A Review of Audio-Visual Speech Recognition

Thum Wei Seong and M. Z. Ibrahim
Applied Electronic and Computer Engineering Cluster Faculty of Electrical & Electronic Engineering, University Malaysia Pahang, 26600 Pekan, Pahang, Malaysia.
zamri@ump.edu.my

***Abstract*—Speech is the most important tool of interaction among human beings. This has inspired researchers to study further on speech recognition and develop a computer system that is able to integrate and understand human speech. But acoustic noisy environment can highly contaminate audio speech and affect the overall recognition performance. Thus, Audio-Visual Speech Recognition (AVSR) is designed to overcome the problems by utilising visual images which are unaffected by noise. The aim of this paper is to discuss the AVSR structures, which includes the front end processes, audio-visual data corpus used, recent works and accuracy estimation methods.**

***Index Terms*—Audio-Visual Speech Recognition; Audio-Visual Data Corpus; Feature Extraction; Model Validation Techniques; Performance Evaluation.**

I. INTRODUCTION

Speech is one of the most significant method of communication between human and his environment. Voice signals and visible lip movements are generated by human speech organs, such as vocal tract and oral cavity systems. Automatic Speech Recognition (ASR) are developed to translate audio and visual information into formats readable by machines. ASV are normally used to translate or convert speech into text or command, for communication between machines and human beings. For a real-time speech recognition application, a machine must be capable to interpret and make an analysis, and subsequently give immediate response to complete the data transfer [1].

The first research work was conducted at Bell Labs in the early era of 1950s [2]. In the research [3], speech recognition was classified as a technique of extracting related information from the input speech signal and to produce precise recognition of the matching text. Computers are able to react by translating human speech into commands, whereby this creates a good interface for human-computer interaction [3]. Speech technology has made the interaction with machines easier compared with some conventional input devices like pointers or keyboards [4].

In real world recognition application, ASR system are normally affected by noisy environment. Noise is always the main impact factor in the research of recognition system [5]. ASR that exploits the visual modality such as speaker's lip movement and the combination of audio modality leads to audio-visual speech recognition (AVSR) systems. AVSR that utilizes audio and visual information can increase the accuracy over a wide range of acoustic conditions. When an audio signal is corrupted by noise, visual information acquired from the speaker helps to improve the speech recognition performance. Integration techniques between audio and visual modalities has always been the main issue of AVSR. Multimodal recognition that combines both modalities has been proved to outperform mono-modal classifiers [6].

This paper is structured as follows. Firstly, all visual front end and feature extraction will be discussed generally in Section 2. Secondly, Section 3 concentrates on the audio-visual speech data corpus. Then, Section 4 describes the types of integration techniques. Accuracy estimation methods such as cross-validation and bootstrap methods are addressed in Section 5. Finally, Section 6 concludes the paper with a summary and discussion on some issues in AVSR.

II. OVERVIEW OF AVSR SYSTEM

The general block diagram of an AVSR is demonstrated in the form of a flow chart as shown in Figure 1, and information on all steps are explained in the following sections. Feature extraction techniques act as an important role in AVSR, which enhances the performance of the speech recognition system. If the essential information of audio and visual features is extracted perfectly, it is projected to achieve an effective AVSR system.

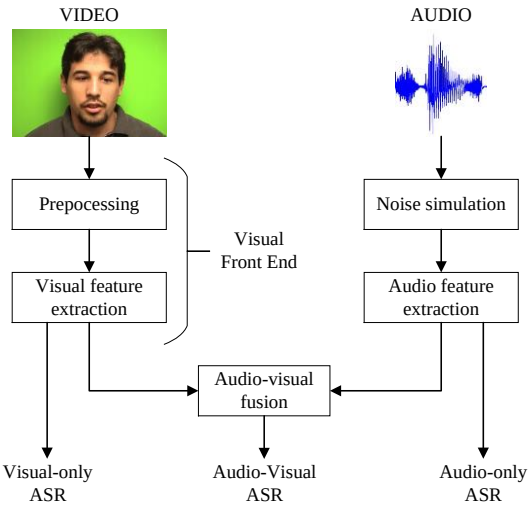


Figure 1: Block diagram of AVSR

A. *Visual Front End*

In recent years, a number of visual front end designs have been recommended in the literature. Studies have shown that the speech visual feature can generally fit into three main categories: 1) appearance-based features; 2) geometrical-based features; and combination of both [7]. Appearance based features expects all pixels within the region of interest (ROI) are informative to speech utterance. To undergo speech classification, the ROI pixel value will experience linear transformation and produce feature vector with reduced dimensionality that contain relevant speech information [8-

10]. The drawback of this feature is that it is sensitive to environment variations, such as subject illumination and head pose. Geometrical-based features can overcome the drawback of appearance-based features [11], as it refers to the speech data restrained in the contours of the speaker's lips, such as width, height and even area of it [12].

