Visemes of Chinese Shaanxi Xi’an Dialect Talking Head

Lu Zhao[1], László Czap[*2]

Institute of Automation and Infocommunication, University of Miskolc
H-3515 Miskolc-Egyetemváros, Hungary
e-mail: [1] qgezhao@uni-miskolc.hu, [2] czap@uni-miskolc.hu

Abstract: Animated 3D articulation models – called talking heads – can be utilized, for instance, in speech assistant systems for children who are hard-of-hearing or when teaching learners of a second language. In this study, the objective is to identify articulation features and a dynamic system for visual representation of speech sounds for a Shaanxi Xi’an dialect talking head. In the first phase of the study, a phonetic alphabet of the dialect (northwest China) is formed following the official Romanization system used for Mandarin (Standard Chinese). After relating the phonemes of the dialect to those of Mandarin, we introduce the SAMPA code developed for the dialect, in addition to the correspondent regularities for whole syllable pronunciation. Secondly, we display the classification of static visemes (phonemes represented in visual form) for the dialect and describe an experiment carried out to articulatory movements of the tongue (features of timing and position) in dialect speech utterances recorded at different tempos. Finally, we discuss the results of an analysis of the images based on spatial-temporal tracking of the tongue movement contour. For definition of each uttered viseme the visual information obtained is classified and then used to create the dynamic viseme system of the tongue for a talking head using the Shaanxi Xi’an dialect of Chinese.

Keywords: Shaanxi Xi’an dialect; talking head; tongue contour tracking; dynamic viseme system; speech assistant system

1 Introduction

In this paper we identify the fundamental aspects needed for forming a talking head for the Shaanxi Xi’an dialect of Chinese [1]. The structure of the talking head system is illustrated in Figure 1. Using the X-SAMPA code created here, visemes are created for the consonants and vowels of the dialect. The viseme library also contains a dominance model (see Section 4.2) for the talking head [2]. Viseme classification is aided by the X-SAMPA code of consonants (C) and vowels (V) and identification of the regularities of their usage in whole-syllable pronunciation. We categorized the static visemes of the Shaanxi Xi’an dialect using the classification method for static visemes of Mandarin (Standard Chinese). Then an experiment was conducted focusing on the timing and position features of articulatory movements of the tongue in VCV and CVC utterances in the Chinese Shaanxi Xi'an dialect.

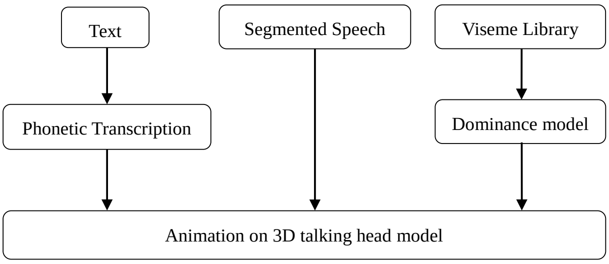


Figure 1
Structure of Shaanxi Xi'an dialect talking head system

Mandarin is a common focus of basic linguistic research in the fields of speech synthesis and speech recognition technology. However, other minority languages and a variety of Chinese dialects are present in modern-day multi-ethnic China. There are various studies relevant to the current topic. For instance, one study [3] has looked at phonetic conversion from Mandarin to the Min dialect of Taiwanese, along with mixed speech synthesis in Chinese with English. Another investigation of speech synthesis involving dialects focused on Tibetan, using a computer readable SAMPA scheme for conversion of text [4]. The Lanzhou, Liaocheng, Shenyang, and Tianjin dialects of Chinese have also been represented in speech synthesis [5].

The talking head being developed will form the foundation of a system to support deaf and hard-of-hearing children in learning to produce speech. It is possible to make the digital face transparent so that tongue placement and movement are clearly displayed – an advantage over a human speaker.

Xi'an, located in northwest China, was a capital during 13 dynasties of ancient China, and remains important today. Its dialect, Shaanxi Xi'an (also called Qin), has been spoken for over 3,000 years and has 8 million speakers today. It is the representative dialect of a large and influential region [6], making it well worth researching. Shaanxi Xi'an displays differences from Mandarin in its vocabulary, grammar, and most particularly in its articulation.

