Hierarchical Picture of existing Audio-Visual Speech Database

Bibish Kumar K T, Sunil John, Muraleedharan K M, R K Sunil Kumar

Abstract: ***Despite the technological improvement and arrival of new methodologies in the different process of a speech-based applications, a parallel development is not observed in the availability of audio-visual speech database. This paper provides a detailed hierarchical picture of the existing audio-visual speech database. Since the performance of a speech-based application deeply depends on the different parameters like the number of speakers, speaker variability, phonetically balanced sentences, recording quality etc. involved in the creation of a database to attain specific task. This paper gave more importance to these parameters involved in the exciting audio-visual speech database rather than the experimental side which need linguistic knowledge about the concerned language in the feature extraction task and classification task. This paper is arranged in such a way that a new face in this realm can capture the needy things to build a speech database in his language. In addition, this paper differs from other review papers in the aspect that it gives equal importance to the available audio-visual speech database in the resourced and under-resourced languages.***

***Keywords*: Audio-Visual speech database, Speech-based applications, Video Parameters and Audio Parameters.**

I. INTRODUCTION

The human brain uses both acoustical and facial information and extracts relevant information to decode the underlined spoken utterances, especially in acoustically hampered condition. Likewise, the intelligent system uses audio-visual sensors, namely microphones and cameras, to capture and analyze it to facilitate an efficient Human-machine interface. There is a profusion of intelligent systems which uses both speech modality and visual modality alone, such as speech recognition, speaker recognition, emotion recognition, forensic applications, lip-reading, lip tracking, lip synchronization, re-dubbing etc. From 1976 [1] onwards these intelligent speech systems have witnessed significant improvement in the target goals by utilizing the progress in computer vision and machine learning area. However, the essential element needed for the robust performance of all these intelligent systems is the availability of suitable speech database.

Because of the difficulties associated with the high volumes of data necessary for the simultaneous recording video and audio and time consumption, the creation of audio-visual databases has limited when compared to the audio-only database. Besides, only a few reported audio-visual databases are open to the public, which is mainly in English. However, the fascination of speech-based applications has made various research groups curious to develop systems in their native language. Language exploration is needed to opt the convenient methods for feature extraction and classification task, which are the essential components of any speech-based applications.

This paper presents a historical sketch in different aspects of available audio-visual speech database from 1988 to the present. Even though many papers have reviewed some of the essential audio-visual speech databases, this paper gives an overall view in the creation of the databases in resourced and under-resourced language [2], and technical advances occur in the recording devices and approaches. A new face in this realm gets a proper insight into the quantity and quality of basic requirements and resources needed to build an audio-visual speech database for a task in his language.

II. KNOWN AUDIO-VISUAL SPEECH DTABASE

Over the years of development and unceasing research in speech-based applications has endorsed the dearth of standard audio-visual speech database. The need for diversity in resources and massive storage capacity is the main challenge that produces the remarkable difference in the statistic between uni-modal and bi-modal speech database. Creation of exhaustive and methodically arranged audio-visual speech database build up a worldwide acceptance in the research community. An audio speech can capture at high quality with relatively low cost, but a high-quality visual speech can capture the dynamic movements of the lip accurately, which is relatively large. Massive storage space needed for the storage and distribution of audio-visual speech database than the audio-only speech database. Instead of giving a detailed description of the characteristics of the existing database, this paper presents a tabular representation of available audio-visual speech database which offers a clear picture of parameters and comparison among different others.

Manuscript published on 30 September 2019
* Correspondence Author

Bibish Kumar K T*, Computer Speech & intelligence Research Centre, Department of Physics, Government College, Madappally, Vadakara, Calicut, Kerala, India. Email: bibishkrishna@gmail.com

Sunil John, Computer Speech & intelligence Research Centre, Department of Physics, Government College, Madappally, Vadakara, Calicut, Kerala, India. Email: suniljohn.e@gmail.com

Muraleedharan K M, Computer Speech & intelligence Research Centre, Department of Physics, Government College, Madappally, Vadakara, Calicut, Kerala, India. Email: muraleedharan.km@gmail.com

