Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

A comprehensive audio-visual corpus for teaching sound Persian phoneme articulation

Azam Bastanfard
IRIB University
19395-1746, Niayesh Ave.
Tehran, Iran
bastanfard@iribu.ir

Maryam Fazel
IRIB University
19395-1746, Niayesh Ave.
Tehran, Iran
mfazzel@yahoo.com

Alireza Abdi Kelishami
Department of Electrical, Computer and IT Engineering
Qazvin Azad University
Qazvin, Iran
alireza.abdi@yahoo.com

Mohammad Aghaahmadi
Department of Electrical, Computer and IT Engineering
Qazvin Azad University
Qazvin, Iran
aghaahmadi@ymail.com

***Abstract*—Building an audio-visual data corpus is one significant step in audio-visual research. One of the most challenging tasks in computer science is computer-aided speech therapy and language learning. Developing computer applications for training and rehabilitation of the handicapped and helping the hearing and speaking-impaired by facial speech synthesis are among the most helpful, state-of-the-art roles of computer technology in today's human-machine interacting systems. To date, there have been no audio-visual corpora in Persian language, in that it makes it difficult or even impossible for researchers to carry out studies in the area. This paper gives an indication of the collected Persian audio-visual data corpus. AVA is a comprehensive, systematic collection of both continuous speech and isolated spoken utterances in Persian language. The goal of this project is to facilitate audio-visual research in the language through this data corpus which is available upon request.**

***Keywords*—Audio visual data, Corpus design, Speech therapy.**

I. INTRODUCTION

In recent years, audio-visual (AV) speech processing has received great attention because of the revolutionary potential it has to improve the human-machine interaction. Multi-modal data can correct the many defects caused by uni-modal data [4]. Computer aided rehabilitation for speaking and hearing-impaired is one of the most prominent and influential goals in computer science. Some distinctions cannot be distinguished with degraded hearing, even when the hearing loss is compensated by hearing aids or cochlear implants [33]. Observations confirm that visual information improves the intelligibility of both synthetic and natural acoustic speech by 15% to 50% [32]. Deaf children can perceive 70% of speech information if both audio and video are presented, while human teachers are insufficient for teaching them, in that many of these children can only understand their teacher. This provides motivation for developing an automated facial animation. There are software like [6] that have been developed for such uses but can be used merely for their target language, in that they have been designed based on a language's articulatory and vismic information which are unlike that of a different language.

Yet, AV data corpora have been collected in many languages, e.g. English, German, Chinese, Turkish, Indian, French, Dutch, Czech and Arabic. This paper presents the collecting process of the first Persian AV data corpus, AVA, which is to help training and rehabilitation of the handicapped and language learners [6,14].

In what follows, section 2 includes a survey of the existing AV datasets and motivation behind a more robust design approach of data corpus. Section 3 introduces the framework of the presented AV data corpus, where the spoken material and the recording setup are entailed. The result comes in the 4th section, followed by suggestions and future work in the area.

II. RELATED WORKS

A number of corpora have been produced over the past few years. Only few databases appear to be designed based on a comprehensive phonemic visemic analysis in recording, which causes them not be perfectly useable.

Some databases are designed for recognizing digits such as CUAVE containing 36 speakers uttering numbers [22,23] and M2VTS [25]. Some are developed to enable speech recognition e.g. Manssa L.K [18], AVOZES [10,11], Mandrain Chinese [17], Czech [5,26] and Dutch [27]. In BANCA [1], a 12 digit number, speaker's name, his address and date of birth is uttered.

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

Much as researches have been done to develop several AV data corpora in various languages, none has emerged as a thoroughly reusable one; for they meet their former goal applications' requirements, where for a latter unforeseen project, they more probably fall short of meeting the new requirements. M2VTS and XM2VTS [19] for example, cover isolated digits. Most do not record the profile view of respondents, which makes it not be useable for some certain applications [22]. Data corpora can be used as computer assistive technology [24].