In order to create visual speech synthesis of the dialect, a transcription system is needed to label its phonetic information [7]. Based on this, it is possible to convert to SAMPA – Speech Assessment Methods Phonetic Alphabet. This is a machine-readable phonetic alphabet developed within the ESPRIT project that was first used with languages in the European Community, but has expanded to a variety of languages worldwide [8] [9]. This prompted the development of X-SAMPA (Extended-SAMPA), which covers all of the International Phonetic Alphabet (IPA) characters and remaps them into 7-bit ASCII, meaning that computer-readable phonetic transcriptions can be generated for any language [10]11]. X-SAMPA analysis of dialect phonemes and comparison with Hungarian and Pinyin ones enables us to derive certain viseme features from these existing systems, meaning that the main tasks are to distinguish unique features and identify regularities in articulation compared to those languages.

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 [12]. 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 that speech [13]. The audiovisual mode is more effective in transmitting articulatory features than any form of unimodal communication [14]. It has been proven 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 [15]. 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 [16]. Visual speech is studied and utilized in the fields of speech recognition [17], speech processing [18] [19], audio-visual speech synthesis [20], virtual talking head animation [21] and lip or tongue synchronization [22].

During the visual perception of speech, it is not the sight of the movement of lips and the face alone that matters; it has been found that the motion of the tongue, even though it is partially hidden, also conveys articulatory information that lip reading alone cannot access [23]. Development of infocommunication systems encourages speech researchers to deal with the speech of hard of hearing people and to study its physiological and acoustic characteristics. Computers can reveal features unseen before. In view of this, there are some pioneer Hungarian applications, such as the "Magic Box" (Varázsdoboz) package, developed by a team led by Klára Vicsi, which offers a tool for correcting pronunciation using spectrograms [24].

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 pronunciation training (CAPT) by Massaro and Cohen, who employed it as a tool in speech therapy and second language learning [25]. 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 [26]. 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 [27] that was later built into an audiovisual talking head that was able to display the normally hidden articulators (tongue, velum) during articulation [28]. 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 [29]. Wik and Engwall described how intra-oral articulations displayed in animation were able to contribute to the perception of speech [30]. 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. [31]. An audio-visual representation of speech processes is also given by talking head developed at the University of Miskolc in Hungary, which is intended primarily to act as an aid in teaching speech to hard-of-hearing children [32] [33].

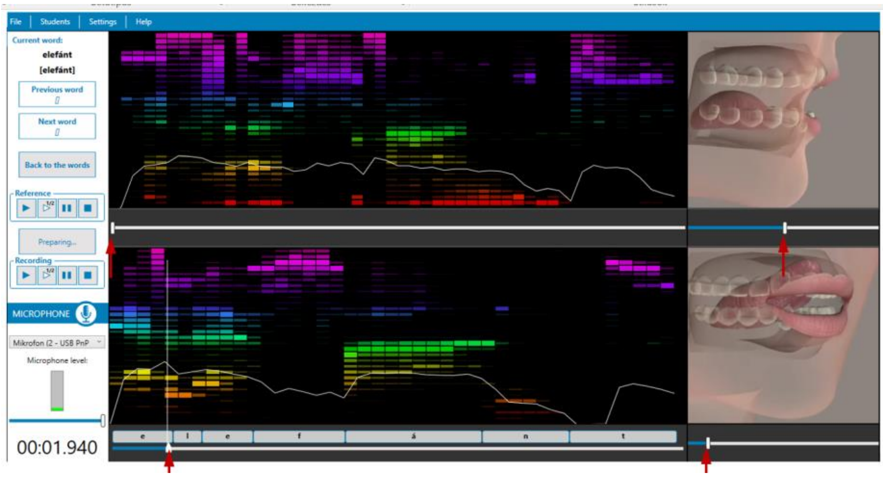


Figure 2

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

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° 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.

In the SA the bar chart and the talking head visually represent the speech signal and articulation, realizing sensor bridging between the modalities [34] [35] [36]. “The sensory information is transformed to an appropriate sensory modality in a way that the user can process it effectively” [37]. 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 [38].

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 assist those practicing on their own [39]. 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.

