Creating Simplified Version of Lip Database based on Front View of Face

Ritesh A. Magre
Dept.of CS & IT
Dr. B.A.M.U. Aurangabad

Ajit S. Ghodke
Dept. of MCA, Sinhagad Institute. SIBACA
Lonavala, University of Pune,
Pune, India

ABSTRACT
Recently lot of work has been done on audio visual speech recognition but less work has been done on visual speech and speaker recognition. This research belongs to human computer interaction (HCI) domain. HCI makes human computer interaction simple. This paper represents the creating of database of visual speech and speaker in English language and preprocessing of it to improve recognition accuracy. We have studied Tulipse1 database, AV Database and CUAVE Database on the basis of these different databases we have created our own database. This is useful for all researchers those are working HCI domain particularly Visual Speech and Speaker Recognition.

Keywords
Speech, Visual Speech, Lip Reading, Lip Database, Visual Speech Recognition, Speaker Recognition, Face detection, Lip Cropping .

1. INTRODUCTION
Visual Speech: Visual speech is defined as a human eye catchable expression of a speaking human face.

Visual Speech Recognition: The recognition of speech from the visual information only is called as visual speech recognition.

Speaker Recognition: Speaker recognition is nothing but identifying speaker by machine or recognizing who is speaking.

Visual Speech database has been created from several years and shared to the researchers for further development. Most of the databases are created in a standard environment and videos are recorded using high definition camera. In this research we have recorded the videos by mobile phone camera with little standards.

The Visual Speech Recognition has attracted researchers from several years. it has developed through the recognition of the isolated words from one speaker to speaker-independent continuous speech. Now a day the recognition rate of system has improved in limited condition. The environment and the type of speech influence the recognition rate. By the environment it is especially meant environmental noise that degrades acoustic signal. In this case the visual part of the speech can be used to increase recognition rate. Such a system is called audio-visual speech recognition (AVSR). This paper deals with design, recording and preprocessing of the audio-visual speech database [1].

In general, the use of visual features jointly with the acoustic information is becoming increasingly important as a technique to improve the speech recognition robustness [2], [3].

2. RELATED WORK
Some efforts have already been taken in the creation of audio visual database. Tulips1 is a twelve subject database of the first four English digits recorded in 8-bit grayscale at 100x75 resolutions [4]. AVLetters includes the English alphabet recorded three times by ten talkers in 25 fps grayscale [5]. DAVID is a larger database including various recordings of thirty-one speakers over five sessions including digits, alphabets, vowel-consonant-vowel syllable utterances, and some video conference commands distributed on multiple SVHS tapes [6]. It is recorded in color and has some lip highlighting. AV Database (Audio-Visual Database by IBM) consist of full-face frontal video and audio of 290 subjects, uttering ViaVoiceTM training scripts that is continuous read speech with mostly erbalized punctuation (dictation style) and vocabulary size of approximately 10,500 words. Transcriptions of all 24,325 database utterances, as well as a pronunciation dictionary are provided. The database video is size 704x480 pixels, interlaced, captured in color at rate of 30 Hz (that is 60 fields per second are available at a resolution of 240 lines) and it is MPEG2 encoded at the relatively high compression ratio about 50:1. High quality wideband audio is synchronously collected with the video at a rate of 16 kHz and at a relatively clean audio environment (quite office, with some background computer noise). The duration of entire database is approximately 50 hrs. it is mentioned that to this date this is the largest audio visual database collected and it constitutes the only one suitable task of continuous large vocabulary, speaker independent audio-visual speech recognition, as all other existing databases are limited to small number of subjects and /or small vocabulary tasks. [7], [8], [9], [10], [11], [12], [13], [14], [15]. CUAVE database project has been initiated at Clemson University, for audio-visual experiments. The main interest is audio-visual processing where visual technique such as lip reading is combined with more traditional audio methods. The multimodal speech processing can be much more robust to noise and has other benefits as well.