R K Sunil Kumar, School of Information Science and Technology, Kannur University, Kerala, India. Email: seuron74@gmail.com

Table. I summarize the historical background of audio-visual speech database in terms of gender distribution, speech corpus, hardware setup and some distinctive features. The history begins with Petajan 1988 for lip-reading digit recognizer

Table- I: Summary of known audio-visual speech database

Database - Year	Speaker (Female, Male)	Corpus - Repetition	Video Parameters (Pixel size, FPS)	Audio Parameters (Sampling Frequency, bit)	Special Features
TULIPS1[4] - 1995	12 (9,3)	First four English digit – twice.	100 x 75, 30 fps. Mouth region.	unknown	Isolated digits.
DAVID [5] - 1996	124	Isolated digits. English-alphabet E-set. Video-conference control commands. 'VCVCV' nonsense utterances.	640 x 480, 30 fps. Full face. Frontal View.	unknown	Speech/Person recognition. Contain 4 corpus with different research theme. Complex background and variable illumination. Contain individual speaker and more than one speaker during recording. Lack of head pose and facial expression variations.
M2VTS [6] – 1997 Multi Modal Verification for Teleservices and Security applications	37	Numbers (0 to 9) – 5 times. French language.	286 x 350, 25 fps. Full face. Frontal view.	48kHz, 16 bit.	Speech verification, face recognition. Mostly French Speakers. Head rotation (left, right, up and down). Presence of glasses and hats.
XM2VTSDB [7] – 1999 Extended M2VTS Database	295	Three sentences (numbers and word) – twice.	720 x 576, 25 fps. Full face. Frontal view. 2 Camera used.	32kHz, 16 bit.	Personal identity verification. Head rotation (left, right, up and down). Recorded in extremely controlled condition. Text dependent. http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/
IBM ViaVoice [8] - 2000	290	Continuous speech with verbalized punctuation. Dictation style.	704 x 480, 30 fps. Full face. Frontal view.	16kHz, 16 bit.	Speaker-Independent Large Vocabulary Continuous Speech Recognition task. Total duration of database- 50 hours.
AMP/CMU [9] – 2001 Advanced Multimedia Processing Lab	10 (3,7)	78 Isolated words (date and time, month, day and miscellaneous) – 10 times.	720 x 480, 25 fps. Full face. Frontal view.	16kHz, 16 bit.	Lip reading. Presence of glasses and hats. Recorded in controlled situation.
AV Letters [10] - 2002	10 (5, 5)	Isolated letters (A to Z) – 3 times. Total 780 utterances.	376 x 288, 25 fps. Full face. Frontal view.	22.05kHz, 16 bit.	Speech recognition.
CUAVE [11] – 2002 Clemson University Audio Visual Experiments	36 (19, 17) Speaker pairs -20	Isolated digits. Connected digits. Total 7000 utterances.	720 x 480, 29.97 fps. Shoulder and head. Frontal and Profile view.	16kHz, 16 bit	Speaker independent digit recognition. Speaker independent database. Contain individual speaker, speaker pairs and moving speakers. Head movement in side-to-side, back-and-back. Presence of glasses, facial hairs and hats. Fit to one DVD-data disk.
VidTIMIT [12] - 2002	43 (19, 24)	10 TIMIT sentences per speaker. First 2 sentences are same for all but remaining 8 are unique.	512 x 384, 25 fps. Frontal view.	32kHz, 16 bit.	Multi-modal person verification. Data acquisition with 3 sessions. Extended head rotation. Change in speaker appearance and voice during each session. Variability in camera zoom factor and background noise. http://conradsanderson.id.au/vidtimit/

