‘vVISWa’ – A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction

Prashant Borde
Vision and Intelligent System Lab
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India

Ramesh Manza
Biomedical Image Processing Lab
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India

Bharti Gawali
SpeechCommunication and Machine Research Lab Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India

Pravin Yannawar
Vision and Intelligent System Lab
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India

ABSTRACT
Automatic Speech Recognition (ASR) by machine is an attractive research topic in signal processing domain and has attracted many researchers to contribute in this area of signal processing and pattern recognition. In recent year, there have been many advances in automatic speech reading system with the inclusion of audio and visual speech features to recognize words under noisy conditions. The objective of audio-visual speech recognition system is to improve recognition accuracy. In order to develop robust AVSR systems under Human Computer Interaction an appropriate simultaneously recorded speech and video data are needed. This paper describes a ‘vVISWa’ (*Visual Vocabulary of Independent Standard Words*) database consists of audio visual data of 48 native speakers and 10 nonnative speakers. These speakers have contributed towards development of corpus in three profiles that is *full frontal*, 45^0 *profile* and *side pose*. This database was primarily designed to deal with Multi-pose Audio Visual Speech Recognition system for three languages that is, ‘Marathi’ (*The Native language of Maharashtra*), ‘Hindi’ (*National Language of India*) and ‘English’ (*Universal language*). This database is multi-pose, multi-lingual database formed in Indian context. This database available by request from http://visbamu.in/viswaDataset.html.

Keywords
Automatic Speech Recognition (ASR), Visual Speech Reading (VSR), Multi-pose Audio Visual Speech Recognition (AVSR) and ‘vVISWa’.

1. INTRODUCTION
Speech is a simplest and natural means of expressing ourselves and to communicate with others. In the recent year, many researcher have been attracted towards design of robust methods for face detection, lip tracking for Audio-Visual Speech Recognition, Bimodal Speaker Recognition and Speaker localization so on. Researchers have adopted speech as medium to interact with machines in-light of natural means of interacting with humans. Human Computer Interface has become potential research domain to build robust speech recognition system which control’s machine activities and processes. Advances in the fields of image processing, computer vision and pattern recognition has led foundation for advance research in automatic lip reading systems. *Visual speech recognition* (VSR) or speech reading could open the door for many novel applications where in absence of audio the speech could be recognized. VSR can be used in many applications such as speaker recognition, human-computer interaction (HCI), sign language recognition, audio-visual speech recognition (AVSR), and security purposes when audio is not available or audible. VSR aims to recognize spoken word(s) by using only the visual information that is produced during speech. The use of visual features in AVSR is motivated by the bimodality of the speech formation and the ability of humans to better distinguish spoken sounds when both audio and video are available. Speech perception is a multimodal process which involves information not only from audio but from visual [1]. Although visual information is insufficient to distinguish languages completely, when combined with audio, it could significantly improve the performance of speech recognition [2]. When there are a limited number of utterances to be identified, it is possible to use visual information only to do speech recognition. Visual information is an important alternative to traditional speech recognition technology in human-machine interactions, especially when audio is unavailable or seriously corrupted by background noise.

This paper addresses towards design of a multi pose AVSR database in Indian context. The work reflecting towards database of isolated words (*Numerals*, *Color*, *Months*, *Fruit Names* and *frequently used dictionary words*) of Marathi, Hindi and English language. The content of this paper is organized as, Section II addresses the *AVSR* dataset, Section III describe with *‘vVISWa’* corpus, section IV report the database parameters considered at the time of design, section V is conclusion of work followed acknowledgement and references.

2. RELATED WORK ON AVUDIO VISUAL SPEECH DATABASE
The work on audio-visual speech recognition system have been initiated with an objective to design robust speech recognition systems using two basic modalities that is ‘acoustic’ and ‘visual’ signal. Various researchers have contributed in this domain and contribution was also be seen for AVSR database. However very few attempts were made in this as far as Indian context is concern, this research work was aimed towards formation of multi-pose audio-visual database of English, Hindi and Marathi continuous and isolated words.