Tongue movement measurement technologies have progressed in a number of stages. An x-ray microbeam system to investigate the effect of rate of speech on tongue-tip and lower-lip velocity profiles when uttering stop consonants was introduced by Adams et al [40]. A non-invasive NMR tagging technique representing tissue as discrete deforming elements was used by Napadow et al. for quantifying the degree of local tongue deformation [41]. Systems involving electromagnetic tracking utilize transmitters placed around the head and sensor coils positioned in the mid-sagittal plane and fixed to a variety of points on the jaw, lips and tongue [42]-[45]. Tongue positions can be revealed by ultrasound imaging, providing 2D images of the tongue surface contour [46]. Tongue dorsum movements were monitored during C-V sequences with varying speech rates with a computerized pulsed-ultrasound system [47]. Stone introduces methods for extracting, displaying and analyzing ultrasound image contours [48]. Ultrasound technology, despite some disadvantages, is a practical way of obtaining an image sequence for tongue motion. This non-invasive technology offers real-time capture rates, is relatively affordable, and can be easily incorporated into experimental setups. Other available methods such as slow motion recording, high cost (MRI), or radiation exposure (X-ray) have major drawbacks [49].

The following section deals with identifying different articulation of consonants and vowels between Mandarin and the Shaanxi Xi’an dialect and extension of the Pinyin Scheme to form a phonetic alphabet. The third section focuses on analysis of the CV phonetic features of the Shaanxi dialect and its phoneme changes compared to Mandarin. In the fourth section we report on an experiment studying articulatory movements of the tongue in terms of timing and position when a Chinese Shaanxi Xi'an dialect speaker is making VCV and CVC utterances of at varied tempos. Methods for modeling the tongue movements and creating the dynamic viseme system are detailed. In addition, we provide an example showing the result of tracing the tongue contour obtained by ultrasound technology for the phonemes of the dialect based on the algorithm we developed in MATLAB. Finally, we summarize our progress towards achieving a talking head for Chinese Shaanxi Xi'an dialect.

2 Phonetic Alphabet of Shaanxi Xi'an Dialect

2.1 Pinyin Scheme for Mandarin Chinese

The Pinyin scheme, the official Romanization system for Mandarin Chinese, contains basic phonemes (56), consonants (23) and simple vowels (36). This leads to 413 potential CV combinations plus special cases. When the four tones of Mandarin are considered, there are about 1,600 unique syllables. Pinyin is illustrated in Table 1 [50] [51].

2.2 Shaanxi Xi'an Dialect and Its Phonetic Alphabet

The Shaanxi Xi'an dialect consists of 26 consonants and 40 simple vowels. Its phonemes represented in IPA can be found in [52] [53] [54]. Its five unique consonants are presented in Table 2: 'pf' and 'pfʰ' are both labiodental plosive fricative consonants, 'v' is a voiced labiodental fricative consonant, 'ŋ' is a velar nasal initial, and 'ɳ' is a retroflex nasal consonant [55] [56].

For establishing the relationship between phonemes and SAMPA code, first we need to identify the articulation of Shaanxi Xi'an dialect in IPA, then transcribe it into X-SAMPA [57] [58] [59]. Results are given in Tables 2 and 3.

Table 3 shows both the simple and compound vowels of the Shaanxi Xi'an dialect, which has 13 unique vowels compared to Mandarin.

X-SAMPA code analysis has shown that 5 vowels and 13 consonants are similar to Hungarian phonemes. The dialect talking head visemes of these phonemes can be easily derived from their Hungarian counterparts.

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		
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

Table 2

X-SAMPA mapping for consonants of Shaanxi Xi'an dialect

Character	IPA	X-SAMPA
追	Pf	pf
吹	pfʰ	pv
味	V	v
爱	ŋ	N
女	ɳ	n`

Table 3

X-SAMPA mapping for vowels of Shaanxi Xi'an dialect

Character	IPA	X-SAMPA	Character	IPA	X-SAMPA
哀	æ	{	恩	ẽ	e~
岩	iæ	i{	因	iẽ	ie~
歪	uæ	u{	温	uẽ	ue~
安	æ̃	{~	晕	yẽ	ye~
烟	iæ̃	i{~	核	ɯ	M
弯	uæ̃	u{~	药	yo	Yo
冤	yæ̃	y{~			

[1] Note on pronunciation: The letters 'y' and 'w' can mark a new syllable: the syllable 'wu' is pronounced as the Pinyin 'u' , 'yi' as the Pinyin 'i' and 'yu' as the Pinyin 'ü'

3 Phonemic Differences between Mandarin and the 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 [60].

3.1 The Correspondence of some Consonants between Shaanxi Xi’an Dialect and Mandarin [61]

Table 4 shows the different consonants used in Shaanxi Xi’an dialect and in Mandarin when articulating the same Chinese character.