DUTAVSC [13] - 2002	8 (1, 7)	POLYPHONE corpus. Phonetically rich sentences. Connected digits. Spelling. Application driven utterance. Dutch language.	384 x 288, 25f fps. Frontal view. Lower face view.	44kHz, 16 bit.	Audio visual speech recognition.
BANCA [14] - 2003	52 (26, 26) for each language class.	Numbers. Names. Addresses. Date of birth. 4 Languages-English, French, Italian and Spanish.	720 x 576, 25 fps. Shoulder and head. Frontal view. 2 Camera used.	32kHz, 12& 16 bit. 2 Microphone used.	Multi-modal identity verification. Recorded in controlled, degraded and adverse condition. Text independent. Lack of head pose and facial expression variations. http://www.ee.surrey.ac.uk/CVSSP/banca/
AVICAR [15] – 2004 Audio-Visual Speech in a Car.	100 (50, 50)	Isolated digits. Isolated letters. Phone numbers. TIMIT sentences. Total 59,000 utterances.	720 x 480, 30 fps. Full face. 4 Frontal views. 4 Camera array.	48kHz, 16 bit. 8 Microphone array.	Speech recognition in car. 60% American English others Latin American, European, East Asian and South Asian. Recorded in 5 noisy condition (automotive noise). http://www.isle.illinois.edu/sst/AVICAR/
AV-TIMIT [16] - 2004	223 (106, 117)	450 TIMIT-SX sentences. Each speaker utter 20 sentences. First sentences is common and other 19 sentences are different.	720 x 480, 30 fps. Full face. Frontal view.	16kHz, 16 bit.	Speaker independent continuous speech recognition. Continuous phonetically balanced speech. Contain multiple speakers. Controlled office environment. Presence of facial hairs, glasses and hats. Recorded in different illumination condition.
AVOZES [17] – 2004 Audio Video OZtralian English Speech	20 (10, 10)	Digits. Words. Phrases. Continuous word. Australian English language. Total of 56 sequences per speaker without repetition.	720 x 480, 29.97fps.	48kHz, 16 bit.	Modular approach database- each module addresses specific task. 6 Modules- 1 general module (speaker independent) and 5 speaker specific module. Native speakers of Australian English. Presence of glasses, facial hairs and lip highlighter. Speaker personal data acquisition mode.
MANDARIN CHINESE [18] – 2004	225	Continuous speech. Chinese language. Total 17,000 utterances.	720 x 576, 25 fps. 768 x 576, 25 fps. 7 Cameras used.	48Hz, 16 bit. 12 Microphones used.	Audio-visual speech/speaker recognition and 3Dface modelling. Recorded in 2 sessions.
IBM Infrared Headset [19] - 2004	79	Continuous digits.	720 x 480, 30 fps. Frontal view. Mouth-chin area.	22kHz, 16 bit.	Audio-Visual Speech Recognition. Use infrared videos. Captured in real-time visual condition.
VALID [20] – 2005	106 (29, 77)	XM2VTS speech corpus.	720 x 576, 25 fps. Full face with shoulder. Frontal view.	32kHz, 16 bit.	Multi-modal speaker/speech recognition. 97 Europeans and 9 Asians. 5 Recording session – 1 controlled and 4 uncontrolled (varying noise, illumination). Presence of facial hairs. Text dependent.
UWB-04-HSCAVC [21] – 2006 University of West Bohemia-2004-Hundred Speakers Czech Audio-Visual Corpus	100 (61, 39)	200 Sentences (50 shared and 150 unique). Slavonic language (Czech and Russian).	720 x 576, 25 fps. Full face. Frontal view.	44kHz, 16 bit. 2 Microphones used.	Audio-visual speech recognition. Visual speech parameterizations. Recorded in laboratory condition. Constant illumination and static head position. Mean age is 22.