It was seen that many researchers have contributed in the area of audio-visual database for audio-visual speech recognition system, some of most promising work have been summarized as Lee et al. [3] introduced the *Audio-Visual speech In a Car* (AVICAR) database. It was recorded in a moving car. The framework employed four cameras in a lateral array on the dashboard of car for video recording purpose which resulted in four synchronized video streams with different views. Due to the limited space in the car, the angles between the views relative to the speaker were modest and the actual degrees

unknown. There were 100 speakers out of which 50 were male and 50 were female actively involved in the recording. Out of 100 speaker, data of 86 speakers were available for researchers to download and carry out their experiments on the same. The acquisition framework handles 5 noise conditions which was set up during recording. Under each condition, each speaker was asked to first speak isolated digits and letters twice. It was followed by 20 phone numbers with 10 digits each and 20 sentences randomly chosen out of 450 TIMIT sentences [4]. Video was recorded at 30 fps with a resolution of 720 × 480 pixels and audio sampled at 16 kHz, 16-bit resolution.

The *AVLetters* database [5] consists of 10 speakers (5 male and 5 female) uttering isolated letters A-Z. Each letter was repeated three times by each speaker during recording. Video was recorded at 25 fps with a resolution of 376 × 288 pixels and audio at 22.5 kHz with a 16-bit resolution. Image data were processed at 80 × 60 size, full-face region was cropped manually by locating center of the mouth in the middle frame of each utterance. Cox et al. [6] collected a higher definition version of the *AVLetter* database, named '*AVLetter2*'. The corpus includes 5 speakers uttering 26 isolated letters seven times. Video was recorded at 50 fps with a resolution of 1920 × 1080 pixels and audio as 16-bit 48 kHz mono.

Hazen et al. [7] produced the *AV-TIMIT* database for their studies of speaker-independent AV-ASR. The corpus contains 4 hr. of AV data collected from 233 speakers (117 male and 106 female). The spoken utterances were chosen from the phonetically balanced TIMIT sentences [4]. Each speaker was asked to read 20 sentences and each sentence read by at least 9 different speakers. The sentence, were chosen so that it was uttered by all the speakers. Video was recorded at 30 fps with a resolution of 720 × 480 pixels and audio sampled at 16 kHz. Patterson et al. [8] recorded the *Clemson University Audio-Visual Experiments* (CUAVE) database that included speaker movement and simultaneous speech from multiple speakers. It consists of two major sections. In the first section there were 36 speakers (17 male and 19 female) involved in the recording. Each speaker was asked to utter 50 isolated digits while standing naturally and another 30 isolated digits while moving side-to-side, back-and-forth, or tilting the head. After that, the speaker was framed from both profile views while uttering 20 isolated digits. Each individual then uttered 60 connected digits while facing the camera again. The second section of the database includes 20 pairs of speakers. For each pair, one speaker was asked to utter a connected-digit sequence, followed by the other speaker and vice versa a second time. For the third time, both speakers uttered their own digit sequences simultaneously. Video was recorded at 30 fps with a resolution of 720 × 480 pixels and audio at 16-bit, mono rate of 16 kHz. The data was fully labeled manually at the millisecond level.

The *IBMSmart-Room* (IBMSR) Database [9] was collected as part of the European project, CHIL. The corpus consists of 38 speakers uttering continuous digit strings. There were two microphones and three cameras used for AV data collection. The cameras were set to frame the speaker from the frontal and both profile views. Video was recorded at 30 fps with a resolution of 368 × 240 pixels and audio at 22 kHz. There were total 1661 utterances included in the corpus. The *Language Independent Lip-Reading* (LILiR) database [10] collected at the University of Surrey consists of 20 speakers uttering 200 sentences from the Resource Management Corpus [11]. The speaker was framed by two HD cameras from the front and profile views and by three SD cameras placed at 30°, 45° and 60°. It is unknown about the video and audio quality. The *MOBIO* database [12] was designed for evaluating face and speaker authentication algorithms on mobile phones. Videos were recorded from a mobile phone held by speakers. Consequently, the microphone and video camera were no longer fixed and were used in an interactive and uncontrolled manner. There are in total 152 speakers each of whom had multiple sessions of video recording. They were asked short-response questions, free-speech questions and to read predefined texts. Video was recorded at 16 fps with a resolution of 640 × 480 pixels.