This was beneficial to researchers working in this area. CUAVE is a speech corpus of over 7000 utterances of continuous, connected and isolated digits. It includes both individual speakers and speaker pairs. The speech is fully labeled and all files are available on DVD. There are 36 individuals in the database that is 17 female and 19 male speakers. All these speakers have different accents, skin, tone, facial hairs, hats and glasses. There are also 20 groups of speakers that may be used for multi-speaker research. The video is compressed at 5000 kbps at MPEG-2 encoding. Audio is included with 44 kHz stereo and 16 kHz mono. The speakers are recorded either moving or standing still while saying connected and continuous digit sequences [16].

3. DATABASE DEVELOPMENT

In this we have recorded videos of 10 subjects which include male and female with different ages. Each subject has spoken 5 sentences from 5 different areas such as College, Government Office, Hospital, House and Restaurant and each sentence is repeated 5 times. We have selected 5 frequently used sentences from each area as shown in table 1. All the sentences are spoken in English language. The total sentences recorded in a database are 1250.

The format of the videos recorded is mp4 and all these videos are recorded by Samsung galaxy nxt mobile phone 13 megapixel camera by hand. The distance of the video recording is 4 feet.

Table 1. List of sentences recorded

College	
1	Who is the in charge
2	You are not regular
3	There is a meeting today
4	What are the courses
5	I am your teacher
Government Office	
1	Where is head of department
2	How long will it take
3	The process is on
4	Please sign here
5	My work is pending
Hospital	
1	May I help you
2	Thank you doctor
3	Please go outside
4	Please maintain silence
5	You will be fine
House	
1	Please bring the tea
2	Take a bath
3	I am watching TV
4	Open the door
5	Where is the news paper
Restaurant	
1	What would you like to order
2	Please come here
3	How can I help you
4	Bring me the menu card
5	Kindly clean the table

3.1 Database preprocessing

The whole audio-visual database is preprocessed and organized. The sample of the database is shown in figure 1 and figure 2.

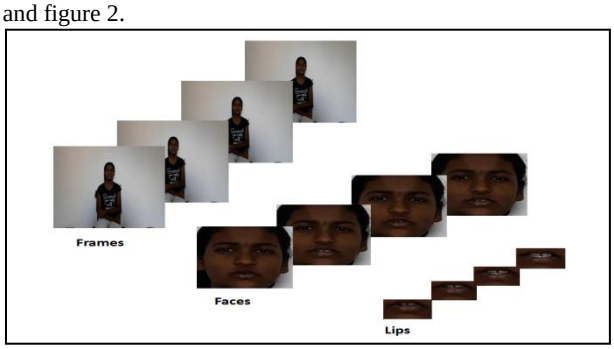

Fig.1 Organised collection of frames, faces and lips.

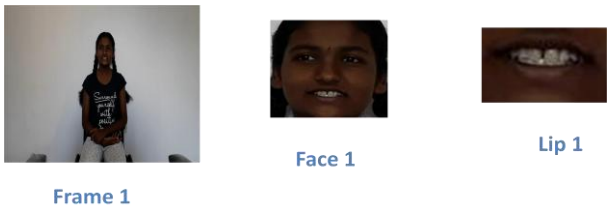

Fig.2 Structure of the database.

3.1.1 Frame Extraction

The frames are extracted from the recorded videos. The number of frames generated varied based on the size of videos and style of speakers for the same sentence spoken. The required frames are selected for further processing.

3.1.2 Face Detection and Cropping

The face detection task is carried out on number of frames extracted from video using viola jones algorithm. While detecting face the merge threshold has to be increased to reduce missed face detection. The detected faces has cropped and stored separately for supportive features required for speech and speaker recognition.

3.1.3 Lip Detection and Cropping

The Lip detection task is carried out on cropped face images which are stored separately. The lips are also detected using viola jones algorithm by adjusting merge threshold to reduce missed lip detection. The detected lips are cropped and stored separately.

