UNIVERSITY OF MISKOLC

FACULTY OF MECHANICAL ENGINEERING AND INFORMATICS

MISKOLCI EGYETEM

UNIVERSITY OF MISKOLC

Dynamic Modeling for Chinese Shaanxi Xi'an Dialect Visual Speech

Summary of PhD dissertation

Lu Zhao

Information Science

'JÓZSEF HATVANY' DOCTORAL SCHOOL

OF INFORMATION SCIENCE, ENGINEERING AND TECHNOLOGY

ACADEMIC SUPERVISOR

Dr. László Czap

Miskolc 2020

Content

Introduction

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. I create articulation features and a dynamic modeling system for visual representation of speech sounds for a Chinese Shaanxi Xi'an Dialect talking head. 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 a speech conversion to other language.

Xi'an was the capital of 13 dynasties in ancient China and it still occupies quite animportant position in northwest China. The Shaanxi Xi'an Dialect, 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.

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 [1]. 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 [2]. 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.

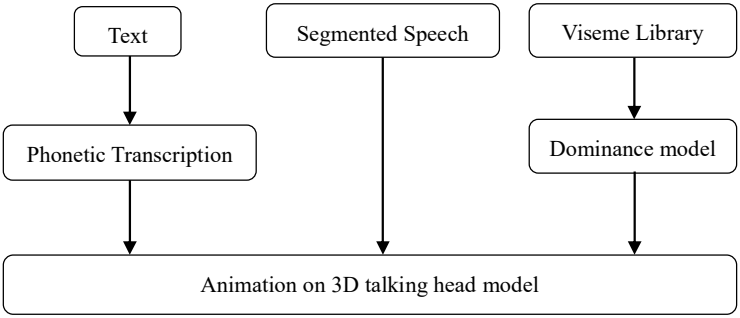


Figure 1. Structure of the Chinese Shaanxi Xi'an Dialect talking head system

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

Chapter 1

Phonetic Aspects 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.

The phonemes of the Chinese Shaanxi Xi'an Dialect are described by IPA, and there are 26consonants and 40 vowels [3][4][5].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 1.1 with gray background: 'pf' and 'pfʰ' 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 [6][7].

It is common to describe Chinese tone by tone type and tone pitch. The five - degree mark method [8] is used to annotate the changes of tone pitch. This Thesis makes comparisons between Mandarin and the Shaanxi Xi'an Dialect according to the tone pitch which given by Peking University [9].

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

The X-SAMPA code derived for the Chinese Shaanxi Xi'an Dialect is based on the Hungarian X-SAMPA code. Table 1.1 shows the X-SAMPA symbols for unique consonants of Chinese Shaanxi Xi'an Dialect.

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

Character	*Pinyin expression*	*IPA*	*X-SAMPA*
追		pf	pf
吹		pfʰ	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. The translation of unique vowels into X-SAMPA code [10][11][12] is created and presented in Table 1.2. The translation of the phonemes 'v' and 'ŋ' to X-SAMPA code is based on the Hungarian phoneme system.

Table 1. 2. X-SAMPA for vowels of the Chinese Shaanxi Xi'an Dialect

Character	*Pinyin expression*	*IPA*	*X-SAMPA*
哀		æ	{
岩		iæ	i{
歪		uæ	u{
安		æ̃	{~
岩		iæ̃	i{~
弯		uæ̃	u{~
冤		yæ̃	y{~
恩		ɛ̃	e~

Character	*Pinyin expression*	*IPA*	*X-SAMPA*
因		iẽ	ie~
温		uẽ	ue~
晕		yẽ	ye~
核		ɯ	M
药		yo	yo

Chinese Shaanxi Xi'an dialect tone translation expression is showed in Table 1.3.

Table 1. 3. 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>

1.2. 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 1.1, Table 1.2, and Table 1.3. 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.

This process is like this: 赵璐(Chinese characters)——zhāolū (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)

Chapter 2

Visemes of the Chinese Shaanxi Xi'an Dialect

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.

2.1. Quantitative description of lip static visemes

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 subject was one adult female – the author of this dissertation – who is a native speaker of Chinese Shaanxi Xi'an Dialect. In my research 26 consonants and 40 vowels as well some prescribed syllables were recorded by a Pentax K-30 camera for the quantitative analysis of visemes for Chinese Shaanxi Xi'an Dialect.