Table 4

Correspondence of some consonants in Mandarin and dialect

Mandarin	Dialect	Examples (Mandarin/ Shaanxi Xi’an dialect)
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;
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/ziong; 精 jīng/zing
q	c	全 quān/cuan; 群 qún/cun；晴 qíng/cing
x	s	选 xuān/suan; 讯 xùn/sun; 削 xūe/suo; 行 xíng/sing

The syllables beginning with the consonants ‘n’ and ‘l’ basically have the same articulation as Mandarin, but in situations such as ‘农(nóng)’, ‘n’ is articulated as ‘l’. The consonants ‘ch’, ‘t’, ‘d’, ‘k’ in Mandarin also have corresponding consonants in dialect. It can also be seen that when a syllable starts with the Mandarin consonants ‘j’, ‘q’, and ‘x’, in dialect these are articulated as ‘g’, ‘k’, and ‘h’, respectively.

A considerable number of non-aspirated consonants (‘b’, ‘d’, ‘g’, ‘j’, ‘z’) in Mandarin are always replaced by the aspirated initials (‘p’, ‘t’, ‘k’, ‘q’, ‘c’, respectively) in Shaanxi Xi’an dialect. We present a series of examples in Table 5 to demonstrate this phenomenon.

Table 5

Correspondence 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 the phonemes ‘zh’, ‘ch’, or ‘sh’ are used at the beginning of the syllables they will be articulated as either ‘zh’, ‘ch’, ‘sh’ or as ‘z’, ‘c’, ‘s’ depending on the following phonemes. When ‘zh’, ‘ch’, ‘sh’ are followed by finials such as ‘a’, ‘ai’, ‘an’, ‘en’ they are articulated ‘z’, ‘c’, ‘s’; otherwise, they are pronounced ‘zh’, ‘ch’, ‘sh’. Some examples are listed to express this rule in Table 6.

Table 6

Correspondence among the phonemes ‘zh’, ‘ch’, ‘sh ’and ‘z’, ‘c’, ‘s’

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 Xi’an dialect speakers articulate differently. ［pf］, ［pfʰ］ and ［f］ are fricatives, voiceless and unaspirated, ［v］ is fricative, voiced (note that the use of square brackets indicates IPA). Table 7 gives some examples to explain this articulation of phonemes.

Table 7

Complex reading of syllables ‘zhu’, ‘chu’, ‘shu’, ‘ru’

Character	猪	追	入	吹
Mandarin	zhǖ	Zhuī	rù	chuī
Dialect	［pfu］	［pfui］	[vu]	［pfʰei］

3.2 Vowel Features of Dialect Compared with Mandarin [62]

When the consonants ‘d’, ‘t’, ‘n’, ‘l’, ‘z’, ‘c’, ‘s’ and ‘zh’, ‘ch’, ‘sh’ appear in the front of the vowel ‘u’, ‘u’ will be changed to ‘ou’. This phenomenon is extremely common, and is widely perceived as an accent feature of the Shaanxi people. Table 8 presents examples of this and other vowel changes to show the corresponding phonemes between the two languages.

Table 8

Correspondance between 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 Xi’an dialect. Table 9 presents examples.

Table 9

Special vowel changes

Character	三	端	捐	电
Mandarin	sān	Duān	jüān	diàn
Dialect	sain	Duain	jüai	die

4 Modeling the Tongue Movement of Chinese Shaanxi Xi'an Dialect Speech

This section describes the process used to gather data and model motion of the tongue for use in the talking head. The process begins with a static viseme classification of the dialect based on the method used to classify Mandarin static visemes [12] [63]. Then we introduce the ultrasound system used to obtain the speech materials (VCV and CVC sequences at different tempos). The visual information thus obtained is used to define the uttered viseme (the visual representation of a phoneme) for the dynamic viseme system for the tongue. This dynamic viseme system forms the fundamentals of a talking head – in this case, an animated articulation model for Shaanxi Xi’an dialect speech [64] [65].

4.1 Static Viseme Classification

We established a set of static visemes reflecting characteristics of the Shaanxi Xi’an dialect and their phonetic composition. When the dialect and Mandarin share a phoneme, we refer to the classification of Mandarin static visemes; when not, we provide dialect-specific phonemes. The viseme system in Standard Chinese is carried out by means of statistical analysis [12]. The viseme classification of the 26 consonants of Chinese Shaanxi Xi’an dialect speech is shown in Table 10.