GRID [22] – 2006	34 (16, 18)	Command sentences. Each sentence contains six-word sequence. Total corpus size 34,000.	720 x 576, 25 fps Full face. Frontal view.	25kHz, 16bit	Speech recognition. Mean age is 27. http://spandh.dcs.shef.ac.uk/gridcorpus/
AV Letters 2 [23] - 2008	5	26 Isolated letters – 7 times.	1920 x 1080, 50 fps. Full face. Frontal view.	48kHz, 16 bit.	Speech recognition. High-definition version of AV Letter database.
UWB-07-ICAVR [24] – 2008 University of West Bohemia-2007-Impaired Conditions audio visual speech Recognition	50 (25, 25)	200 Sentences (50 shared and 150 unique). Czech language. Total 10,000 continuous utterances.	720 x 576, 50 fps (high quality). 640 x 480, 30 fps (low quality). 2 Camera used.	44kHz, 16 bit. 2 Microphone used.	6 types of illumination condition Average age is 22.
IV2 [25] – 2008	300	15 French sentences.	780 x 576, 25 fps (high quality). 640 x 480, 25 fps (low quality). 2 Camera used. Full face. Frontal and profile view.	2 Microphone used.	Face recognition. Majority data acquisition within single session. Pose, expression, illumination and glass variability. Different illumination levels and orientations. Iris image, 3D laser scanner face data.
DXM2VTS [26] – 2008 Damascened XM2VTS	295	XM2VTS database. Additional videos containing several degradation level of background noise.	720 x 576, 25 fps. Full face. Frontal view. 2 Camera used.	32kHz, 16 bit.	Face recognition. Internal video distortion (blur, salt and pepper and rotation). External video distortion (zooming and dynamic background noise).
IBM Smart-Room [27] – 2008	38	Connected digit strings. Total 1661 utterances.	368 x 240, 30 fps. 3 Cameras used. Full face. Frontal and profile view.	22kHz, 16-bit. 2 Microphones used.	Lip-reading system.
HIT-AVDB-II [28] – 2008 Harbin Institute of Technology Audio Visual Speech Database II	30 (15, 15)	Digits. Chinese poems. Tongue twisters of Chinese and English. Greek alphabets. Music notes. Mandarin vowels.	720 x 576, 25 fps. 4 Cameras used. 4 views- frontal, profile, 30^0 and 60^0.	unknown	Visual speech, Biometrics, lip tracking and multi view. Database witness emotions, fast mouth movements and tunes. Recorded in 3 sessions in different day time to capture varying speaker appearances and background. Presences of spectacles and hair ornaments.
OuluVS [29] - 2009	20 (3, 17)	10 daily use short phrases – 9 times. Total 817 sequences.	720 x 576, 25 fps. Full face. Frontal view.	No audio	Visual only speech recognition.
WAPUSK20 [30] - 2010	20 (9, 11)	100 GRID database sentence. Total 2000 sentences.	640 x 480, 32 fps. Full face. Frontal view.	16kHz, 32 bit. 4 audio channels	Stereoscopic video. Speakers – 2 England, 1 Greece, 1 Kazakhstan and 1 Spain. All other native German speakers. Mean age is 29.

BL-Database [31] – 2011 Blue Lips-Database	17 (8, 9)	238 sentences. French language. Diphone rich utterances.	576 x 720, 25 fps (front view camera). 640 x 480, 30 fps (side view camera). 640 x 480, 30 fps (depth camera). 3 Cameras used. Full face. Frontal and side view.	44.1kHz, 16 bit. 2 Microphones used.	Audio-Visual speech recognition. Recorded in 2 sessions- First sessions for 2D analysis and second session for 3D analysis of mouth movement. Native French speakers. Blue lipstick used.
UNMC-VIER [32] - 2011	123 (49, 74)	11 XM2VTS sentences. Sequence of numerals.	708 x 640, 25 fps (3 high quality cameras). 320 x 240, 29 fps (low quality camera). 320 x 240, 15 fps (low quality camera). 5 Cameras used. Full face. Frontal and left profile view.	48kHz, 16 bit (From high qualitycamera) . 44kHz, 16 bit (From low quality camera). 32kHz, 16 bit (From low quality camera). 22kHz, 16 bit (Audio device).	Audio-Visual speech/speaker recognition. Contain multiple visual variation in the same video recording. Visual variations- illumination, expression, head pose, background and resolution. Head rotation- left-to-right and up and down. Spoken in normal and slow speech pace. Recorded in controlled and uncontrolled environment with different devices. Speakers- 116 Asians, 4 Africans and 3 Europeans.
AusTalk [33] - 2011	1000	Multiple read and spontaneous speech tasks. Australian English	640 x 480, 48 fps	Three high quality microphones are used	Applied for Speaker verification, Audio-Visual speech Recognition, Forensic Speaker recognition. Total duration- 3000 hours. http://austalk.edu.au
MoBio [34] - 2012	152 (52, 100)	32 questions (short response questions, short response free speech, set speech, and free speech).	640 x 480, 16-30 fps.	48kHz, 16 bit.	Mobile-based biometric system. Database almost captured from mobile devices. High variability in pose and illumination.
LILiR [35] – 2012 Language Independent Lip Reading	20	Resource management corpus. Total 200 sentences per speaker.	5 cameras used (2 HD and 3 SD). Full face. 5 views- frontal, profile, 30^0, 45^0 and 60^0.	unknown	Continuous speech recognition.
AVAS [36] – 2013 Audio-Visual Arabic Speech	50	36 daily words. 13 casual phrases. Arabic language.	640 x 480, 30 fps. Full face. Frontal view.	48kHz, 16 bit.	Audio-Visual speech/speaker recognition. Visual variations-4 illumination condition and 5 head pose variations. First database in Arabic language.
LUNA-V [37] – 2014 Loughborough University Audio-Visual data corpus	10 (1,9)	English digits- 5 times. Sentences were taken from TIMIT.	1920 x 1080, 25 fps. Full face. Frontal view.	16kHz,16 bits.	Audio-visual speech recognition. Phone Recognition.