Some corpora such as [12] are built for covering emotional speaking faces and others in different environments such as car [16] or inside smart meeting rooms [21]. A survey of audio visual databases is prepared in [3] listing many corpora, where they cover mono-modal or limited, small spoken material. Other research [34] discusses speech characteristics of the deaf; reviews computer-based speech training aids to the hearing-impaired.

Needless to say, no data corpus has been designed fully reusable. But, moving towards taking some basic factors into consideration to make it more reusable can be considerably time and cost reducing in future development. The proposed method explains a new AV corpus aiming at assisting speech therapists for teaching sound Persian phoneme articulation. In this project, we aimed at facilitating the future works in the area for all researchers by choosing the framework described in next sections.

III. Database Design Framework

In order to efficiently collect the data, a framework [20] was considered, in which the targeted application of data corpus had to be declared in the first step. The applications are:

- Develop and build:
  - application for teaching Persian phoneme pronunciation to the hearing and speaking-impaired [6] and children suffering from delay speech and language (DSL) [7,35];
  - application for language learners studying Persian as their second language or for Iranian children living worldwide in non-Persian speaking environments [7];
- Persian viseme classification;
- Talking head generation [29].

One significant point that should be taken into consideration is that, the designed data corpora have been always developed based on their target application. A thorough analysis was made after considering an appropriate lighting and acoustic environment in a professional studio for recording. The following are to be discussed here: the spoken material, respondent selection and recording setup.

As stated, because there had been no other Persian AV data corpora, and in order to satisfy the essential need of the various applications in the language, the decision was to develop AVA such that it would cover all Persian phonemes in existing and possible syllables, other than continuous speech and digits.

A. *The spoken material*

Here, a research had to be made to collect and classify the Persian language phonemes; where in other designed corpora, this is an available preliminary data. In Persian, there are 23 consonants and 6 vowels. In figure 1, Persian and other languages' vowels [28,15] are shown in IPA [31]. The vertical and horizontal axes represent mouth opening size and place of articulation respectively. Symbols in dark italic font represent Persian vowels.

We covered digits and all utterances within cv, vc, cvc, vcv and cvcc syllables in the language. As studied and stated in [9], 20 common sentences are chosen, where the phoneme proportion is balanced. This means that phonemes' occurrence frequency in the sentences is like that in Persian language's daily speech. Serial digits are included as well. Two examples are shown in table 1.

Table 2 presents all possible and captured forms of consonant /b/ in various combinations. AVA covers this combination for all consonants. In cv and vc cases an example is given. The collected place and manner of articulation of consonants [13] in Persian language are shown in table 3.

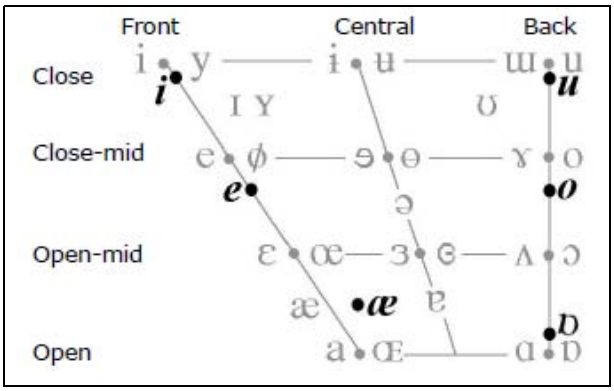

Figure 1. Vowels in IPA and Persian language

Table 1 Example of two common sentences

IPA form	**/sobhe hæmegi: be kheir/**
Persian form	صبح همگی بخیر
IPA form	**/sæbr d ʃte b ʃ/**
Persian form	صبر داشته باش

Table 2 Syllables of consonant /b/

c=b	**/u/**	**/o/**	**/ /**	**/æ/**	**/e/**	**/i/**
cv	bu /bud/	bo /boro/	b /b b /	bæ /bæle/	be /bede/	bi /bit /
vc	ub /xub/	ob /sob/	b /x b/	æb /ædæb/	eb /teb/	ib /sib/
cvc	cuc	coc	c c	cæc	cec	cic
vcv	vbv	vbv	vbv	vbv	vbv	vbv
cvcc	cucc	cocc	c cc	cæcc	cecc	cicc