Table 10[2]

Static consonant visemes classification for Shaanxi Xi’an 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		

[2] ‘v’, ‘pf’, ‘pf^h’, ‘ɳ’, ‘ŋ’ are phonemes illustrated by IPA and others are a Romanized expression of phonemes.

The 40 vowels of the dialect [66] are classified into Static vowel viseme groups in Table 11. Fifteen basic static vowel visemes are classified; compound vowels are a combination of visemes for single vowels, as illustrated in Table 11.

Table 11[3]

Static vowel visemes classification and examples for Shaanxi Xi’an dialect

Simple Vowel	a, ang	啊，昂	er	儿
	æ, æ̃	哀，安	i	衣
	ao	奥	u	乌
	o	喔	ü	迂
	ou	欧	-i(-i front)	是
	e,eng	鹅，鞥	-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			

4.2 Dominance Classification Concept

While some parameters reach their target values during pronunciation, others do not, especially during fast speech. Grouping was performed according to the dominance of each feature determining tongue position and lip share and the articulation features of each speech sound were placed in a dominance class [66]. This differs from the standard approach, which classifies only the dominance of the phonemes. Four dominant grades emerge from the parametric model features:

- stable — co-articulation has no effect (e.g. tongue position of alveolar plosives, lip shapes of bilabials),
- dominant — co-articulation has only a slight effect (e.g. lip shapes of vowels),
- flexible —neighboring sounds affect the feature (e.g. tongue positions of vowels),
- uncertain — the neighborhood defines the feature (e.g. tongue position of bilabials, lip shapes of ‘h’ and ‘r’).

[3] ‘æ’, ‘iæ’, ‘uæ’, ‘æ̃’, ‘iæ̃’, ‘uæ̃’, ‘yæ̃’, ‘ẽ’, ‘iẽ’, ‘uẽ’, ‘yẽ’, ‘ɯ’, ‘yo’ are phonemes illustrate by IPA and others are a Romanized expression of phonemes.

4.3 Tongue Movement Contour Tracking

In this study we record a small-scale visual speech database using combinations of the consonants and vowels of 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 Chinese Shaanxi Xi’an dialect determined through dynamic analysis.

4.3.1 Subjects and Speech Material

The subject was one adult female – the first author of this paper – who is a native speaker of Chinese Shaanxi Xi’an dialect.

Two structures were investigated, VCV and CVC (‘V’ indicates vowel while ‘C’ indicates consonant), covering all phonemes involved in our experiment (the 26 consonants and 40 vowels of Shaanxi Xi’an dialect). In the VCV structure, ‘e’ and ‘a’ (eCe and aCa) 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, ‘k’ and ‘t’ are the two phonemes used (kVk and tVt) 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 ‘a’ and ‘k’ and front position when articulating ‘e’ and ‘t’.

4.3.2 Tongue Movement Recording Method

In order to follow the motion of the tongue we 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.

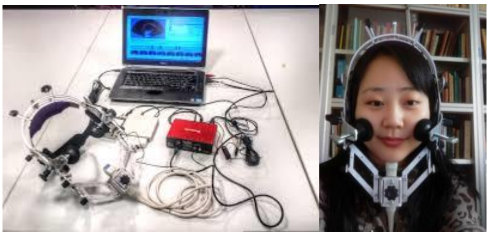


Figure 3

Left: ‘Micro' Ultrasound System. Right: Probe Stabilization Headset installation

A photograph of this instrumentation is presented on the left side of Figure 3, 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 [67]. The headset fixes the subject's head while capturing the images and also fixes the ultrasound transducer under the chin [49].

4.4 Tongue Movement Contour Tracking

Problems revealed in biomedical image analysis such as user fatigue, user bias, and difficulty in reproducing results also may occur when manually tracking tongue contours [68]. 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 shapes that represent clusters of contrastive phonemes [69]. 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 [70], our dynamic viseme model blends dominance and parameter values.

4.4.1 Dominance Classification for the Shaanxi Xi'an Dialect

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 we 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 4, (b) and (c) show the tongue contour comparison in the frame belonging to the burst of 't' and 'p' in the two structures.