AGH [38] - 2015	166 (one third female)	Isolated words and Numbers. Polish language. Total 117,450 words.	1920 x 1080, 50 fps.	44.1kHz, 16 bit.	Automatic speech recognition and Text-to-speech systems. Largest audio-visual Polish corpus.
OuluVS2 [39] - 2015	53 (13, 40)	Continuous digits. Phrases. TIMIT sentences.	1920 x 1080, 30 fps (From 5 HD camera). 640 x 480, 100 fps (From frontal HS camera). Six cameras used. Full face. 5 views- frontal, profile, 30^0, 45^0 and 60^0.	High quality audio.	Multi-view audio-visual database. No native English speakers. Speakers- European, Chinese, Indian/Pakistan, Arabian and African. Neutral facial expression and static head pose. Simultaneous recording by 6 cameras from 5 different views. http://www.ee.oulu.fi/research/imag/OuluVS2/
TCD-TIMIT [40] - 2015	62 (30, 32)	6913 phonetically rich TIMIT sentences.	1920 x 1080, 30fps. 2 Cameras used. 2 views- frontal and 30^0.	16kHz, 16 bit.	Two class of speakers- 3 professional lip speakers (female) and non-lip speakers. Presence of glasses and piercings. https://sigmedia.tcd.ie/TCDTIMIT/
vVISWa [41] – 2016 Visual Vocabulary of Independent Standard Words	58 (20, 38)	Isolated words – 10 times. Continuous words – 10 times. Marathi, Hindi and English languages. Total 2, 96,960 words.	720 x 576, 25 fps. 3 Cameras used. 3 views- frontal, profile and 30^0.	unknown	Multi-pose audio visual speech recognition system for 3 languages. Speakers- native (20F & 28M) and non-native (Iraq and Yemen) (10M). Presence of glasses and caps. Induced mode of data acquisition contain speakers with lipstick.
MODALITY [42] - 2017	35 (9, 26)	168 commands.	1080 x 1920, 100 fps. 2 Cameras used. Full face. Partial front views.	44.1kHz, 16 bit. Array of 8 microphones used.	Audio-visual speech recognition. Speakers- native and non-native English speakers. Noise varying recording setup.
AVID [43] - 2017	10 (5, 5)	1040 Sentences. Influenced by GRID. Indonesian language.	1280 x 962, 48 fps. Full face. Front view.	44.1kHz, 16 bit.	Multimodal ASR system. First database in Indonesian language.
Audio-Visual Lombard Speech [44] - 2018	54	Plain and Lombard speech. Total 5400 utterances.	720 x 480, 24 fps (from frontal web cam). 864 x 480, 30 fps (from side webcam). Frontal view.	48kHz, 24 bits.	Characterization of across-speaker variation.