In the research [11], experiment have been done to test the robustness of the appearance-based and geometrical-based features with the head pose of the visual image artificially rotated by ±20° with an increment of 5°, and the brightness of the image adjusted by ±20% with an increment of 4%. They work using hue, saturation and value (HSV) for skin detection so that the colour model resembles closer to how real human perceive colour [13] and then convex hull algorithm was applied for lip feature extraction. The experiment concluded that geometrical-based features is more robust to environmental changes (head pose and illumination) compared with appearance-based features.

B. Audio Feature Extraction

An overview has been done on the features extraction method such as linear Predictive Coefficient (LPC) [14], Principal Component Analysis (PCA) [15, 16], Linear Discriminate Analysis (LDA) [17], Independent Component Analysis (ICA) [18] and Mel-Frequency Cepstrum Coefficients (MFCC) [14, 15].

In the paper of [19], the speech recognition system is established using different feature techniques, such as linear predictive coding (LPC) and mel-frequency cepstral coefficient (MFCC). Data corpus of 35 Hindi words with 5 samples per word taken from 3 female and 2 male speakers was used in this research. Data corpus was separated into train database and test database and was tested in two different systems: speaker dependent system and speaker independent system. In speaker dependent environment, MFCC feature extraction is seen to perform well with HMM as classifier compared with LPC. Thus, this paper concludes that MFCC perform better than LCP in most instances, however, it achieves poorer performance in speaker independent environment compared with LPC feature extraction.

III. AUDIO-VISUAL SPEECH DATA CORPUS

The choice of audio-visual data corpus can significantly affect the performance of speech recognition. Although there is plenty of current AVSR data corpora accessible, but several of it having imperfect features in terms of recording quality, number of participants and word coverage. The CUAVE database is a famous database used for speech recognition. It allow researchers to use this database as a baseline, allowing comparison of performances between AVSR methodologies to be made and verified independently.

Previous speech data corpus, Tulips1 [20] and AVletters [21] were developed in 1995 and 1998 with resolution of 100x75 pixels and 80x60 pixels respectively. Later, the newer speech database CUAVE [22] with resolution of 750x576 pixels was developed in 2002. Now, visual features can be extracted with more detailed information in higher definition data corpus. The new database that has been recently established is called the Loughborough University Audio-Visual data corpus (LUNA-V). A comparison evaluation of AVSR using LUNA-V and CUAVE data corpus has been conducted and proved that higher resolution images contribute significant improvement to the performance of visual-only speech recognition [23].

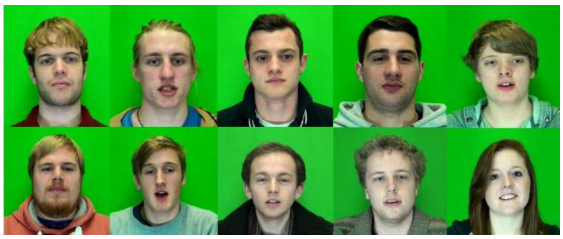


Figure 2: Sample frames of all 10 speakers in LUNA-V data corpus

The LUNA-V database [23] consists of 10 speakers (9 males and 1 female, shown in Figure 2) for the time being and in the future, more speakers will be added to produce a more statistically reliable result. For the first part, each speaker utters digit ‘zero’ to ‘nine’, five times, in English, and the second part, some sentences were adopted from the famous TIMIT database. The video is recorded at 25fps with high resolution of 1280 x 720 and audio at 16 kHz. There are a total of 170 sentences including 1820 words which are available in 10 separate video files (one for each speaker).

Table 1
The Sentences Collected in LUNA-V Data Corpus

Sentences	Contents
1	She had your dark suit in greasy wash water all year.
2	Each untimely income loss coincided with the breakdown of a heating system part.
3	The easy going zoologist relaxed throughout voyage.
4	The same shelter could be built into an embankment or below ground level.