Table 3 Consonants in Persian language

Place of articulatio / Manner of articulatio	Labial	Alveolar	Post alveolar	Palatal	Velar	Uvular	Glottal
Nasal	m	n			[ŋ]		
Plosive	p b	t d			K g	ɢ	[ʔ]
Affricate			tʃ dʒ				
Fricative	f v	s z	ʃ ʒ	x ɣ			h
Tap	[ɾ]						
Trill	r						
Approximant	l		j				

B. Respondent Selection

The selected first two speakers were preferred to have the standard Persian accent, Tehrani. According to Bench 1995 [2], were people divided into four groups of young and old male and female, young ladies convey most audio-visual information. Hence, lip reading of young ladies is the simplest aware of sound speaking techniques, and a computer engineer were captured. Here, a linguist was chosen, resulting in producing higher standard phonemes, because the designed data corpus had to be, firstly, fully usable for developing the teaching Persian phoneme pronunciation. The other person was captured, for it was preferred to have an ordinary prototype; because, for talking head generation [8], the head has to look like an ordinary talking face. Neither of the speakers have articulation nor voice disorder, make ups nor cosmetic surgery.

C. Recording Setup

Having an appropriate recording setup is an important factor in collecting a more reliable data corpus; however depending on the application for which the corpus is developed, these conditions differ. In our approach, the lighting and acoustic conditions are considered with great care. This setup is comprised of the following segments:

- Studio layout and speaker position
- Recording equipment
- Prompts (slides)

This setup is described in the following.

1) Studio Layout and Speaker Position

In order to collect the data corpus in an appropriate professional environment, the IRIBU, Iran's broadcasting university, TV studios was chosen. The studio is an ideal place for audio-video recordings either for its perfect lighting or its acoustic environment. This condition was selected in that it should meet the stated application requirements. To have the minimum shades on speaker's face 4 portable projectors other than various hanging projectors were used. A blue curtain is the background of the three cameras filming different views of the speaker.

Figure 2. Studio setup in different perspectives

The speaker sits on a fixed chair, while the first and second cameras record the profile and frontal view, respectively, while the third one records the lip area. The first two capture the speaker's shoulders and head, and the third camera has the entire chin up to the middle of the nose in frame for all speakers. For both respondents, the position of the cameras and light projectors were fixed, as presented in Fig. 2.

2) Recording Equipment

As stated, three cameras were used, two Canon XL2 cameras and one PD150 SONY. The first XL2 camera records a profile view of the speaker, named Cam 1 for quick reference. The other XL2 camera records the full view, Cam2; while the SONY camera, Cam 3, shoots the lip area with a slight difference in location to Cam 2. Recording the video from both frontal and profile view enables face 3D modelling, as well as anterior and posterior lip movement identification while uttering different phonemes with the same visemes (the visual form of phoneme), as in /m/ and /p/. Before data recording, camers' lens focus, white-balance and brightness were automatically calibrated. Prompts are shown on an LCD located close to the front camera. Two other monitors were used, to control the filming process. The first one monitors the video output from Cam 3 to the speaker to keep her face in frame by some easy eye movements. The other monitor is positioned in the control room. This one is used to control the video of the speaker and to correct the probable problems in the recording process. The location of the cameras and lights are shown in Fig. 3 and 4.

Other than the XL2 cameras' sound recording system, an AKG collar microphone is used to record the speaker sound. The microphone is placed under the scarf, heading outwards to minimize the noise made by clothes. The AKG output is plugged into a mixer, and from the mixer to both the SONY camera and the MOTU external sound card in the control room.