Zhao et al. [13] recorded, *OuluVS* database for visual-only ASR. It consists of 10 daily-use English phrases uttered by 20 speakers (17 male and 3 female). Each utterance was repeated by a speaker up to nine times. Video was recorded at 25 fps with a resolution of 720 × 576 pixels. The *Grid AV* corpus was collected by Cooke et al. [14]. There are 34 speakers (18 male and 16 female) involved in its recording. There are multiple words that could be chosen at each position, resulting 1000 sentences per speaker in total. Video was recorded at 25 fps with a resolution of 720 × 576 pixels (a lower quality version (360 × 288) also available) and audio down-sampled to 25 kHz with the peak SNR varying across speakers from 44 to 58 dB. During recording, subjects were asked to speak sufficiently quickly to fit each sentence into a 3-second time window. The *XM2VTSDB* database [15] was collected at the University of Surrey for personal identification. There were 295 subjects involved and the recording consisted of four sessions. In each section, each subject was asked to speak two continuous digit strings and one phonetically balanced sentence. The utterances remained the same in all the four sections. Table 1 shows, summary of reputed audio-visual databases reported in literature.

Table 1: Reported Audio-Visual Databases

Database Name	Author/Agency	Speakers	Corpus
Audio-Visual speech In a Car (AVICAR)	Lee et al. [3]	100 (50 Male, 50 Female)	20 phone numbers with 10 digits each, 20 sentences randomly chosen out of 450 TIMIT sentences for each speaker
AVLetters	[5]	10 (5 Male, 5 Female)	A-Z Letters
AVLetter2	Cox et al. [6]	5	26 isolated letter seven times
AV-TIMIT	Hazen et al. [7]	233 (117 Male, 106 Female)	Utterances from phonetically balanced TIMIT sentences
Clemson University Audio-Visual Experiments	Patterson, et.al [8]	36 (17 Male, 19 female)	Isolated and Continuous digits with head movement in side-to-

(CUAVE)			side, back-forth and tilting.
IBMSmart-Room	European project [9]	38	Continuous digit string
Language Independent Lip-Reading (LILiR)	University of Surrey [10]	20	Resource management Corpus
MOBIO	Idiap Research Institute [12]	152 (100 Male, 52 Female)	short-response questions, free-speech questions
OuluVS	Zhao et al. [13]	20 (17 Male, 3 Female)	10 daily use English phrases per speaker with nine repetition each
Grid AV	Cooke et al. [14]	34 (18 Male, 16 Female)	1000 Sentences per speaker
XM2VTSDB	University of Surrey [15]	295	Continuous Digits

3. 'vVISWa' DATABASE

The database entitled 'vVISWa' is abbreviation for *Visual Vocabulary of Independent Standard Words.* This database was aimed to provide an audio-visual corpus of isolated as well as continuous words uttered by native and non-native speakers in Indian context. Unlike [3][5][6][7][8][9][10][12] [13][14][15] uni-language database, 'vVISWa' database was primarily designed to deal with Multi-pose Audio Visual Speech Recognition system for three languages that is, 'Marathi' (*The Native language of Maharashtra*), 'Hindi' (*National Language of India*) and 'English' (*Universal language*).

3.1 Experimental Setup

The entire database was acquired using Sony PJ660, High definition (HDR) digital cam-coder. The figure 1 shows top view placement multiple cameras at acquisition process. There are three cameras used for acquiring simultaneous input corresponding to utterance of speaker. These cameras acquiring visual recording of speaker in close-open-close mode. The speakers/subjects were instructed to utter continuous isolated words in Marathi, Hindi, and English language. The set of standard dictionary words of language were considered for the same. The visual utterance was recorded at resolution of 720 x 576 in (*.avi) format at 25fps frame rate in three angle comprising full frontal using camera 'C1', 45^o face using camera 'C2' and side pose using camera 'C3'. The distance between the speaker and the camera was 1 meter. Two fluorescent lamps was used, the light come from these lamps was filtered by white curtains and a dark gray background was used to avid the reflection of light. The video sequences used for acquiring the database was collected in the laboratory in a closed environment.