IV. AUDIO-VISUAL INTEGRATION

There remains an absence of clear up regarding the utilization of phrasing relating to the levels of integration in AVSR. Audio-visual integration contributes to a major research topic in AVSR, targeting the combination of audio and visual modalities informative streams into a multi modal classifier which performs better than both audio only and visual only recognition.

Research from the literature found that speech recognition strongly depends on the correlation between the audio and visual signals that are used to enhance understanding between human and machine. For integration between audio and visual modalities, it can be grouped into three main approaches: feature fusion, modal fusion and decision fusion. Figure 3 shows the stages of different fusion allocated using HMM in AVSR system.

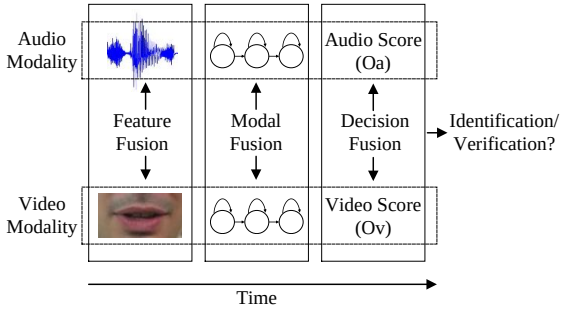


Figure 3: Three types of fusion techniques in AVSR system

A. Feature Fusion

Feature fusion as shown in Figure 4, can be defined as the apposition of audio and visual speech signals to be processed as a single observation for learning and classification [24]. This method assumes there is direct dependence in between audio and visual modalities. However, this feature fusion method suffers from two aspects. First, audio and visual features acquired having large number of information, so after the feature fusion the combined feature vector often become extremely large. Second, since this feature fusion is at the early stage of the whole speech recognition, so either the audio or visual modalities is corrupted, and so does the entire speech modality, thus this causes misclassification of the whole speech recognition.

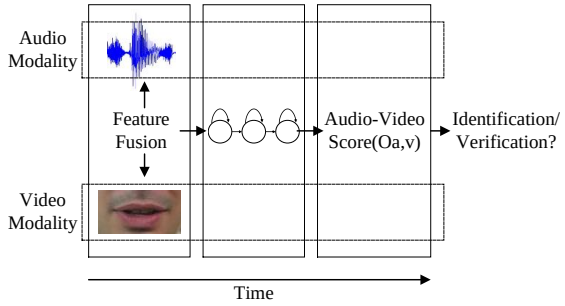


Figure 4: Feature fusion

B. Modal Fusion

Modal fusion as shown in Figure 5, is able to provide synchronization between two acoustic and visual modality and offers protection of corruption in either modality. It is a higher level integration than feature fusion. It integrates both modalities and then classifies them independently. The middle integration approach can be modeled by multi-stream Hidden Markov Models (HMMs) that utilizes two or more separate streams of audio and visual observation [24], as it delivers independence between streams statically with loose temporal dependence dynamically.

Streams can be integrated simply in the case of assuming it to be completely synchronous and characterized by HMMs with the similar topologies. However, it is not constantly synchronous all the time as they do not have the same frame rate. Modal fusion based approaches have been reviewed and proved to be able to achieve greater performance in continuous audio-visual speech recognition [25].

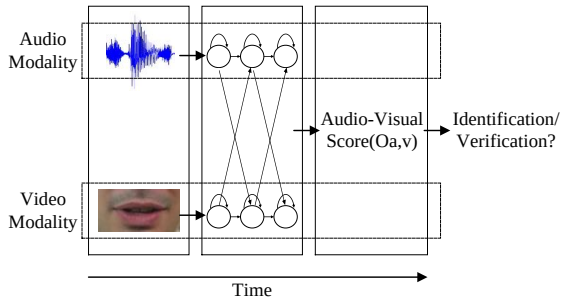


Figure 5: Modal fusion

C. Decision Fusion

The decision fusion shown in Figure 6 conceives whole independence between audio and visual speech modalities. There is no interaction between two modalities during the classification process, with only the final classifier scores being incorporate at the end. The output score can be combined easier than combining the feature vectors during early integration. Decision fusion technique fulfill the asynchronous classification of both modalities and is able to highlight the significance of a modality reliability on the corresponding quality of two signals, but it is unable to benefit from the correlation of both modalities at early integration stage [24]. Decision fusion generally takes place after the spoken utterance is completed, so this becomes the drawback of this approach and then results in the delay to generate the classification result and leads to unnatural interaction sessions [26].