This process allows control of a wide range of motions using a set of parameters associated with different articulation functions. So, we need to analyze the images obtained of the 26 static lip visemes based on the parameters in Figure 2.1. We can establish the facial animation on lip with these parameters for lip visemes.

pf,pv ɲ ŋ

Figure 2. 1. Consonants 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. Figure 2.2 shows the similarity of the same viseme in the speaker's photograph and the 3D model.

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

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.

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

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. I use the 'Micro' ultrasound system (Articulate Instruments Ltd.). to make the recordings as shown in Figure 2.3. The subject was also the author of this thesis.

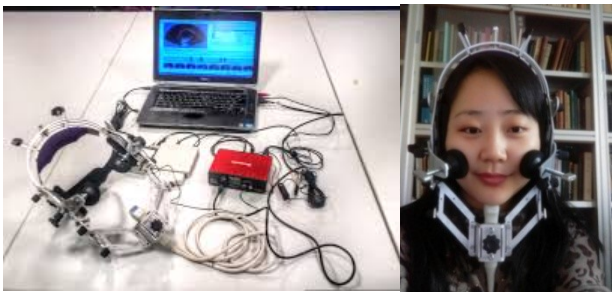
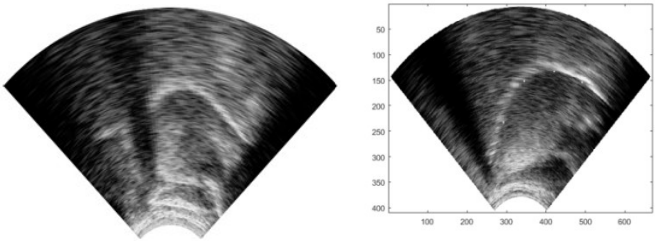
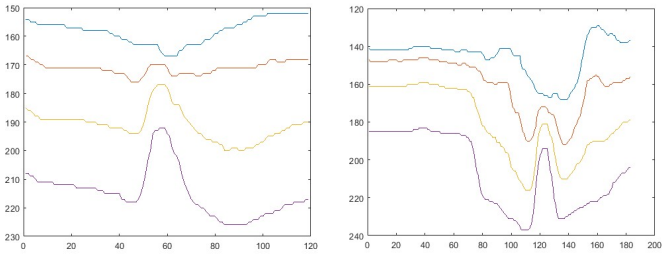


Figure 2. 3. Left: 'Micro' Ultrasound system. Right: Probe stabilization headset installation

After recording the ultrasound images of phonemes and sentences, I can obtain tongue feature points by processing these images. The contour of static tongue visemes was obtained by dealing with the ultrasound images recorded in the experiment based on the algorithm I developed.



(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. 4. Tongue contour tracing in Matlab

All of the tongue feature points tracking for each frame of the speech are gained based on this algorithm. Figure 2.4 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.

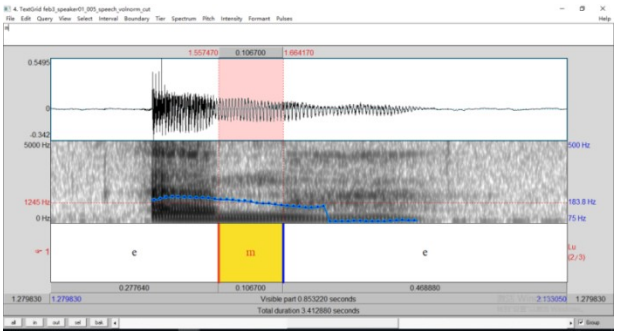


Figure 2. 5. Central frame of 'm' in the phrase 'eme' chosen in Praat

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’ in Figure 2.5.

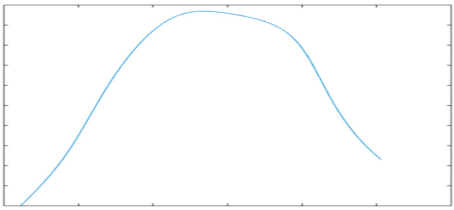


Figure 2. 6. Tongue contour of central frame of 'm' in the phrase 'eme'

As we can see, the tongue contour of the central frame is showed in Figure 2.6. We can get the static viseme of the tongue for each phoneme at each frame though this processing method.