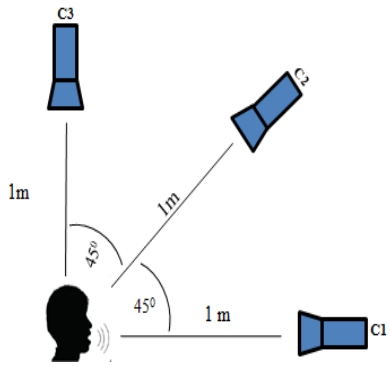

Fig. 1 Acquisition of utterances

3.2 Recording Process

At recording, subjects were asked to sit on a chair and get comfortable with the environment of data acquisition room. Speaker was instructed about complete acquisition process at the beginning. Every speaker have to adjust or get comfortable seating so as to find a central position relative to the cameras that would ensure similar angles of the camera views. This process helps in comprised finding in optimum position to seat, right height of the chair and an un-rotated head position of the speaker. Each speaker was asked to read and utter isolated word presented on a visual prompter (screen) using power point presentation. Usually time window of 2s were considered for recording of each utterance, each utterance were uttered for ten repetitions. The arrangement of acquisition setup was presented in of Figure 2.

Fig. 2 shows the recording environment of 'vVISWa' Dataset.

4. DATA ACQUISITION AND CORPUS

The 'vVISWa' audio visual speech database was developed at Vision and Intelligent System Lab (VIS) of Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University. This corpus is recorded with readings of large set of English, Marathi and Hindi Isolated words. Total 58 speakers have contributed in the formation of corpus out of which 48 speakers were native and 10 speakers were non-native that is, they are from Iraq and Yemen (*foreign, non-native and Arabic language was their first language of communication*). There were 20 female and 28 Male speakers in native speakers set and all non-native speakers were male. The database it-self carries visual complexity such as out of 28 male speakers four speakers were wearing glasses, some were wearing cap etc. Visual profile obtained in the acquisition is as shown in figure 3. A continuous video recording was made of each subject, rather than a few snapshots from each recording session. The corpus consisting isolated numerals (*Marathi, Hindi, English and Arabic*), Colors, Months, Fruits, City Names and frequently used dictionary words (*in Marathi, Hindi and English*) were selected for the constitution of database. Table 1 shows the

database generated during acquisition of data form native speakers.

Each speaker was asked to utter 10 repetitions of the target words of corpus. The audio visual data was acquired through three channels. Each audio-visual utterance was recorded for two seconds and sampling rate for visual signal is 25fps. Male and Female speakers have been asked to utter the set of numeral words in closed natural environment under 'close-open-close' mode and data is recorded for three channels. The data stream from each user have been collected for target corpus in Marathi, Hindi and English languages. The volume of data for three channel *(Front profile, 45^0 profile, and Side pose profile)* in three language is shown in table 1. The total volume of corpus by native speakers is 2,78,300 audio visual words. The total samples in respective class of isolated word in a corpus was estimated as,

$$Total\ Sample = NU \times NW \times NWR \times C$$

Where NU: Number of Users, NW: Number of Words, NWR: Number of Word Repetitions and C: Channel.

The volume (9000 samples) of non-native speakers uttering numerals set in English, Hindi and Arabic is as shown in table 2.