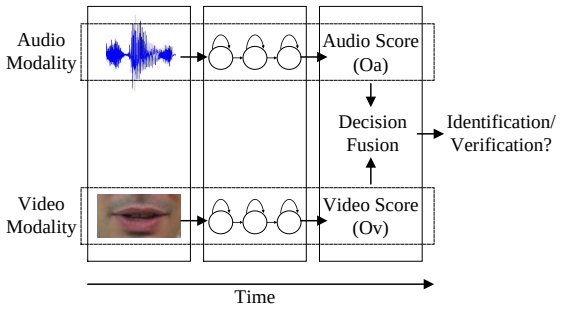


Figure 6: Decision fusion

V. Accuracy Estimation Methods

There are two general methods in evaluating the classifier's accuracy: the holdout method and the cross-validation. The holdout method is to divide all the samples randomly into two independent subsets (training set and test set). Training set is used for the training of classifier and the test set is used to verify and evaluate the performance of the trained classifier. Training set and test set will be divided into two-thirds (2/3) and one-thirds (1/3) respectively.

However, this method has a drawback of which, since this is a train-and-test experiment, thus the holdout estimation may be misleading if the training set samples contain corrupted data. The restrictions of the holdout method can be overcome by cross validation. Cross validation is a technique to estimate the results of a statistical analysis that sums up into an independent data. To reduce the variability, several turns of cross validation are performed to average the validation results. Cross validation is commonly partitioned into three categories, which are: random subsampling, K-fold cross-validation and leave one out cross-validation (LOOCV).

A. Random Subsampling

Random subsampling shown in Figure 7 performs K number data splits of dataset and each splits will randomly select a fixed number of example without any replacement. After that, classifier will be retrain with training examples and the results will be evaluated with test examples.

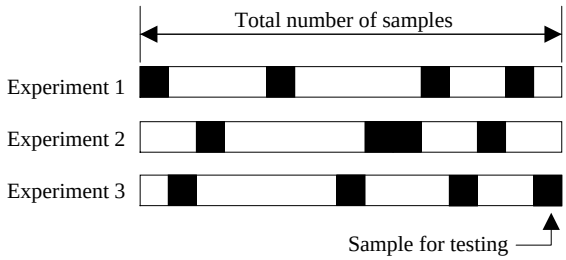


Figure 7: Illustration diagram of random subsampling

B. K-Fold Cross-Validation

K-fold cross validation is to split the dataset into k-equal parts. For the rest k-1 parts will be undergoing training and the one part that is left out will be for testing purposes. Based on Figure 8, it shows that this process will be repeated for k times by changing the test part one-by-one until all the other K parts are ultimately used for both training and testing.

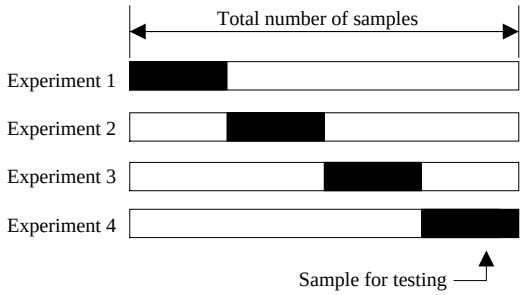


Figure 8: Illustration diagram of 4-fold cross validation

C. Leave-One-Out Cross Validation (LOOCV)

LOOCV is the extreme example of K-fold cross validation, where the K is equivalent to the total number of observations (N) as shown in Figure 9. So if the dataset is having N examples, the process will be repeated by N times. The classifier will be trained for N times using N-1 parts, and the one outstanding part will be used for testing purposes. In previous work [41], LOOCV has been proved by researchers as the utmost unbiased model validation technique.

