired perceivers[C]. STiLL-Speech Technology in Language Learning. 1998.

[21] Massaro D W, Light J: Read my tongue movements: bimodal learning to perceive and produce non-native speech [C]. Eighth European Conference on Speech Communication and Technology. 2003. CD-ROM 4 pp.

[22] Badin P, Bailly G, Reveret L: Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images[J]. Journal of Phonetics, 2002, 30(3): 533-553.

[23] Badin P, Elisei F, Bailly G: An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data[J]. Articulated Motion and Deformable Objects, 2008: 132-143.

[24] Fagel S, Madany K: A 3-D virtual head as a tool for speech therapy for children[C]. Ninth Annual Conference of the International Speech Communication Association. 2008.

[25] Wik P, Engwall O: Can visualization of internal articulators support speech perception?[C]. INTERSPEECH. 2008: 2627-2630.

[26] Beskow J, Engwall O, Granström B: Visualization of speech and audio for hearing impaired persons[J]. Technology and Disability, 2008, 20(2): 97-107.

[27] Czap László , Illés Béla , Varga Attila: Concept of a Speech Assistant System[C]. 4th Word Congress on Software Engineering WCSE 2013. Hong Kong, China, 2013.12.03 -2013.12.04. Hong Kong: IEEE Com-puter Society, pp. 207-211.

[28] Czap László: Speech Assistant System. INTERSPEECH 2014[C]. 15th Annual Conference of the Interna-tional Speech Communication Association. Singapore, 2014.09.14 -2014.09.18. Singapore: International Speech Communication Association (ISCA), pp. 1486-1487.

[29] Baranyi P, Csapo A, Sallai G: Cognitive Infocommunications (CogInfoCom)[M]. Springer, 2015.

[30] Baranyi P, Csapó Á: Definition and synergies of cognitive infocommunications[J]. Acta Polytechnica Hungarica, 2012, 9(1): 67-83.

[31] Vinciarelli A, Pantic M, Heylen D, et al.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing[J]. IEEE Transactions on Affective Computing, 2011, 3(1): 69-87.

[32] Sallai G: The cradle of cognitive infocommunications[J]. Acta Polytechnica Hungarica, 2012, 9(1): 171-181.

[33] Massaro D W, Cohen M M: Evaluation and integration of visual and auditory information in speech perception[J]. Journal of Experimental Psychology: Human Perception and Performance, 1983, 9(5): 753.

[34] Kovács S, Tóth Á, Czap L: Fuzzy model based user adaptive framework for consonant articulation and pronunciation therapy in Hungarian hearing-impaired education[C]. 2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom). IEEE, 2014: 361-366.

[35] https://en.wikipedia.org/wiki/Pinyin

[36] Zein P.: Mandarin Chinese Phonetics. http://www.zein.se/patrick/chinen8p.html, accessed: 11.07.2017

[37] Zhou Youguang: Basic knowledge of Hanyu Pinyin Schedule [M]. Language Publishing House. 1995. (In Chinese)

[38] Sun Lixin: Xi'an dialect research [M]. Xi'an publishing house, 2007. (In Chinese)

[39] Kang Jizhen: An Experimental Study of Phonetics in Xi'an Dialect[C]. Northwest University. 2015. (In Chinese)

[40] Guo Weitong: Analysis of Acoustic Features and Modeling of Prosody in Xi'an Dialect. [C]. Northwest Normal University. 2009. (In Chinese)

[41] Yuan Jiahua: Outline of Chinese Dialects [M]. Text Reform Press, 1983. (In Chinese)

[42] Chinese Dialect Vocabularies [M]. Text Reform Press, 1989. (In Chinese)

[43] Shi Feng: Study on the five degree value method [J]. Journal of Tianjin Normal University (SOCIAL SCIENCE EDITION), 1990 (3): 67-72. (In Chinese)

[44] Li A: Chinese prosody and prosodic labeling of spontaneous speech[C]. Speech Prosody 2002, International Conference. 2002.

[45] Wu Z: From Traditional Chinese Phonology to Modern Speech Processing--Realization of Tone and Intonation in Standard Chinese [J]. Language Teaching & Linguistic Studies, 2002.

[46] Yuan Jiahua: Outline of Chinese Dialects [M]. Text Reform Press, 1983. (In Chinese)

[47] Chinese dialect Vocabularies [M]. Text Reform Press, 1989. (In Chinese)

[48] Yuan Jiahua: Outline of Chinese Dialects [M]. Text Reform Press, 1983. (In Chinese)

[49] Mao-Peng M A: Acoustic Study of the Tones of Xi'an Dialect [J]. Journal of Yanan University, 2005. (In Chinese)

[50] Guo Jinfu: A survey on the Chinese tone and intonation[M]. Beijing Language Institute. (In Chinese)

[51] Zhu W, Zhang J: Manual segmentation & labeling in Chinese speech database[C]. The first China-Japan Workshop on Spoken Language Processing (CJSLP'97). 1997.

[52] Chen, X., Li, A., Sun, G., Hua, W., & Yu, Z.: An application of SAMPA-C for standard Chinese[C].Sixth International Conference on Spoken Language Processing. 2000.