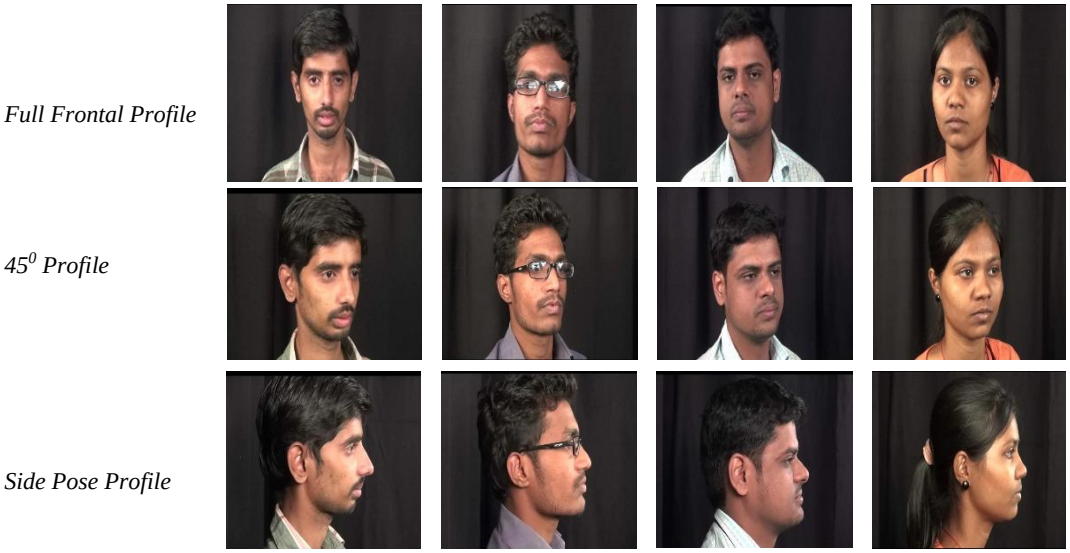

Fig. 3 Visual Profiles of Native Speakers in natural mode

Fig. 4 Visual Profile of Non-Native Speakers in natural mode

Table 2: Corpus of Native-Speaker

Isolated words Class	No of words	Corpus Set	Speakers	Total Samples for three channel
Numerals	10	English {Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine}	48	14,400
		मराठी {शुन्य, एक, दोन, तीन, चार, पाच, सहा, सात, आठ, नऊ}	48	14,400
		हिंदी { शुन्य, एक, दो, तीन, चार, पांच,छः, सात, आठ, नौ }	48	14,000
		Arabic {'صفر' إلى 'تسعة'} (Non Native Speaker)	10	3,000

City Names	10	English/Marathi/Hindi{'Aurangabad', 'Beed', 'Hingoli', 'Jalgaon', 'Latur', 'Mumbai', 'Osmanabad', 'Parbhani', 'Pune', 'Satara' }	12	3,600
Months	12	English {January, February, March, April, May, June, July, August, September, October, November, December}	48	17,280
		मराठी{जानेवारी, फेब्रुवारी, मार्च, एप्रील, मे, जून, जुलै, ऑगस्ट, सप्टेंबर, ऑक्टोबर, नोव्हेंबर, डिसेंबर}	48	17,280
		हिंदी {जनवरी, फरवरी, मार्च, अप्रैल, मै, जून, जुलै, अगस्त, सितंबर,ऑक्टोबर, नवंबर, दिसंबर }	48	17,280
Colors	11	English {Red, Green, Blue, Yellow, Purple, Brown, Magenta, Orange, White, Black, Pink}	48	15,840
		मराठी { लाल, हिरवा, निळा, पिवळा, जांभळा, कथ्था, केशरी, पांढरा, काळा, गुलाबी}	48	15,840
		हिंदी {लाल, हरा, नीला, पिला, बैगणी, भुरा, नारंगी, सफेद, काला, गुलाबी}	48	15,840
Fruits	10	English {Apple, Grape, Banana, Pineapple, Jackfruit, Watermelon, Guava, Sweet lime, Oranges, Pomegranate, Mango}	48	14,400
		मराठी {सफरचंद, द्राक्ष, केळ, अननस, फणस, टरबुज, पेरू, मोसंबी, संत्री, डाळिंब, अंबा}	48	14,400
		हिंदी {सेब,अंगूर, केला, अननस, फणस, तरबुज, अमरूद, मोसंबी, संतरा, अनार, आम}	48	14,400
Day to day communication words	20	English { Not, Go, Our, Day, Night, Morning, Afternoon, Something , Give, Many, Again, Increase, Here, Near, There, Other, Name, Language, Therefore, Time }	48	28,800
		मराठी {अभिप्रेत, अनुभव, अपेक्षित, आरामदायक, दूरदृष्टी, कामगिरी, महत्वपूर्ण, मजबूत, नमस्कार, पराक्रम, परिणाम, पुरातन, सभाग्रह, समावेश, समाविष्ठ, सुंदरता, स्वतंत्रता, स्वयंपाक, उदाहरण, विद्यापीठ}	48	28,800
		हिंदी { नहीं, जाना, अपने, दिन, रात ,सुबह ,दोपहर ,कुछ ,दिया, बहुत ,फिर ,अधिक, यहाँ , पास ,वहाँ ,अन्य, नाम, भाषा, इसलिए, समय} [16]	48	28,800
		Total Volume of Corpus by Native Speaker		2,78,360