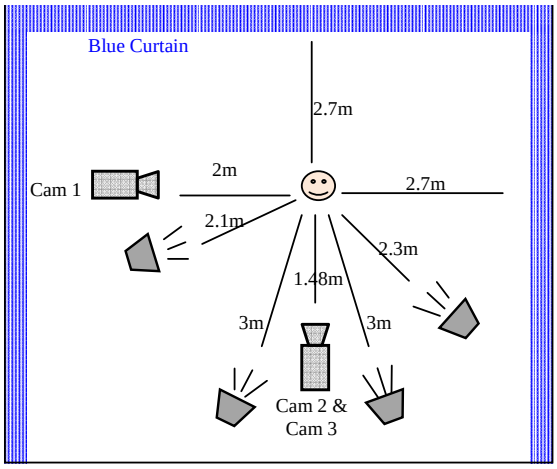

Figure 3. Recording equipment, vertical view

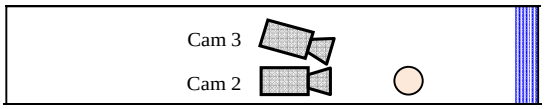

Figure 4. Cam 2 & cam 3, horizontal veiw

The MOTU output is then recorded using Adobe Audition software. The recorded sound is in WAVE, 16 bps and 48 KHz (mono). The video is captured in Digital Video (DV) format, 25 Hz frame rate and 720x576 pixels of resolution. The entire data is stored on 40 4-GB DVDs covering all current and possible phoneme combination in Persian language.

3) Prompts

The spoken materials that the speaker had to utter are prepared in power point slides. Firstly, 20 common, frequent sentences [9] are pronounced, then all possible forms of cv, vc, cvc, vcv words and all existing forms of cvcc in the language are covered. The vowels are pronounced separately. In the end, each speaker reads an isolated digit sequence. Prompts are shown on an LCD next to the front camera and are 5500 slides, where both respondents had 11000 slides together. To avoid unwanted phonemic and visemic coarticulation and for the speaker not to get tired, each slide is passed in 5 seconds of interval.

IV. Conclusion

In this paper we presented AVA, the first audio-visual data corpus in Persian language. Previously, there had been no conducted audio-visual researches in the area for Persian language. The objective was to build a data corpus with high quality recordings in which reusability is a major concern other than its usability. Usability is a factor that would be brought up by thorough initial analysis and completeness of corpus design. Reusability, as stated in the context, is a quality upon which the data corpus designed for certain applications, could later be extended and updated so that it would be considerably energy and cost-reducing.

We recorded the data with 3 cameras filming from different aspects to enable 3D modelling. Pictures from the corpus are illustrated in Fig. 5, where visemes in different views from the

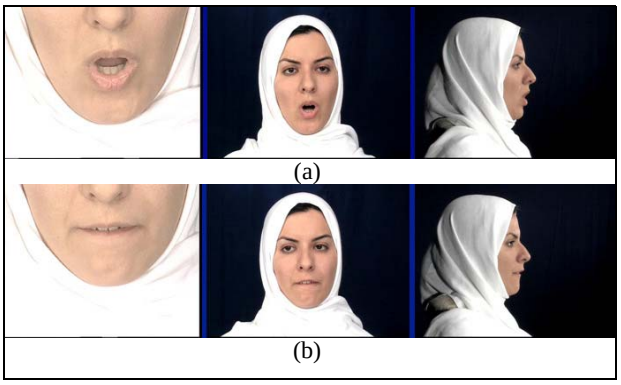

Figure 5. Pictures from the corpus

three cameras. In this figure, (a) represents all three perspectives of speaker pronouncing /b n/; where (b) illustrates the speaker when uttering /me**v**/.

The data was captured in IRIBU studio with great care in lighting and noiselessness. A great concern was to develop the corpus so that it could be used by other researchers to conduct studies in the area. Removing the face from shadows by devising several projectors, minimizing the noise and covering all possible utterances in the language were considered to accomplish this objective. The data corpus is available upon request.

V. Future Work

As depicted, AVA is designed and developed such that it would better satisfy the requirements of the initial targeted applications. This made us cover the entire possible phoneme combinations in the language. Here, for each speaker 5500 utterances had to be pronounced taking 15 hours for two speakers, uttering 11000 utterances. This was a boring task for the speaker; however, it was done to meet the requirements of the stated applications. For further development of AVA, the decision is to increase the number of respondents to more speakers and to decrease the time interval taken for each speaker, enabling recording more respondents in that it makes the data corpus more robust for more of today's state-of-the-art technologies and other applications as in AV automatic speech recognition (AVSR) and lip synchronization [35] in Persian language.