[53] Zhang Jialu: SAMPA_SC for standard Chinese (Putonghua)[J]. Acta Acustica, 2009, 34:82-86.(In Chinese)

[54] SAMPA: Computer Readable Phonetic Alphabet. http://www.phon.ucl.ac.uk/home/sampa/home.htm; accessed: 11.07.2017

[55] Wells J C: Computer-coding the IPA: a proposed extension of SAMPA[J]. Revised draft, 1995, 4(28): 1995.

[56] Zhao Ling. The corresponding rules of the initial consonants of Mandarin and Shaanxi Dialect [J]. Journal of Baoji University of Arts and Sciences (Social Science Edition) (1): 127-130. (In Chinese)

[57] Sun Lixin: Classification of the Xi'an Dialect [J]. Dialect,Vol. 2: 106-124. 1997. (In Chinese)

[58] Lu Tuanhua: A Comparison between the Phonetic Features of Xi'an Dialect and the Pronunciation of Mandarin[J] ,Kaoshi Zhoukan, No.9,2010. (In Chinese)

[59] Yang Jinfeng: Corresponding speech sound in west Shannxi Dialect and Mandarin[J]. Journal of Xianyang Teachers'College.Vol.18, No.5, 2003. (In Chinese)

[60] Wang Yuding: Research on the types of phonetic changes in Xi'an Dialect. [J].Journal social of Yanan university science edition. Vol.17, No.2, 1995. (In Chinese)

[61] Czap, László, and Lu Zhao: Phonetic aspects of Chinese Shaanxi Xi'an dialect [C]. 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, 2017.

[62] Wang A, Bao H, Chen J: Primary research on the viseme system in standard Chinese[C]. Proc. Internat. Symp. on Chinese Spoken Language Processing, ISCSLP, Beijing, China, October. 2000: 13-15.

[63] Bailly G, Badin P: Seeing tongue movements from outside[C]. Seventh International Conference on Spoken Language Processing. 2002.

[64] Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P.: Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise[J]. The Journal of the Acoustical Society of America, 1998, 103(6): 3677-3689.

[65] Erber N P.: Auditory-visual perception of speech[J]. Journal of speech and hearing disorders, 1975, 40(4): 481-492.

[66] Werda, Salah, Walid Mahdi, and Abdelmajid Ben Hamadou: Lip localization and viseme classification for visual speech recognition[J]. arXiv preprint arXiv:1301.4558 (2013).

[67] Massaro, D. W., Beskow, J., Cohen, M. M., Fry, C. L., & Rodgriguez, T: Picture my voice: Audio to visual speech synthesis using artificial neural networks[J]. AVSP'99-International Conference on Auditory-Visual Speech Processing. 1999.

[68] Železný, M., Krňoul, Z., Císař, P., & Matoušek, J: Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis[J]. Signal Processing 86.12 (2006): 3657-3673.

[69] Zorić, Goranka, and Igor S. Pandžić: Real-time language independent lip synchronization method using a genetic algorithm[J]. Signal processing 86.12 (2006): 3644-3656.

[70] https://en.wikipedia.org/wiki/Viseme

[71] Bothe H H, Wieden E A.: A neurofuzzy approach for modeling lips movements[C]. Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference. IEEE, 1994: 234-237.

[72] Le Goff B, Benoît C.: A text-to-audiovisual-speech synthesizer for french[C]. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. IEEE, 1996, 4: 2163-2166.

[73] Ezzat T, Poggio T. Miketalk: A talking facial display based on morphing visemes[C]. Proceedings Computer Animation'98 (Cat. No. 98EX169). IEEE, 1998: 96-102.

[74] Lande C, Francini G. An MPEG-4 facial animation system driven by synthetic speech[C]. mmm. IEEE, 1998: 203.

[75] Qi Jie: Text-driven lip-motion synthesis system[J]. Computer Engineering and Design, Vol (1): 31-34. 1998. (In Chinese)

[76] Zhong Xiao, Zhou Changle, Yu Ruizhen, et al.: A Method of Oral Shape Recognition for Chinese Speech Recognition[J]. Journal of Software, 1999, 10(2): 205-209. (In Chinese)

[77] Czap L, Pintér J M, Baksa-Varga E.: Features and results of a speech improvement experiment on hard of hearing children[J]. Speech Communication, 2019, 106: 7-20.

[78] Adams S G, Weismer G, Kent R D.: Speaking rate and speech movement velocity profiles[J]. Journal of Speech, Language, and Hearing Research, 1993, 36(1): 41-54.

[79] Napadow, V. J., Chen, Q., Wedeen, V. J., & Gilbert, R. J.: Intramural mechanics of the human tongue in association with physiological deformations[J]. Journal of biomechanics, 1999, 32(1): 1-12.

[80] Dromey, C., Nissen, S., Nohr, P., & Fletcher, S. G.: Measuring tongue movements during speech: Adaptation of a magnetic jaw-tracking system[J]. Speech Communication, 2006, 48(5): 463-473.