III. CONCLUSION

A diversity of standard database has reported, and most of them claim to be beneficial for the specific task. For the speech recognition task, the most favourable speech material is continuous speech when compared to isolated speech for speaker verification task. Speaker recognition task requires a large size and high variability of speaker population when compared to the speech recognition task. Multi-view visual speech performs better in lip reading task. So, the database to be invented should serve more than one goal so that it will be useful for alternative research works.

The resolution and frame rate of the camera chosen by means of a trade-off involving computational complexity and high-quality visual speech information. The database should be captured with uniform distribution to avoid gender imbalance. The main criteria needed for speech database construction is that it should have a large phonetically balanced speech corpus uttered by many unique speakers in an uncontrolled environment. It is mandatory to figure out the peculiarities of the language of the database and its linguistic background and comparing it with other groups of languages which help to resolve the issues arose during the creation of the database in under-resourced languages.

REFERENCES

1. McGurk, Harry, and John MacDonald. "Hearing lips and seeing voices." *Nature* 264, no. 5588 (1976): 746.
2. Besacier, Laurent, Etienne Barnard, Alexey Karpov, and Tanja Schultz. "Automatic speech recognition for under-resourced languages: A survey." *Speech Communication* 56 (2014): 85-100.
3. Petajan, Eric, Bradford Bischoff, David Bodoff, and N. Michael Brooke. "An improved automatic lipreading system to enhance speech recognition." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 19-25. ACM, 1988.
4. Movellan, Javier R. "Visual speech recognition with stochastic networks." In Advances in neural information processing systems, pp. 851-858. 1995.
5. Chibelushi, C. C., S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston. "Design issues for a digital audio-visual integrated database." (1996): 7-7.
6. Pigeon, Stéphane, and Luc Vandendorpe. "The M2VTS multimodal face database (release 1.00)." In International Conference on Audio-and Video-Based Biometric Person Authentication, pp. 403-409. Springer, Berlin, Heidelberg, 1997.
7. Messer, Kieron, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. "XM2VTSDB: The extended M2VTS database." In Second international conference on audio and video-based biometric person authentication, vol. 964, pp. 965-966. 1999.
8. Neti, Chalapathy, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari. Audio visual speech recognition. No. REP_WORK. IDIAP, 2000.
9. Chen, Tsuhan. "Audiovisual speech processing." IEEE Signal Processing Magazine 18, no. 1 (2001): 9-21.
10. Matthews, Iain, Timothy F. Cootes, J. Andrew Bangham, Stephen Cox, and Richard Harvey. "Extraction of visual features for lipreading." IEEE Transactions on Pattern Analysis and Machine Intelligence 24, no. 2 (2002): 198-213.
11. Patterson, Eric K., Sabri Gurbuz, Zekeriya Tufekci, and John N. Gowdy. "CUAVE: A new audio-visual database for multimodal human-computer interface research." In Proceedings of International Conference on Acoustics, Speech and Signal Processing (CASSP, pp. II-2017. IEEE, 2002.
12. Sanderson, Conrad. The vidtimit database. No. EPFL-REPORT-82748. IDIAP, 2002.
13. ChiŃu, Alin G., and Leon JM Rothkrantz. "Building a data corpus for audio-visual speech recognition." In Euromedia, pp. 88-92. 2007.
14. Bailly-Bailliére, Enrique, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas et al. "The BANCA database and evaluation protocol." In International conference on Audio-and video-based biometric person authentication, pp. 625-638. Springer, Berlin, Heidelberg, 2003.
15. Lee, Bowon, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang. "AVICAR: Audio-visual speech corpus in a car environment." In Eighth International Conference on Spoken Language Processing. 2004.