3.1.4 Noise Removal

The noise removal required to improve the quality of image and to get the more required information from the image. In this we have removed the noise from the extracted lip frames. The median filter is used to smoothen the extracted lip frames and high pass filter is used to sharpen the lip frame images.

4. CONCLUSION

In this paper the suitable images are selected from the set of images for archiving better results. a new database is created to support the development of visual speech and speaker recognition system. The concentration is on the lip and it is valuable for visual speech parameterization. In the given experiment for some subjects there is a error in the detection of lip. In the future we will try to improve the overall lip detection rate using other more robust techniques.

5. ACKNOWLEDGEMENTS

The 10 subjects including male, female has given their valuable time for video recording. Each subject spoke 125 sentences.

6. REFERENCES

[1]. Jana Trojanova´, Marek Hru´ z, Pavel Campr, Milosˇ Zˇ elezny "Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition" Department of Cybernetics, Faculty of Applied Sciences, University OF West Bhoemia Univerzitni
22, 306 14, Plzen, Czech Republic.

[2]. Paterson, E. K.: Audio Visual Speech Recognition for Difficult Environments. Ph.D. thesis, Clemson University(2002).

[3]. Weber, K., Ikbal, S., Bengio, S., and Bourlard, H.: Robust Speech Recognition and Feature Extraction Using HMM2. Computer Speech & Language, 17 (2003) 2–3.

[4] J.R. Movellan, "Visual speech recognition with stochastic networks," in Advances in Neutral Information Processing Systems, G. Tesauro, D. Toruetzky, and T. Leen, Eds., vol. 7. MIT Press, Cambridge, 1995.

[5] I. Matthews, Features for Audio-Visual Speech Recognition, Ph.D. thesis, School of Information Systems, University of East Anglia, UK, 1998.

[6] C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston, "Design issues for a digital audiovisual integrated database," in IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, Savoy Place, London, Nov. 1996, number 1996/213, pp. 7/1–7/7.

[7] A Adjoudani and C Benoit, On the integration of auditory and visual parameters in an HMM-based ASR, In Stork and Henneke |11|, pages 461-471 .

[8] C Bergler, H Hild , S Manke and A Waibel, Improving connected letter recognition by lipreading, In Proc International Conference on Acoustics, Speech and Signal Processing, volume 1, pages 557-560, Minneapolis, 1993.

[9] M T Chan, Y Zhang and T S Huang, Real time lip tracking and bimodal continuous speech recognition, In Proc IEEE 2nd workshop on multimedia signal processing, pages 65-70, Redondo Beach, 1998.

[10] C C Chibelushi, F Deravi and J S D Mason, Survey of audio-visual speech database, technical report, Department of electrical and electronic engineering, University of Wales, Swansea, 1996.

[11] I Matthews, T Cootes, S Cox, R Harvey and J A Bangham, Lipreading using shape, shading and scale, In Proceedings of Workshop on Audio visual speech processing , pages 73-78, Terrigal 1998.

[12] K Messar, J Matas, J Kittler, J Luettin and G Maitre, XM2VTS: The extended M2VTS database, In Proc. 2nd International conference on audio and video based biometric person authentication (AVBPA) page 72-76, Washington 1999.

[13] J R Movellan and G Chadderdon, Channel seperability in audio visual integration of speech: A Bayesian approach, In Stork and Hennecke |11|, pages 473-487.

[14] E D Petjan, Automatic Lipreading to enhance speech recognition, In Proc Global Telecommunication Conference (GLOBCOM), pages 265-272, Atlanta 1984.

[15] P Teissier, J Robert-Ribes and J L Schwartz, Comparing models for audio visual fusion in noisy vowel recognition tasks, IEEE transaction on speech and audio processing, 7(6): 629-642, 1999.

[16] CUAVE Database, Clemson university database for Audio visual experiments http://www.ece.clemson.edu/speech/cuave.htm.