PAVID-CVs: Persian Audio-Visual Database of CV syllables

Mahsa Hedayatipour
Faculty of Computer Science and Engineering, Shahid Beheshti University
Tehran, Iran
m.hedayatipour@mail.sbu.ac.ir

Yasser Shekofteh
Faculty of Computer Science and Engineering, Shahid Beheshti University
Tehran, Iran
y_shekofteh@sbu.ac.ir

Mohsen Ebrahimi Moghaddam
Faculty of Computer Science and Engineering, Shahid Beheshti University
Tehran, Iran
m_moghadam@sbu.ac.ir

Abstract— **Lip-reading is a visual speech recognition process. In this process, recognizing the smaller units of speech can be the basis for recognizing the larger units of a language such as words. In this paper, we have introduced a Persian (Farsi) Audio-Visual Database of CV syllables, named PAVID-CVs, as a set of isolated two-phoneme visyllable and isolated words related to the visyllables, which include only Persian CV syllables, for lip-reading or audio-visual speech recognition purposes such as isolated word recognition. This dataset can be used for machine learning-based methods due to its useful tagged information. Here, we explain the steps of preparing the database. It contains about 30 hours data from 40 speakers. Initial experiments are done utilizing hidden Markov models (HMM) as a visyllable classifier. Then, these models have been used for visual recognition of 6 Persian words with different numbers of syllables and an accuracy of 47.37% was obtained in a speaker-independent experiment.**

Keywords— ***Visual Speech Recognition, Lip Reading, CV syllables, Visyllable, Audio-Visual Database, Persian/Farsi Language.***

I. Introduction

Speech is a phonetic sequence that is organized according to specific patterns. Some organs of the body are used in the production of human voices. These organs include the lungs, trachea, larynx, pharynx, nasal cavity, palate, tongue, and lips. The variability of the shape and volume of the mouth distinguishes many of the phonetic characteristics of speech sounds. These changes are also shown in the shape of the lips. Lip-reading, the process of visual speech recognition, is the way of understanding speech by interpreting the movements of the speaker's lips. It is a very challenging task. In this process, by extracting the features of lip images and proper pattern recognition and machine learning methods, lip movements can be interpreted to the text of the speech. Extraction of these features can be implemented using image processing methods or by machine learning methods [1].

There are various applications for lip-reading systems. For example, they can help intelligence agencies utilizing a camera to track the contents of a conversation without having audio information. Another application is to use lip-reading instead of a keyboard to enter information into the computer [2]. Although the goal of the lip-reading system is to recognize words or sentences, so we need to model each word. If word models are used to distinguish a list of words, the number of models needed for training will increase to cover a large number of words. However, if we use smaller speech units such as phoneme or syllable, proper modeling of such units is required and the number of models required for training is greatly reduced [1]. According to [1], the lip-reading of long phrases is easier and more accurate than shorter ones. Also, as mentioned in [3] syllable-based lip-reading model has indicated good results in recognition of out of vocabulary words. Therefore, syllable-based modeling of speech units is proposed in this paper instead of phone-based units. As will be explained in the second section of this paper, Persian two-phoneme syllables can be a good choice for recognizing words or longer phrases. In order to recognize CV syllables, we need a database that contains Persian CV syllables and words. Existing Persian databases do not cover all the needs in this research field and their specifics do not provide a more comprehensive recognition. For this reason, we decided to provide a database containing Persian two-phoneme syllables and commonly used words that are composed of them by using more speakers in a non-studio environment with conditions close to everyday life. This database can be used in the lip-reading and speech processing studies of the Persian language, speech therapy, and talking face that needs these conditions.

The structure of this article is as follows: The second part includes a brief statistical review of Persian syllables and explains the reasons that motivated us to choose Persian two-phoneme syllables for this database. In the third section, some of the existing databases are introduced and their specifications are briefly described. In the fourth part, the database created in this research, the process of word selection and audio and video recording are described. In the fifth section, using some of the data in the database, models for recognizing Persian two-phoneme syllables are trained and their test results are presented. Also, these models are used to recognize some Persian words. In the last section, the article is summarized and concluded.