16. Hazen, Timothy J., Kate Saenko, Chia-Hao La, and James R. Glass. "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments." In Proceedings of the 6th international conference on Multimodal interfaces, pp. 235-242. ACM, 2004.
17. Goecke, Roland, J. Bruce Millar, A. Zelinsky, and J. Robert-Ribes. "A detailed description of the AVOZES data corpus." In Proc. 10th Austral. Int. Conf. Speech Science & Technology, pp. 486-491. 2004.
18. Liangi, Luhong, Yu Luo, Feiyue Huang, and Ara V. Nefian. "A multi-stream audio-video large-vocabulary mandarin chinese speech database." In Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on, vol. 3, pp. 1787-1790. IEEE, 2004.
19. Huang, Jing, Gerasimos Potamianos, Jonathan Connell, and Chalapathy Neti. "Audio-visual speech recognition using an infrared headset." Speech Communication 44, no. 1-4 (2004): 83-96.
20. Fox, Niall A., Brian A. O'Mullane, and Richard B. Reilly. "VALID: A new practical audio-visual database, and comparative results." In International Conference on Audio-and Video-Based Biometric Person Authentication, pp. 777-786. Springer, Berlin, Heidelberg, 2005.
21. Císař, Petr, Jan Zelinka, Miloš Železný, Alexey Karpov, and Andrey Ronzhin. "Audio-Visual speech recognition for Slavonic languages (Czech and Russian)." In Proc. of 11-th International Conference SPECOM-2006, St. Petersburg:—Anatoliya. 2006.
22. Cooke, Martin, Jon Barker, Stuart Cunningham, and Xu Shao. "An audio-visual corpus for speech perception and automatic speech recognition." The Journal of the Acoustical Society of America 120, no. 5 (2006): 2421-2424.
23. Cox, Stephen J., Richard W. Harvey, Yuxuan Lan, Jacob L. Newman, and Barry-John Theobald. "The challenge of multispeaker lip-reading." In AVSP, pp. 179-184. 2008.
24. Trojanová, Jana, Marek Hrúz, Pavel Campr, and Miloš Železný. "Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition." In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). 2008.
25. Petrovska-Delacrétaz, Dijana, Sylvie Lelandais, Joseph Colineau, Liming Chen, Bernadette Dorizzi, M. Ardabilian, E. Krichen et al. "The IV 2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic, and Talking Face Data), and the IV 2-2007 Evaluation Campaign." In Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on, pp. 1-7. IEEE, 2008.
26. Teferi, Dereje, and Josef Bigun. "Evaluation protocol for the dxm2vts database and performance comparison of face detection and face tracking on video." In ICPR 2008 19th International Conference on Pattern Recognition, pp. 1-4. IEEE, 2008.
27. Lucey, Patrick J., Gerasimons Potamianos, and Sridha Sridharan. "Patch-based analysis of visual speech from multiple views." (2008): 69-74.
28. Yao, Xiaoxin Lin1 Hongxun, and Xiaopeng Hong1 Qian Wang. "HIT-AVDB-II: A new multi-view and extreme feature cases contained audio-visual database for biometrics." (2008).
29. Zhao, Guoying, Mark Barnard, and Matti Pietikainen. "Lipreading with local spatiotemporal descriptors." IEEE Transactions on Multimedia 11, no. 7 (2009): 1254-1265.
30. Vorwerk, Alexander, Xiaohui Wang, Dorothea Kolossa, Steffen Zeiler, and Reinhold Orglmeister. "WAPUSK20-A Database for Robust Audiovisual Speech Recognition." In LREC. 2010.
31. Benezeth, Yannick, Grégoire Bachman, Guylaine Le-Jan, Nathan Souviraà-Labastie, and Frédéric Bimbot. "BL-Database: A French audiovisual database for speech driven lip animation systems." PhD diss., INRIA, 2011.
32. Wong, Yee Wan, Sue Inn Ch'ng, Kah Phooi Seng, Li-Minn Ang, Siew Wen Chin, Wei Jen Chew, and King Hann Lim. "A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities." Pattern Recognition Letters 32, no. 13 (2011): 1503-1510.
33. Burnham, Denis, Dominique Estival, Steven Fazio, Jette Viethen, Felicity Cox, Robert Dale, Steve Cassidy et al. "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box." (2011).
34. McCool, Christopher, Sebastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matejka, Jan Cernocký, 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.
35. Lan, Yuxuan, Barry-John Theobald, and Richard Harvey. "View independent computer lip-reading." In Multimedia and Expo (ICME), 2012 IEEE International Conference on, pp. 432-437. IEEE, 2012.