[81] Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., Jackson, M. T.: Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements[J]. The Journal of the Acoustical Society of America, 1992, 92(6): 3078-3096.

[82] Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B.: Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract[J]. Brain and Language, 1987, 31(1): 26-35.

[83] Dromey, C., Nissen, S., Nohr, P., & Fletcher, S. G.: Measuring tongue movements during speech: Adaptation of a magnetic jaw-tracking system[J]. Speech Communication, 2006, 48(5): 463-473.

[84] Lundberg A J, Stone M.: Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data[J]. The Journal of the Acoustical Society of America, 1999, 106(5): 2858-2867.

[85] Ostry D J, Munhall K G. Control of rate and duration of speech movements[J]. The Journal of the Acoustical Society of America, 1985, 77(2): 640-648.

[86] Stone M.: A guide to analysing tongue motion from ultrasound images[J]. Clinical linguistics & phonetics, 2005, 19(6-7): 455-501.

[87] Akgul Y S, Kambhamettu C, Stone M.: Automatic extraction and tracking of the tongue contours[J]. IEEE Transactions on Medical Imaging, 1999, 18(10): 1035-1045.

[88] Scobbie J M, Wrench A A, van der Linden M: Head-Probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement[C]. Proceedings of the 8th International Seminar on Speech Production. 2008: 373-376.

[89] Akgul Y S, Kambhamettu C, Stone M: Automatic extraction and tracking of the tongue contours[J]. IEEE Transactions on Medical Imaging, 1999, 18(10): 1035-1045.

[90] Z Lu, Czap L.: Modelling the tongue movement of Chinese Shaanxi Xi'an dialect speech[C]. 2018 19th International Carpathian Control Conference (ICCC). IEEE, 2018: 98-103.

[91] Waters K, Levergood T M.: DECface: An automatic lip-synchronization algorithm for synthetic faces[M]. Digital Equipment Corporation, Cambridge Research Laborartory, 1993.

[92] Kshirsagar S, Magnenat-Thalmann N.: Lip synchronization using linear predictive analysis[C]. 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532). IEEE, 2000, 2: 1077-1080.

[93] Morishima S.: Facial animation synthesis for human-machine communication system[J]. Human-Computer Interaction: Software and Hardware Interfaces, 1993: 1085-1090.

[94] Cohen M M, Massaro D W.: Modeling coarticulation in synthetic visual speech[M]. Models and techniques in Computer Animation. Springer, Tokyo, 1993: 139-156.

[95] Bregler C, Covell M, Slaney M.: Video Rewrite: driving visual speech with audio[C]. Siggraph. 1997, 97: 353-360.

[96] Pelachaud C, Badler N I, Steedman M.: Generating facial expressions for speech[J]. Cognitive Science, 1996, 20(1): 1-46.

[97] Zhiming W, Lianhong C, Haizhou A.: Text-to-visual speech in Chinese based on data-driven approach[J]. 2003.

[98] Liu N, Fang N, Kamata S.: 3D reconstruction from a single image for a Chinese talking face[C]. TENCON 2010-2010 IEEE Region 10 Conference. IEEE, 2010: 1613-1616.

[99] Zhou Wei. Realistic 3D face animation research of Chinese speech synchronization [D]. University of Science and Technology of China. 2008. (In Chinese)

[100] Sztahó D, Kiss G, Czap L, Vicsi K: A computer-assisted prosody pronunciation teaching system[C]. WOCCI. 2014: 45-49.

[101] Wu, Z., Zhang, S., Cai, L., & Meng, H. M.: Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar[C]. Ninth International Conference on Spoken Language Processing. 2006.

[102] Zhao H, Tang C.: Visual speech synthesis based on Chinese dynamic visemes[C]. 2008 International Conference on Information and Automation. IEEE, 2008: 139-143.

[103] D Sztahó, G Kiss, L Czap, K Vicsi: A Computer-Assisted Prosody Pronunciation Teaching System[C]. WOCCI 2014 Satellite Workshop of Interspeech Singapore 2014. Singapore, 2014.09.19.

[104] Cosi P, Fusaro A, Tisato G.: LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model[C]. Eighth European Conference on Speech Communication and Technology. 2003.

[105] Xu K., Csapó T. G., Roussel P., Denby B.: A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization[J]. Journal of the Acoustical Society of America. 2016 Vol. 139, 5 pp. 154-160.

[106] Taylor S L, Mahler M, Theobald B J: Dynamic units of visual speech[C]. Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation. Eurographics Association, 2012: 275-284.

[107] Aghaahmadi M, Dehshibi M M, Bastanfard A, Fazlali M: Clustering Persian viseme using phoneme subspace for developing visual speech application[J]. Multimedia Tools and Applications, 2013, 65(3): 521-541.

[108] Joanne Cleland,Caitlin Mccron&James M. Scobbie:Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds. Clinical Linguistics & Phonetics, Volume 27, 2013 - Issue 4, pp 299-311.

[109] Zhao L, Czap L.: Visemes of Chinese Shaanxi Xi'an Dialect Talking Head[J]. Acta Polytechnica Hungarica, 2019, 16(5): 173-193.