II. Statistical study of Persian two-phoneme syllables

Distinctive sounds produced by human speech organs are called phones. Phonemes are abstract units of language and have no meaning of their own, but by changing them they change the meaning of a word. Phonemes have two categories in general: consonant and vowel. The consonants indicated by the symbol C are sounds when they are uttered, the mouth, tongue, and lips are not necessarily uniform and there is usually obstruction in the passage of exhaust air. The vowels indicated by symbol V are sounds that the mouth, tongue, and lips are perfectly uniform and there is no obstruction to the exit of air. In Persian, only consonants can be placed at the beginning of a word. The vowels are the core of the syllable and never appear at the beginning of the word [4].

978-1-6654-3365-5/21/$31.00 ©2021 IEEE

In general, there are 29 phonemes in Persian, which include 6 vowels and 23 consonants [5]. A syllable in Persian is a continuous phonetic string consisting of a vowel and one to three consonants. Continuous phonetic string means that the components of a syllable are produced in a production process without pause. Persian syllables have three serial combinations: CV, CVC, and CVCC. The syllables, like phonemes, affect the movements and appearance of the lips [4]. Corresponding to phonemes in speech, visemes are considered as small units in the field of visual speech. The viseme is the appearance of the lips when pronouncing a phoneme and it is the smallest visual unit of speech [6]. Correspondingly, the appearance of the lips when pronouncing a syllable is called visyllable and can be considered as a visual unit of speech [7]. The recognition of visyllables in the word structure can be used for visual speech recognition.

In the study of Persian syllables, it has been determined that 57.75% of all syllables in Persian have a CV syllable structure and the frequency of CVC and CVCC syllables is 38.09% and 4.16%, respectively. Also, in many cases, when pronouncing Persian words, the final consonant of CVCC syllables is removed from the two-consonant parts CVCC or CVC is converted to CV syllable. For example, kargar as a CVC+CVC word can be converted to karegar containing two CV syllables and one CVC syllable. All this evidence shows that the syllable CV is a frequently used syllable in the Persian language [8]. Also, in our studies, it was found that 2946 unique words of BijanKhan corpus [9] only composed of CV syllables. Due to the mentioned statistical facts and the great importance of CV syllables in Persian, it was decided that the recognition of these syllables as widely used syllables in Persian speech should be considered and for this purpose a database appropriate to this need should be prepared.

III. Overview of existing audio and visual databases

Many audio and visual databases have been created for various applications in different languages. Each database is designed for applications intended by its creators. The difference in the specifics of these datasets is in the number, gender, language, and age range of speakers. They also differ in the quality and quantity of content, the recording environment conditions, and the recording quality of the dataset. Some datasets contain only numbers, letters, some words, sentences, while others consist of combinations.

AVletters [10] is an audio-visual database of the English alphabet expressed by 5 men and 5 women. AVICAR audio-visual database [11] is considered as an instance of a dataset recorded in a non-studio environment. The videos of this database were recorded inside a moving car and 50 men and 50 women expressed English letters, isolated digits, connected digits e.g. phone numbers, and some English sentences. The CUAVE audio-visual database [12] is designed to improve research in two important areas: audio-visual speech recognition that is resistant to the speaker's facial movements, and simultaneous speech recognition of different speakers. This database contains isolated and connected digits expressed by 19 men and 17 women. The database included speaker movement during recording and the simultaneous speech by two speakers. Moreover, the GRID audio-visual database [13] contains 1000 sentences uttered by 34 speakers (18 men and 16 women).

So far, 3 databases have been presented in audio-visual research of Persian speech. AVA audio-visual database [14] was suitable for speech therapy and people who want to learn Persian and classify Persian visemes. It contains numbers and all expressions in the form of syllables: CV, VC, CVC, VCV, CVCC and vowels in the language. Also, 20 common sentences from Sara's lip-reading test in which the number of phonemes is balanced have been included. Two female speakers were used and filmed in a professional studio.