Table 3: Corpus of non-Native speaker

Language	Isolated Word Class Numerals	Number of Speaker	Volume of Corpus All Channels
Arabic	{ 'शुन्य' ते 'नऊ' }	10	3,000
Hindi	{ 'शुन्य' से 'नौ' }	10	3,000
English	{ 'Zero' To 'Nine'}	10	3,000
		Total Volume of Corpus	**9,000**

On the Similar ground, the dataset of induced mode is also collected. In this mode of data acquisition each speaker was requested to apply florescent red color or red color lipstick on their lips at the time of acquisition. Visual profiles of induced color mode is as shown in figure 5. This data was acquired in three channel with all aforesaid specification. The objective of this dataset is to explore the possibility of automatic tracking and generation of accurate deformation of mouth at the time of word utterance. This form of data found suitable for the design of lip-reading based on shape deformations by color based segmentation of mouth.

Fig. 5 Visual Profile of speakers in induced color mode

Table 4 Volume of Corpus in Induced color Mode

Isolated word Class	Number of Speaker	Volume of Data
Marathi Language Day to day commonly used Words	10	6000
Months	10	3600
	Total Volume of Data	**9600**

The total corpus of *'vVISWa'* data set including Natural and Induced color with native and non-native is 2,96,960 and will be available free of cost for research purpose and evaluation purpose to the researcher in the area of multi-pose audio-visual speech recognition. The database is made available on CD-ROM as per requirement or via ftp service form www.visbamu.in website.

4.1 Experiments on vVISWa dataset

The corpus is evaluated by introducing defined set of experiments. The corpus was processed for sampling and preprocessing of all samples. The facial components such as mouth, eyes and nose were localized using Viola-Jones Algorithm [17]. This algorithm detect the face using detector based on *AdaBoost* classifier that uses cascades of weak classifiers to boost. The mouth region from each frame was automatically isolated and were segmented from frame to constitute mouth vector processing and feature extraction. From the experiment it was seen that '*Viola-Jones'* work fine for full frontal visual profile, but when it was tested over 45° and side pose visual stream where some portion of mouth was not visible, the performance of algorithm was found to be degraded. Amarsinh Varpe et.al has discussed isolation of Region of Interest (ROI) for Multi-pose AVSR using '*vVISWa*' dataset and it was measured performance of skin color based detection of ROI over 'Viola-Jones' algorithm was concluded that 'skin color based detection of ROI' was found better as compared with 'Viola-Jones' algorithm under multi-pose AVSR scenario [18]. Similarly Prashant Borde et.al has discussed the contribution of visual features that are computed through Zernike moments in association with MFCC for recognition of isolated city words from 'vVISWa' dataset [19]. Besides this reported work on the database, this 'vVISWa' database have processed for various appearance based feature methods as well as performance of K-Means, Random Forest and Hidden Markov Model classifier are also tested. This database is open for the further investigation by the researchers in this field.

5. CONCLUSION

This paper presents comprehensive information about 'vVISWa' database and its significance in design robust multi-pose audio visual speech recognition. This paper also provides information about the complexities involved in corpus design and enriched existing dataset with induced color set of visual words. The presented database contains the isolated words as well as continuous words of Marathi, Hindi and English language for multi-pose audio-visual experiments. There are several applications where 'vVISWa' dataset may be used like person identification, authentication, AVSR, Multi pose AVSR and moreover it can be extended to Multi-pose speaker recognition, lip synchronization, language identification and Human Computer Interactions.