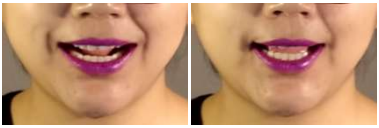
Chapter 3

Dynamic Modeling of the Chinese Shaanxi Xi'an Dialect Speech

I will synthesize a realistic speech animation based on the rule set and the visual dynamic articulation model based on the dominance model [13]. Features of the parametric model can be divided into four dominant grades: stable, dominant, flexible, uncertain.

3.1. The interaction between phonemes and the corresponding lip shape

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. 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.



Lip visemes of central frame of 'a' in the phrase 'shash' (left) and 'sas' (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. The consonants are divided into four different dominance grades based on the degree of influnence by the vowels before and after it.

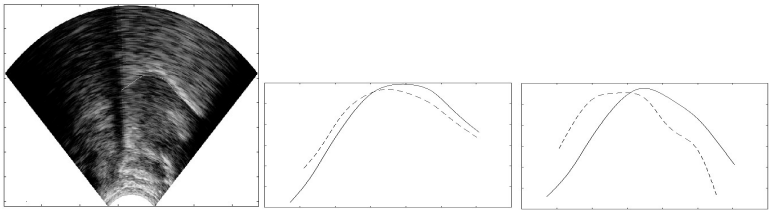
Table 3. 1 shows the different dominance grades of lip visemes of the Chinese Shaanxi Xi'an Dialect.

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, ɲ	flexible	dominant
vowels	dominant	dominant

3.2. Dynamic tongue viseme classification

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.



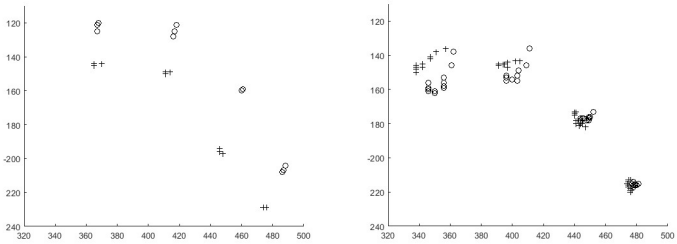
a b c

Figure 3. 2.

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’ (- -)

In Figure 3.2, (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.

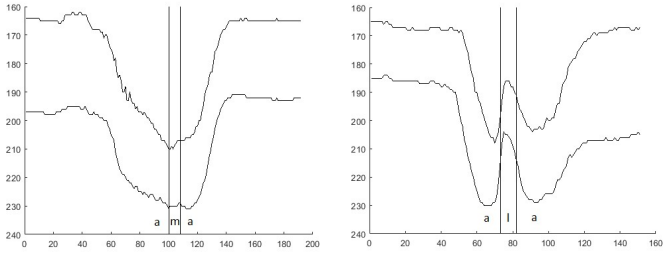


a b

Figure 3. 3.

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

In Figure 3.2, 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. In Figure 3.3 (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 3.3 (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.



a b

Figure 3. 4.

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

Figure 3.4 shows the vertical positions of the two front feature points while the VCV words ‘ama’ and ‘ala’ are being uttered.

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.

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, pfʰ	dominant
ŋ, ɲ, h	dominant
vowels	dominant

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

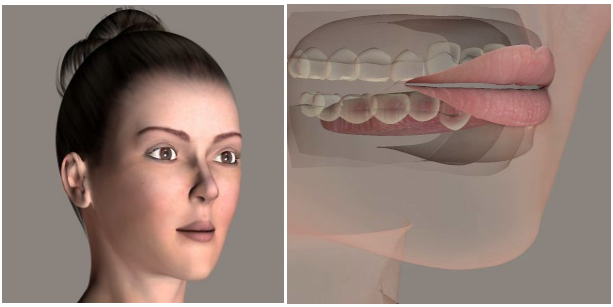


Figure 3. 5. Face animation of Chinese Shaanxi Xi'an Dialect talking head in software Poser Pro (left) and Hungarian speech assistant system (right)

Figure 3. 5. shows the 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 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"[14].

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. 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. 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.

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 and video 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.

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 journals

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 journals

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

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

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

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

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

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

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

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

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

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

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

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

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

[14] 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.