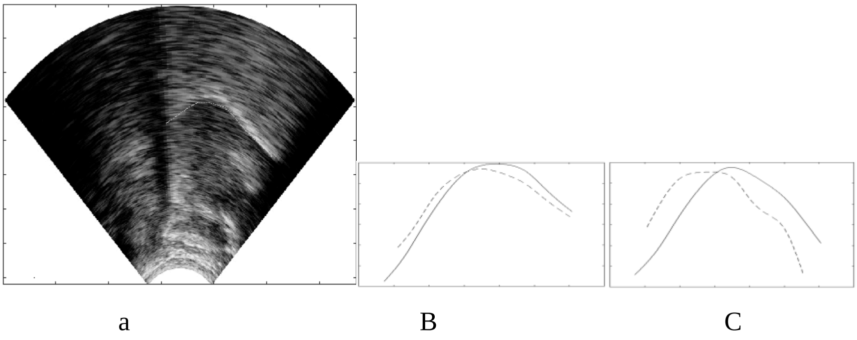


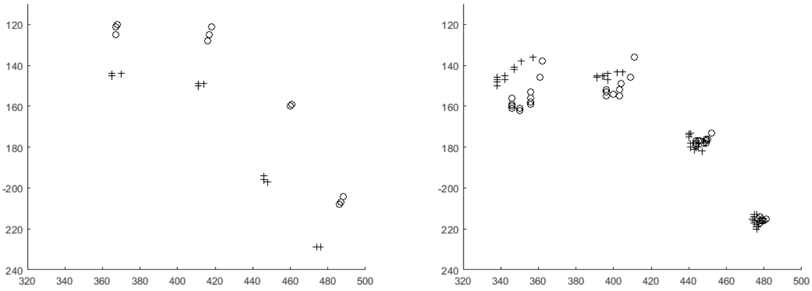
Figure 4

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' (- -)

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 our future research, we plan to focus on a sequence of frames in order for more accurate classification of the dominance grade of viseme features.

In Figure 4, the complete tongue contour that can be seen in the ultrasound image is shown after automatic contour tracking, the uneven curve being 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 [71]. In Figure 5 (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 burst (altogether 36 ms). Figure 5 (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 4, 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 5

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

The movement of the tongue during speech can be described with the changes in the coordinates of the feature points. Figure 6 shows the vertical positions of the two front feature points while 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 possible the investigation of the interpolation between key frames.

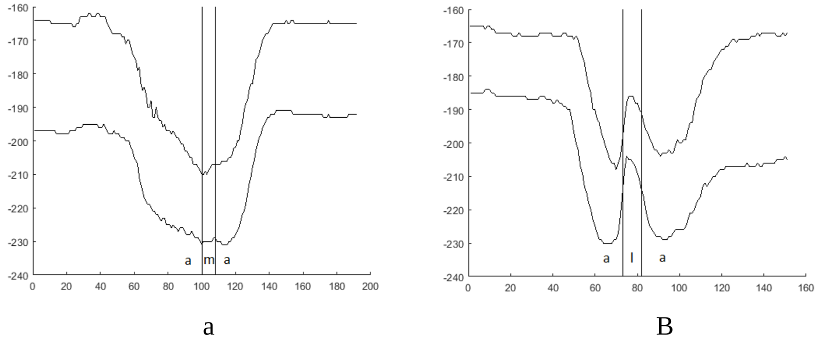


Figure 6
Vertical position of the first two feature points of the tongue while the words a: ’ama’ and b: ’ala’ are being uttered

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 shape of the lips but are of uncertain character as regards the position of the tongue.

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.

5 Conclusion and Future Work

This paper presents a method for the phonetic transcription of the Shaanxi Xi’an dialect of Chinese and the conversion of its basic phomes 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. The purpose is to obtain the fundamental data needed to create a talking head for the Shaanxi Xi’an dialect. The classification method for Mandarin static visemes was applied to static viseme classification of Chinese Shaanxi Xi’an dialect speech. We 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 41 vowels. The 'Micro' ultrasound system recorded the utterances and the Assistant Advanced software formed JPG images and MP4 videos. We 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 in progress 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 4-6: 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.

The long-term objective is to create a dynamic articulation model that can be applied to animate articulation for 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.

Acknowledgement

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.