6. ACKNOWLEDGEMENTS

Authors would like to thank DST-SERB (Department of Science and Technology-Science and Engineering Research Board) for research project entitled, "Analysis of visual speech from multiple camera angle for enhancement of speech recognition" sanctioned under FAST TRACK SCHEME FOR YOUNG SCIENTIST vide letter no SERB/F/1766/2013-14 for their support to carry out this research work. Author would also like to extend their sincere thanks to Dr. Babasaheb Ambedkar Marathwada University for their support.

7. REFERENCES

[1] H. McGurk and J. MacDonald. Hearing lips and seeing voices. Nature, pages 746–748, September 1976.

[2] G. Potamianos, C. Neti, and G. Gravier. Recent advances in the automatic recognition of audio-visual speech. Proceedings of the IEEE, 91(9):1306–1326, 2003.

[3] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys,M. Liu, T. Huang, AVICAR: audio-visual speech corpus in a car environment, Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2004, pp. 380–383.

[4] V. Zue, S. Sene, J. Glass, Speech database development: TIMIT and beyond, Speech Commun. 9 (4) (1990) 351–356.

[5] I. Matthews, T. Cootes, J. Bangham, S. Cox, R. Harvey, Extraction of visual features for lip reading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.

[6] S. Cox, R. Harvey, Y. Lan, J. Newman, B. Theobald, The challenge of multi speaker lip-reading, Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP), 2008, pp. 179–184.

[7] T. Hazen, K. Saenko, C. La, J. Glass, A segment-based audio-visual speech recognizer: data collection, development, and initial experiments, Proc. Int. Conf. Multimodal, Interfaces, 2004, pp. 235–242.

[8] E. Patterson, S. Gurbuz, Z. Tufekci, J. Gowdy, CUAVE: a new audio-visual database for multimodal human-computer interface research, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 2, 2002, pp. 2017–2020.

[9] P. Lucey, G. Potaminanos, S. Sridharan, Patch-based analysis of visual speech from multiple views, Proc. Int. Conf. Auditory–Visual Speech Process. (AVSP), 2008, pp. 69–74.

[10] http://www.ee.surrey.ac.uk/Projects/LILiR/index.html.

[11] P. Price, W. Fisher, J. Bernstein, D. Pallett, Resource Management RM2 2.0, Linguistic Data Consortium, Philadelphia, 1993.

[12] McCool, Chris, Sebastien Marcel, Abdenour Hadid, Matti Pietikainen, Pavel Matejka, Jan Cernocky, Norman Poh et al. "Bi-modal person recognition on a mobile phone: using mobile phone data." In Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on, pp. 635-640. IEEE, 2012.

[13] G. Zhao, M. Barnard, M. Pietikäinen, Lipreading with local spatiotemporal descriptors, IEEE Trans. Multimedia 11 (7) (2009) 1254–1265.

[14] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, J. Acoust. Soc. Am. 120 (5) (2008) 2421–2424.

[15] K. Messer, J.Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, Proc. Int. Conf. Audio, Video-Based Biometrics Person Authentication (AVBPA), 1999.

[16] Resource Centre for Indian Language Technology Solutions (CFILT), IIT Bombay, http://www.cfilt.iitb.ac.in/

[17] P.A.M.J. Viola. "Rapid Object Detection Using a Boosted Cascade of Simple Features," in Proc. IEEE Conf. Computer vision and Pattern Recognition.

[18] Amarsinh Varpe, Prashant Borde, Pallavi Pardeshi, Sadhana Sukale, Pravin Yannawar, "Analysis of Induced Color for Automatic Detection of ROI Multipose AVSR System", Springer International conference on Information System Design and Intelligent Application, 10.1007/978-81-322-2247-7_54, pp 525-538, 2015.

[19] Borde Prashant, Amarsinh Varpe, Ramesh Manza, and Pravin Yannawar. "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition." International Journal of Speech Technology (2014): 1-9.