The AVA II audio-visual database [15], which was built for lip-reading applications, used AVA data while the number of speakers increased to 7 females and 7 males. It was filmed in a professional studio with 3 cameras from different angles. SFAVD audio-visual database [16] was designed with the aim of producing a talking face, which includes 600 Persian sentences and each sentence consists of 5 to 20 words. Linguistic rules and how they affect the appearance of the lips were also considered. These words are selected in four steps from the PEYKARE and FARSDAT databases, and the selected sentences cover all commonly used words, all phones, diaphones, and common syllables. A male speaker has been filmed, shining light on his face to reduce shadows in the lip area. In addition to lip-reading, the database can be used for other applications as well. It should be noted that the introduced Persian datasets are not publicly available.

IV. Introduce the proposed audio-visual database

As previously mentioned, a Persian audio-visual dataset including CV syllables has not been available to cover all of the visyllables and the related words. Therefore it was decided to build a comprehensive audio-visual database, considering linguistic rules and appropriate to the aim of machine learning approaches in the research field of audio-visual speech recognition. Also, the database was recorded with a larger number of speakers in a non-studio environment, so that this data can be used in applications with natural conditions close to everyday life.

A. Database Content

This database consists of three parts: The first part contains various forms of CV syllables that are obtained from all Persian consonants and Persian vowels, which is a total of 138 combinations. In the second part, 52 Persian words have been selected. The choice of the words was based on the famous Persian corpora: Bijankhan and Faranet. The Faranet corpus contains 55,700 words with their pronunciation in phonetic symbols. Initially, using the phonetic part of the corpus, the multi-part words were broken into one part and 85,787 words were obtained, then by removing the duplicate words, 17,000 words remain unique. Then words consisting of consonant and vowels (CV) syllables were extracted and 4685 words were obtained. In order to make it possible to use these data more widely in the future for similar applications, we tried to select words based on the importance and amount of use in Persian.

In the next step, the frequency of CV words in the BijanKhan corpus was the criterion for selection. Of these 4,685 words, 2,946 have at least one repetition in BijanKhan. In addition, the minimum number of words covering all CV syllables was selected, which was 52 words. In the third part, from the remaining words (2,946-52), the most frequent words are selected which brings the number of syllable repetitions in the dataset to at least 10. Thus 577 words were selected. From these 577 words, 500 more frequent words were selected and divided into 10 groups of 50 words. These words form the third part of the data set. The first part (138 CV syllables) and the second part (52 CV based-words) of the dataset are expressed by all speakers and from the third part, every 50 words are expressed by one of 10 groups (each group consisting of 2 males and 2 females).

B. Database recording specification

This database was recorded in the environment with natural daylight at the time between 9 to 12 AM. From the windows in front of the speakers in the recording room, sunlight was shining into the recording environment. A green curtain hangs in the background of the speaker. The camera was placed at a distance of 93 cm from the green screen and at a distance of 63 cm from the speaker so that the speaker's shoulders and face were in camera framing. Video recording was performed using the SONY HDR-PJ410 HANDYCAM camera at a rate of 50 frames per second with a resolution of 1920 x 1080 pixels. In addition, the original sampling rate used for the speech/voice recording was 16kHz. There is also natural ambient sound in this database. There was a screen in front of each speaker who viewed and said the words on the screen in order by clicking on the computer mouse. Examples of recording conditions are shown in Figure 1.

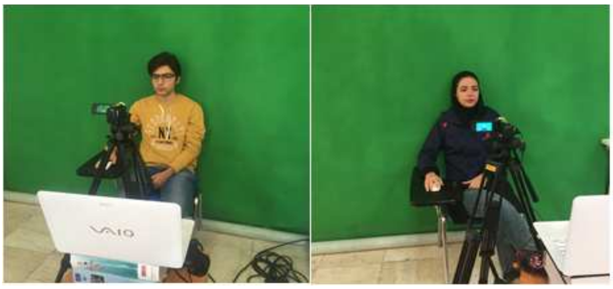

Figure 1: Screenshots of database recording conditions

C. Statistical specifications

The speakers are 20 males and 20 females, from 20 to 40 years old, some of whom are wearing glasses (Women with light makeup and men without beards and mustaches). The speakers utter each of the three parts of the data from beginning to end and repeat this 4 times. The speaker closes his/her mouth as a silence point before uttering the next word. Each set of words was recorded continuously for each speaker. Thus, each data was recorded 4 times at intervals. The total recording time for each speaker was a maximum of 45 minutes and about 30 hours of recording for all speakers in total.