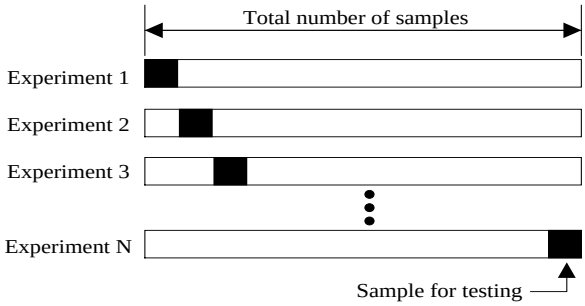


Figure 9: Illustration diagram of LOOCV

D. Bootstrap Validation

Bootstrap is a resampling technique with replacement. It randomly selects N samples (with replacement) and uses this set for training purposes and the remaining set are used for testing. Table 2 shows the replacement process for each experiment. The example below shows the complete set with samples X_1, X_2, X_3, X_4 and X_5. For experiment 2, once X_2 and X_4 are selected as test set, the rest X_1, X_3 and X_5 will be the train set with 2 samples repeated, where the train set now becomes X_1, X_3, X_3, X_5 and X_5.

Table 2
Illustration Sample Diagram of Bootstrap Validation

Experiment Set	Training Set	Testing Set
Set 1	X1 X2 X2 X4 X5	X3
Set 2	X1 X3 X3 X5 X5	X2 X4
Set 3	X2 X2 X2 X4 X5	X1 X3
	.	
	.	
	.	
Set K	X1 X3 X3 X3 X3	X2 X4 X5

**Complete dataset = X1, X2, X3, X4, X5

VI. Summary and Discussion

In this review paper, the brief overview of basic techniques for AVSR for the past twenty years have been discussed. There are some issues relevant to the training and testing of the AVSR system, such as the existing audio-visual dataset used for experimental testing, integration techniques used for audio and visual modality and accuracy estimation methods.

There are some common limitations that is faced by audio-visual data corpus. For instance, previous speech data corpus, which is often very poor in quality, especially for video data. Since nowadays camera technology has improved greatly, thus capturing videos with high resolution is not an issue anymore. Speech database built must have reasonable number of respondents which represent the generality of the results. Even with around 200-300 respondents, their age, gender, race and dialect should be recorded carefully, in case there is a bias from the database related to the unbalanced of age, gender and race of respondents.

Besides, there are also some challenges while performing the fusion of audio and visual modality during speech recognition. There is a problem in estimating the weight for each modality under varying conditions, as the effectiveness of each modality will vary in different environments. Synchronization between the audio and video stream is another issue, which the acoustic noise and visual feature does not essentially happen exactly at the same time. Handling asynchrony between audio and visual modality is a serious problem in real-world applications and more works should be done to discuss and assess this issue properly in the future.

As mention earlier, there are some validation methods addressed in Section 5. According to the research [50], bias and variance were being investigated to test each validation techniques. The paper concluded that the out-of-sample bootstrap validation yield the least biased result compare with other validation methods. It is stated that the hold-out validation tend to produce results with more bias and variance, however, this method is still widely used in real-world applications due to it being computationally cheaper compare with k-fold validation and other methods. So theoretically, out-of-sample bootstrap method is more precise and generalization error better than the rest, but practically its training process is time consuming and computationally expensive.

APPENDIX

Table 3
Performance of NU-InNets and NU-ResNets: Tested with THFOOF-50 Dataset

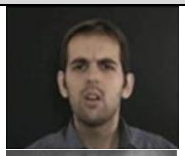
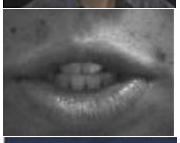
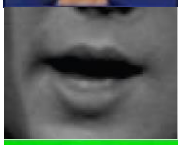
Database	Subject	Audio Quality (Sample Rate)	Video Quality (Resolution)	Recent Work	Accessibility	Typical Image
M2VTS	25 males, 12 females	48kHz, 16bits Controlled audio	286 x 350, 25fps Passport view	[27, 28]	Yes	
TULIPS1	7 males, 5 females	11.1kHz, 8bits Controlled audio	100 x 75, 8bits, 30fps Mouth region	[29, 30]	Yes	
VidTIMIT	24 males, 19 females	32kHz, 16bits Controlled audio	512 x 384, 24bits, 25fps Upper body	[31, 32]	Yes	
CUAVE	19 males, 17 females	44kHz, 16bits Controlled audio	720 x 480, 24bits, 29.97fps Passport view	[29, 33, 34]	Yes	
XM2VTS	295 (unknown gender)	32KHz, 16bits Controlled audio	720 x 576, Passport view	[35 – 38]	Yes	
AVletters	5 males, 5 females	22kHz, 16bits Controlled audio	80 x 60, 8 bits, 25fps Mouth region	[39, 40]	Yes	
LUNA-V	9 males, 1 female	16kHz, 16bits Controlled audio	1280 x 720, 25fps, Passport view	[23]	Yes	