Retrieval Number: C6483098319/2019©BEIESP
DOI:10.35940/ijrte.C6483.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

36. Samar Antar Alaa Sagheer."Audio-Visual Arabic Speech (AVAS) Database for Human-Computer interaction applications". In International Journal of Advanced Reasearch in Computer Science and Software Engineering, vol: 3, 2013.
37. Ibrahim, Zamri. "A novel lip geometry approach for audio-visual speech recognition." PhD diss., © Mohd Zamri bin Ibrahim, 2014.
38. Żelasko, Piotr, Bartosz Ziółko, Tomasz Jadczyk, and Dawid Skurzok. "AGH corpus of Polish speech." Language Resources and Evaluation 50, no. 3 (2015): 585-601.
39. Anina, Iryna, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis." In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, vol. 1, pp. 1-5. IEEE, 2015.
40. Harte, Naomi, and Eoin Gillen. "TCD-TIMIT: An audio-visual corpus of continuous speech." IEEE Transactions on Multimedia 17, no. 5 (2015): 603-615.
41. Borde, Prashant, Ramesh Manza, Bharti Gawali, and Pravin Yannawar. "'vVISWa'–A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction." International Journal of Computer Applications 137, no. 4 (2016).
42. Czyzewski, Andrzej, Bozena Kostek, Piotr Bratoszewski, Jozef Kotus, and Marcin Szykulski. "An audio-visual corpus for multimodal automatic speech recognition." Journal of Intelligent Information Systems 49, no. 2 (2017): 167-192.
43. Maulana, Muhammad Rizki Aulia Rahman, and Mohamad Ivan Fanany. "Indonesian audio-visual speech corpus for multimodal automatic speech recognition." In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 381-386. IEEE, 2017.
44. Alghamdi, Najwa, Steve Maddock, Ricard Marxer, Jon Barker, and Guy J. Brown. "A corpus of audio-visual Lombard speech with frontal and profile views." The Journal of the Acoustical Society of America 143, no. 6 (2018): EL523-EL529.

AUTHORS PROFILE

Bibish Kumar K T earned his Master's Degree in Physics from the University of Calicut in 2010. He is an active research scholar working in the area of audio visual speech analysis under the supervision of Dr. R K Sunil Kumar. He has six research publications in peer-reviewed journals and international and national conferences to his credit. He joined in PhD program in 2015. He guided 30 post graduate projects in the area of audio and visual speech processing. He developed audio visual Malayalam speech data base. His area of interests are Digital Processing of Speech and Image and Digital Electronics.

Sunil John earned his Master's degree in Physics from the University of Calicut in 2008. He is an active research scholar working in the area of multi accent speech analysis under the supervision of Dr. R K Sunil Kumar. He co-authored three books on computational physics, Methodology of science and Physics practical for undergraduate. He has three research publications in peer-reviewed journals and international and national conferences to his credit. He is currently working as Assistant Professor in the department of physics, St. Mary's College, Sulthan Bathery, Kerala, India. His area of interests are Digital Signal Processing and Computational Physics.

Muraleedharan K M earned his Master's in Physics from the University of Calicut in 2003. He is an active research scholar working in the area of Nonlinear modelling of speech production system under the supervision of Dr. R K Sunil Kumar. He has five research publications in peer-reviewed international journals and national conferences to his credit. He is currently working as Assistant Professor in the department of physics, SARBTM Government College, Koyilandy, Kerala, India. His area of interests are Digital Signal Processing, Nonlinear dynamics and Classical Mechanics. He guided 15 post graduate projects in the area of Digital Signal processing

Dr. R K Sunil Kumar earned his PhD in Speech Signal Processing from University of Calicut, Kerala, India in 2004. He has published several research papers in peer-reviewed journals and national and international conferences in the area of signal processing and artificial neural networks. His research interest includes speech signal processing and neural networks and active noise cancellation. He worked as Assistant Professor in the department of physics, Government College Madappally, Kerala, India for four years. Currently he is working as assistant professor in the department of IT, Kannur University, Kerala, India. He is guiding four Research Scholars in the area of speech processing.