D. Extract data from video

Using Avidemux 2.7 software by a human agent, continuously recorded videos from people were cut into smaller videos that contain a syllable or a word. In this database, each speaker uttered 138 CV syllables and 102 words (52+50), so a total of 38,400 (4*40*240) videos are saved. Each of these extracted videos is saved as an avi format video file. The name of each file includes the name of the syllable/word, the number assigned to the respective speaker, and the order in which the syllable or word is repeated by the speaker. For example, the name of the video file bi_10_3 indicates that it is the third iteration of the CV syllable 'bi' by speaker number 10. Data samples that have been uttered by one of speakers are available[1].

V. Using the database to recognize syllables and words

To recognize CV syllables based on the method mentioned in [17], a vector of geometric features such as height, width, and area of the lips, which are calculated in pixels, has been used. First, the video of each syllable is converted to a sequence of frames, and then, using the Viola-Jones algorithm [18], the lip segments in these frames are identified. Examples of lip images are shown in Figure 2.

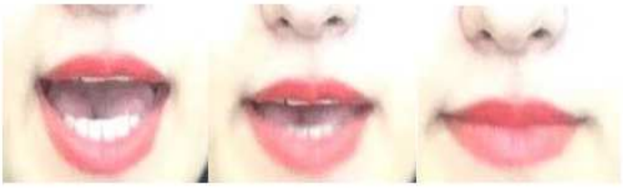

Figure 2: Images of the lip area

Next step is to separate the lip segment in the image. For this purpose, a combination of components a and b of the Lab color space, component H of the Psudo-Hue color space, and component U of the LUX color space are used as an input to the SFCM algorithm [19]. The largest white area in the black and white image resulting from the output of the SFCM algorithm is also selected as the lip segment. Examples of segmented output images are shown in Figure 3:

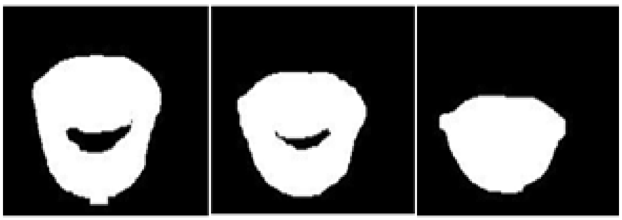

Figure 3: Images of the lip segments

After separating the lip segments, it is time to extract the desired features. By counting the maximum number of white pixels along the vertical and horizontal axes, the height and width features are extracted. Similarly, by counting all the white pixels, the area feature of the lip segment is extracted.

Experiments to recognize CV syllables were performed by hidden Markov models (HMMs). This is done using HTK software and its tools in the training and testing process. For example, by initializing the model parameters using the HINIT function and determining the number of states as 3 or 5 and the number of Gaussian mixture models as 1 according to the baseline article [17], the HREST and HEREST functions are then used to retrain the model. Features and labels of each video are read and utilized to create the HMM model for any visyllable. In the test phase, the HVITE function is used and the test is performed using a grammar. The test results are then presented based on the HRESULT function.

Samples of 6 people were used for the test and the rest of the samples of other people were used for training and validation as a speaker independent framework. The obtained accuracy in 138 classes and 3-state model is equal

[1] https://github.com/mahsahed/PAVID-CVs

to 3.04% and in 5-state model is equal to 4.06%. The reason for this low accuracy is due to the uniform shape of some syllables, here named visyllables, and it is necessary to make another classification to put syllables that have similar shapes in the same class. For this purpose, clustering of the visyllables was done by a man who was a completely unhearing person and all of Persian CV visyllable were grouped into 9 clusters [20]. These clusters were used as a new data label in the training step. At this stage, accuracy in 9-class mode utilizing HMM of 3-states was equal to 39.92% and in 5-state model was equal to 44.94%.