Table 4
Summary of Recent Works on Publicly Available AVSR Speech Database

No	Database	Year	Feature Extraction Technique	Classification	Training and Testing	Task	Integration technique	Result accuracy	Ref.
1	M2VTS	2005	LDA - PCA	MHMM	Train : 75 Test : 25	Speaker recognition	Modal	96.57 %	[42]
2	XM2VTS	2014	DCT - MFCC	MSHMM	Train : 200 Test : 95	Digit recognition	Modal	≈ 89 %	[43]
3	CUAVE	2015	MFCC	HMM	Train : 60 Test : 40	Digit recognition	Feature	94 %	[26]
4	CUAVE	2014	MFCC	HMM	n/a	Digit recognition	-	95 %	[44]
5	CUAVE	2013	DCT-MFCC	DBN	Train : 70 Test : 37	Digit recognition	Feature	WER = 1.4	[45]
6	VidTIMIT	2010	DCT-MFCC	GMM (single-state of HMM)	Train : 344 Test : 86	Person recognition	Hybrid feature-decision	EER = 5.23	[46]
7	Tulips1	2010	LDB	HMM-SVM	Left-one-out CV	Speech recognition	-	EER = 1.74	[30]
8	GRID	2013	MFCC	CHMM	n/a	Speech recognition	Modal	96.37 %	[47]
9	GRID	2014	RASTA - PLP	CHMM	Train : 800 Test : 200	Speech recognition	Modal	96.7 %	[48]
10	LUNA-V	2014	HSV colour filter + border following + convex hull technique - MFCC	HMM	Train : 30 Test : 20	Digit recognition	-	92.5 % (Visual-only)	[49]

ACKNOWLEDGMENT

This work was supported by Universiti Malaysia Pahang and funded by the Ministry of Higher Education Malaysia under FRGS Grant RDU160108.