References

[1] Czap L, Mátyás J: Hungarian talking head. Proceedings of Forum Acusticum 4th European Congress on Acoustics. Budapest, Hungary, 2005, pp. 2655-2658

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

[3] 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

[4] 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)

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

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

[7] Lu Tuanhua: Comparison of Phonetic Features and Pronunciation of Mandarin in Xi'an Dialect [J] Journal of Test Sciences, 2010, 9: 23-24 (In Chinese)

[8] Arora K K, Arora S, Singla S R: SAMPA for Hindi and Punjabi based on their Acoustic and Phonetic Characteristics [C]//Proc. International Oriental COCOSDA 2007 Conference (Hanoi, Vietnam. 2007: 17-22

[9] Kabir H, Saleem A M: Speech assessment methods phonetic alphabet (SAMPA): Analysis of Urdu [J] CRULP Annual Student Report published in Akhbar-e-Urdu, 2002

[10] Tseng C, Chou F: Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan [J] Journal of the Acoustical Society of Japan (E) 1999, 20(3): 215-223

[11] Zu Y, Chen Y, Zhang Y: A super phonetic system and multi-dialect Chinese speech corpus for speech recognition [C] International Symposium on Chinese Spoken Language Processing, 2002

[12] 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

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

[14] 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

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

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

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

[18] 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

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

[20] 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

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

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

[23] Montgomery D: Do dyslexics have difficulty accessing articulatory information? [J] Psychological Research, 1981, 43(2): 235-243

[24] Vicsi, K., Hacki, T.: 'CoKo' - Computerised audiovisual feedback speech-tutoring system for children with articulation disorders and impaired hearing. (In German: 'CoKo' - Computergestützter Sprechkorrektor mit audiovisueller Selbstkontrolle für artikulationsgestörte und hörbehinderte Kinder.) Sprache-Stimme-Gehör 20, 141-149, 1996

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

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

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

[32] Czap L, Pintér J M, Baksa-Varga E: Features and Results of a Speech Improvement Experiment on Hard of Hearing Children Speech Communication 2019 106 pp. 7-20, 14 p.

[33] Czap, L.: Speech Assistant System. INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association,

Singapore, International Speech Communication Association (ISCA) pp. 1486-1487

[34] Baranyi, P., Csapo, Á., Sallai, Gy.: Cognitive Infocommunications, Springer, 2015

[35] Baranyi P., Csapó Á.: Definition and synergies of Cognitive Infocommunications, Acta Polytechnica Hungarica, 9 (1) pp. 67-83, 2012

[36] CogInfoCom - Cognitive Infocommunications www.coginfocom.hu, accessed: 29.05.2018

[37] Sallai, G.: The Cradle of the Cognitive Infocommunications, Acta Polytechnica Hungarica 9 (1) pp. 171-181

[38] D. W. Massaro, M. M. Cohen: Integration of visual and auditory information in speech perception, Journal of Experimental Psychology Human Perception & Performance November 1983, pp. 753-771

[39] Kovács, S., Tóth, Á., Czap, L. “Fuzzy model based user adaptive framework for consonant articulation and pronunciation therapy in Hungarian hearing-impaired education.” CogInfoCom 2014: Proceedings. 2014, pp. 361-366

[40] 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

[41] Napadow V J, Chen Q, Wedeen V J: Intramural mechanics of the human tongue in association with physiological deformations [J] Journal of Biomechanics, 1999, 32(1): 1-12

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

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

[44] Schönle P W, Gräbe K, Wenig P: 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

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

[46] 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

[47] 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

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

[49] 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

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

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

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

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

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

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

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

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

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

[59] Wells J: Computer-coding the IPA: a proposed extension of SAMPA. http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm, accessed: 11.07.2017

[60] Zhao L: The correspondent regularities between Mandarin and Shaanxi Dialect [J] Journal of Baoji University of Arts & Sciences, 2008, Vol. 28, 1 (In Chinese)

[61] Yang Jinfeng: Corresponding speech sound in west Shannxi Dialect and Mandarin. [J] Journal of Xianyang Teachers’ College. 2003 Vol. 18, 5 (In Chinese)

[62] Wang Y: Research on the types of phonetic changes in Xi’an Dialect. [J] Journal of Yanan University (Social Science Edition) 1995, Vol. 17, 2 (In Chinese)

[63] Wu Z, Zhang S, Cai L: Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar[C]//INTERSPEECH. 2006 (4): 1802-1805

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

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

[66] Czap L, Zhao L: Phonetic Aspects of Chinese Shaanxi Xi'an Dialect. 8th International Conference on Cognitive InfoCommunications: CogInfoCom. Debrecen, Hungary, Piscataway: IEEE Computer Society, 2017, pp. 51-56

[67] 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

[68] 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, Journal of the Acoustical Society of America. 2016 Vol. 139, 5 pp. 154-160

[69] 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

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

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