The combination of visyllable models and 2-gram language models based on visyllables is used in visual speech recognition of Persian words. For this purpose, six different words were selected from the database; each of them represents the words including two, three, four, five, six, or seven syllables. By constructing a language model of visyllables based on these words and using the visyllable models, a speaker independent experiment is done to recognize isolated words and an accuracy of 47.37% is obtained for recognition of the words.

VI. Conclusion and future work

In this paper, the audio-visual database of Persian two-phoneme syllables was described. The database contained audios and videos of speakers who had expressed the Persian CV syllables and the common words that contain these syllables. Recording conditions were considered close to the daily living environment. An example use of this database, building hidden Markov models for Persian CV syllables, and utilizing them to recognize Persian words is presented in this article. Although this database was created for the development of Persian lip-reading, it can also be used in other applications of audio and visual speech processing. It can also be used in speech therapy, teaching people who suffering hearing weakness, and make talking face. More speakers than existing databases and non-studio environments make this database suitable for researches leading to the development of commercial products.

References

[1] K. Thangthai and R. W. Harvey, "Building Large-vocabulary Speaker-independent Lipreading Systems.," in *INTERSPEECH*, 2018, pp. 2648–2652.

[2] N. Akhter and A. Chakrabarty, "A Survey-based Study on Lip Segmentation Techniques for Lip Reading Applications," 2016.

[3] A. Kurniawan and S. Suyanto, "Syllable-Based Indonesian Lip Reading Model," in 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1–6.

[4] Y. Samareh, "Phonology of Farsi Language," Markaze Nashre Daneshgahi publisher, Tehran, 1986. (in Persian)

[5] M.M. Homayounpour, "Text to Speech Research", E-Book, 2011. (in Persian)

[6] A. Bastanfard, M. Aghaahmadi, A. A. kelishami, M. Fazel, and M. Moghadam, "Persian Viseme Classification for Developing Visual Speech Training Application," in *Advances in Multimedia Information Processing - PCM 2009*, vol. 5879, P. Muneesawang, F. Wu, I. Kumazawa, A. Roeksabutr, M. Liao, and X. Tang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1080–1085.

[7] Y. Pei and H. Zha, "Stylized synthesis of facial speech motions," *Computer Animation and Virtual Worlds*, vol. 18, no. 4–5, pp. 517–526, 2007.

[8] M.Eslami and S.Rhimi,"Persian phonetic system in the mirror of statistics", Languag and Linguistics, vol. 9, no. 18, pp. 65–90, 2013. (in Persian)

[9] M. Bijankhan, J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi, "Lessons from building a Persian written corpus: Peykare," *Language resources and evaluation*, vol. 45, no. 2, pp. 143–164, 2011.

[10] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, Feb. 2002, doi: 10.1109/34.982900.

[11] B. Lee *et al.*, "AVICAR: Audio-visual speech corpus in a car environment," 2004.

[12] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Orlando, FL, USA, May 2002, p. II-2017-II–2020.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[14] A. Bastanfard, A. A. Kelishami, M. Fazel, and M. Aghaahmadi, "A comprehensive audio-visual corpus for teaching sound persian phoneme articulation," in *2009 IEEE International Conference on Systems, Man and Cybernetics*, 2009, pp. 169–174.

[15] A. Bastanfard, M. Fazel, A. A. Kelishami, and M. Aghaahmadi, "The Persian linguistic based audio-visual data corpus, AVA II, considering coarticulation," in *International Conference on Multimedia Modeling*, 2010, pp. 284–294.

[16] Z. Naraghi and M. Jamzad, "SFAVD: Sharif Farsi audio visual database," in *The 5th Conference on Information and Knowledge Technology*, 2013, pp. 417–421.

[17] S. S. Morade and S. Patnaik, "A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition," *optik*, vol. 125, no. 18, pp. 5181–5186, 2014.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, 2001, vol. 1, p. I-511-I-518, doi: 10.1109/CVPR.2001.990517.

[19] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *computerized medical imaging and graphics*, vol. 30, no. 1, pp. 9–15, 2006.

[20] M. Hedayatipour, Y. Shekofteh, and M. Ebrahimi Moghadam. 'Clustering of Persian CV Visyllables for Lipreading Application." The 11th conference on Information and Knowledge Thechnology, 2020. (in Persian)