References

[1] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J. P. Thiran. "The BANCA Database and Evaluation Protocol" In Proceedings of Audio and Video Based Biometric Person Authentication", Springer Berlin/Heidelberg, Volume 2688, pp. 625-638, 2003.

[2] Bench J, Daly N, Dayle J, Lind C. "Choosing talkers for the BKB/A Speechreading Test: a procedure with observations on talker age and gender" ,British Journal of Audiology. Volume 29 ,Issue 3,pp. 172-187, Jun 1995.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] C.C. Chibelushi, F. Deravi, J.S.D. Mason, "Survey of audio visual speech databases", Tech. Rep., Department of Electrical and Electronic Engineering, University of Wales, Swansea, UK, 1996.

[4] Alin G. ChiŃu and Leon J.M. Rothkrantz. "Building a Data Corpus for Audio-Visual Speech Recognition" AGC, pp. 88-92, April 2007.

[5] P. Cisar, M. Zelezny, Z. Krnoul, J. Kanis, J. Zelinka, L. Müller “Design and recording of Czech speech corpus for audio-visual continuous speech recognition”. Proceedings of the Auditory-Visual Speech Processing International Conference 2005, AVSP2005, p. 1-4,, Vancouver Island, 2005.

[6] Dr Speech, A Training software system, http://www.drspeech.com.

[7] O. Engwall, O. Bälter, A.-M. Öster and H. Kjellström “Designing the human–machine interface of the computer-based speech training system ARTUR based on early user tests”. Behaviour and Information Technology.Volume 25 ,Issue 4, pp. 353–365. 2006.

[8] T. Ezzat and T. Poggio, “Visual Speech Synthesis by Morphing Visemes”, International Journal of Computer Vision, Volume 38, pp: 45-57, 2000.

[9] Gita Movalleli “Sara Lip-Reading Test: Construction, Evaluation and operating on a group of people with hearing disorder” MSc Thesis, Department of Rehabilitation in Tehran University of medical science, In Persian, 2002.

[10] R. Goecke, J. B. Millar, A. Zelinsky, and J. Robert-Ribes, “A detailed description of the AVOZES data corpus”, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01), pp. 486-491, Salt Lake City, Utah, USA, May 2001.

[11] R. Goecke, J. B. Millar, “The Audio-Video Australian English Speech Data Corpus AVOZES” Proceedings of the 8th International Conference on Spoken Language Processing, ICSLP2004, Volume III, pp. 2525-2528, 2004.

[12] M. Grimm and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database", ICME 2008, IEEE, pp. 865-868, April 2008.

[13] C. Jahani, “The Glottal Plosive: A Phoneme in Spoken Modern Persian or Not?” In Éva Ágnes Csató, Bo Isaksson, and Carina Jahani. Linguistic Convergence and Areal Diffusion: Case studies from Iranian, Semitic and Turkic. London: RoutledgeCurzon. pp. 79–96. ISBN 0-415-30804-6, 2005.

[14] H. Kjellstrom, O. Engwall, “Audiovisual-to-articulatory inversion”. Speech Communication, Volume 51, Issue 3, pp. 195-202, 2009.

[15] Peter Ladefoged, “Vowels and Consonants”, Blackwell Publishers pub, 2nd. Ed, ISBN: 978-1-4051-2458-4.l , 2004.

[16] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, T. Huang “AVICAR: Audio- Visual Speech Corpus in a Car Environment” In Proceedings of International Conference on Spoken Language Processing – INTERSPEECH, Jeju Island, Korea, October 4-8, 2004.

[17] L. Liangi, Y. Luo, F. Huang and Nefian, A.V, “A multi-stream audio-video large-vocabulary mandarin Chinese speech database” In IEEE International Conference on Multimedia and Expo, Volume 3, pp. 1787 – 1790, 2004.

[18] Lynn K. Marassa and Charissa R. Lansing “Visual Word Recognition in Two Facial Motion Conditions: full-face versus Lip plus Mandible” Journal of speech and hearing Research. Dec; 38(6): pp. 1387-94. 1995.

[19] K. Messer, J. Matas, J. Kittler, J. Luettin, and G Maitre, “XM2VTSDB: the extended M2VTS database,” in Proceedings of the 2nd International Conference on Audio-and Video-Based Biometric Person Authentication”, (AVBPA '99), pp. 72-77, Washington, DC, USA, March 1999.

[20] J.B. Millar, M. Wagner, and R. Goecke, “Aspects of Speaking-Face Data Corpus Design Methodology” In Proc. 8th Int. Conf. Spoken Language Processing ICSLP, Volume II, Jeju, Korea, pp. 1157–1160, 2004.

[21] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S.M. Chu, A. Tyagi, J.R. Casas, J. Turmo, L. Christoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, “The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms”, Journal of Language Resources and Evaluation, Springer, Volume 41, pp: 389-407, 2008.

[22] E. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, “Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus” EURASIP Journal on Applied Signal Processing, Volume 2002, pp. 1189– 1201, 2002.

[23] E. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, “CUAVE: a new audio-visual database for multimodal human computer-interface research,” in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), Volume. 2, pp. 2017-2020, Orlando, Fla, USA, May 2002.

[24] V. Pera, A. Moura, D. Freitas, “LPFAV2: a New Multi-Modal Database for Developing Speech Recognition Systems for an Assistive Technology Application”, SPECOM’2004: 9th Conference Speech and Computer St. Petersburg, Russia September 20-22, 2004.

[25] S. Pigeon and L. Vandendorpe, “The m2vts multimodal face database (release 1.00). AVBPA ’97: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, pp. 403–409, London, UK. Springer-Verlag, 1997.

[26] J. Trojanová, M. Hrúz, P. Campr, and M. Železný, “Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition” Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008.

[27] J.C. Wojdeł, P.Wiggers, and L.J.M. Rothkrantz, “An audiovisual corpus for multimodal speech recognition in Dutch language”, Proceedings of the International Conference on Spoken Language Processing, ICSLP2002 Denver CO, USA, September, pp. 1917-1920, 2002.

[28] Yadollah Samareh, Persian phonetics, Markaze nashre daneshgahi pub, Tehran, in Persian, 1998.

[29] T. Yotsukura, S. Nakamura and S. Morishima, “Construction of audio-visual speech corpus using motion-capture system and corpus based facial animation”, The IEICE Transaction on Information and System E 88-D, 11, pp. 2377—2483, 2005.

[30] I. Kirschning, T. Toledo, "Language Training for Hearing Impaired Children with CSLU Vocabulary Tutor", Journal: WSEAS Transactions on Information Science and Applications, Issue 1, Vol. 1, pp. 20-25.July 2004.

[31] International Phonetic Association. “Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet”, Cambridge: Cambridge University Press. pp. 124–125. ISBN 978-0521637510, 1999.

[32] R. Möttönen, J-L. Olivés, J. Kulju and M. Sams Parameterized Visual Speech Synthesis and Its Evaluation. Tampere, Finland, Proc. of EUSIPCO 2000.

[33] Cohen, Michael M., Beskow, Jonas, and Massaro, Dominic W. “Recent Developments in facial animation: An inside view”. In proceedings of auditory visual speechperception, Pages 201—206. Terrigal-Sydney Australia, December, 1998.

[34] L. Bernstein, M. Goldstein, J. Mashie, "Speech training aids for hearing-impaired individuals", Journal of Rehabilitation Research and Development,Volume 25,pp.53-62, 1988.

[35] G. Zorić and S.I. Pandžić, “Real-time language independent lip synchronization method using a genetic algorithm”, Signal Processing, Volume 86, Issue 12, pp: 3644-3656, 2006.