REFERENCES

[1] Sawakare, P. A., Deshmukh, R. R. & Shrishrimal, P. P. Speech Recognition Techniques: A Review. 6 (2015), 1693–1698.
[2] Morgan, N. Deep and wide: Multiple layers in automatic speech recognition. IEEE Trans. Audio, Speech Lang. Process. 20 (2012), 7–13.
[3] Ghadage, Y. H. & Shelke, S. D. Speech to Text Conversion for Multilingual Languages (2016), 236–240.
[4] Kulkarni, D. S., Deshmukh, R. R., Shrishrimal, P. P., Waghmare, S. D. & Science, C. HTK Based Speech Recognition Systems for Indian Regional languages : A Review. (2016).
[5] Tian, C., Ji, W. & Yuan, Y. Auxiliary Multimodal LSTM for Audio-visual Speech Recognition and Lipreading(2017), 1–9.
[6] Islam, M. & Rahman, F. Hybrid Feature and Decision Fusion Based Audio-Visual Speaker Identification in Challenging Environment. Int. J. Comput. Appl. 9 (2010), 9–15.
[7] Potamianos, G., Neti, C., Luettin, J. & Matthews, I. Audio-Visual Automatic Speech Recognition : An Overview. Issues Vis. Audiov. Speech Process (2004), 1–30.
[8] Galatas, G., Potamianos, G. & Makedon, F. Audio-visual speech recognition incorporating facial depth information captured by the Kinect (2012), 2714–2717.
[9] Navarathna, R., Dean, D., Sridharan, S. & Lucey, P. Multiple cameras for audio-visual speech recognition in an automotive environment. Computer Speech and Language 27 (2013), 911–927.
[10] Palecek, K. & Chaloupka, J. Audio-visual speech recognition in noisy audio environments. 2013 36th Int. Conf. Telecommun. Signal Process (2013), 484–487.
[11] Ibrahim, M. Z. & Mulvaney, D. J. Robust geometrical-based lip-reading using hidden Markov models. IEEE EuroCon (2013), 2011–2016.
[12] Ibrahim, M. Z. & Mulvaney, D. J. A lip geometry approach for feature-fusion based audio-visual speech recognition. ISCCSP 2014 - 2014 6th Int. Symp. Commun. Control Signal Process. Proc (2014), 644–647.
[13] Oliveira, V. A. & Conci, A. in H. Pedrini, & J. Marques de Carvalho, Workshops of Sibgrapi (2009), 1–2.
[14] Dave, N. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. Int. J. Adv. Res. Eng. Technol. 1 (2013), 1–5.
[15] Ittichaichareon, C. Speech recognition using MFCC. Conf. Computer (2012), 135–138.
[16] Hongbing Hu, Stephen. A, Z. Dimensionality reduction methods for HMM phonetic recognition (2010), 4854–4857.
[17] Mohamed, A. et al. Deep belief networks using discriminative features for phone recognition. Acoust. Speech Signal Process. (ICASSP), IEEE Int. Conf (2011). 5060–5063.
[18] Shrawankar, U. & Thakare, V. Feature Extraction for a Speech Recognition System in Noisy Environment: A Study. Comput. Eng. Appl. (ICCEA), Second Int. Conf. 1 (2010), 358–361.
[19] Tripathy, S., Baranwal, N. & Nandi, G. C. A MFCC based Hindi speech recognition technique using HTK Toolkit. 2013 IEEE 2nd Int. Conf. Image Inf. Process. IEEE ICIIP (2013), 539–544.
[20] Luettin, J., Thacker, N. a. & Beet, S. W. Visual speech recognition using active shape models and hidden Markov models. IEEE Int. Conf. Acoust. Speech, Signal Process. 2 (1996), 817–820.
[21] Matthews, I. Features for audio-visual speech recognition. Citeseer (1998).
[22] Patterson, E. K., Gurbuz, S., Tufekci, Z. & Gowdy, J. N. CUAVE: A new audio-visual database for multimodal human-computer interface research. IEEE Int. Conf. Acoust. Speech, Signal Process.2 (2002), II-2017-II-2020.
[23] M. Z. Ibrahim, "A novel lip geometry approach for audio-visual speech recognition," Loughborough University (2014).
[24] Katsaggelos, A. K., Bahaadini, S. & Molina, R. Audiovisual Fusion: Challenges and New Approaches. Proc. IEEE 103 (2015), 1635–1653.
[25] Huang, P. Sen, Zhuang, X. & Hasegawa-Johnson, M. Improving acoustic event detection using generalizable visual features and multi-modality modeling. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc (2011). 349–352.
[26] Ibrahim, M. Z., Mulvaney, D. J. & Abas, M. F. Feature-fusion based audio-visual speech recognition using lip geometry features in noisy enviroment. 10 (2015), 17521–17527.
[27] Sarvestani, R. R. & Boostani, R. FF-SKPCCA: Kernel probabilistic canonical correlation analysis. Appl. Intell (2016). 438–454.
[28] Saeed, U. Person identification using behavioral features from lip motion. 2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG (2011), 155–160.
[29] Morade, S. S. & Patnaik, S. Comparison of classifiers for lip reading with CUAVE and TULIPS database. Optik (Stuttg). 126 (2015), 5753–5761.
[30] Kambiz Rahbar. Independent-Speaker Isolated Word Speech Recognition Based on Mean-Shift Framing Using Hybrid HMM/SVM Classifier (2010). 156–161.
[31] Makrem, B. Structuring Visual Information for Person Detection in Video : Application to VIDTIMIT database (2016), 233–237.
[32] Soto, P. et al. Single Sample Face Recognition from Video via Stacked Supervised Auto-encoder Single Sample Face Recognition from Video via Stacked Supervised Auto-encoder (2016).
[33] Morade, S. S. Visual Lip Reading using 3D-DCT and 3D-DWT and LSDA. 136 (2016), 7–15.
[34] Foteini Patrona, Alexandros Iosifidis, Anastasios Tefas, Nikolaos Nikolaidis, and I. P. Visual voice activity detection based on spatiotemporal information and bag of words. Int. Conf. Image Process (2015). 2334–2338.
[35] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S. & Pantic, M. A semi-automatic methodology for facial landmark annotation. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work (2013), 896–903.
[36] Naser Damer, Alexander Opel, A. N. Biometric source weighting in multi-biometric fusion : towards a generalized and robust solution (2013).
[37] Ouamane, A., Messaoud, B., Guessoum, A., Hadid, A. & Cheriet, M. Multi scale multi descriptor local binary features and exponential discriminant analysis for robust face authentication. IEEE Int. Conf. Image Process. ICIP (2014). 313–317.
[38] Li, Z., Imai, J. I. & Kaneko, M. Face and expression recognition based on bag of words method considering holistic and local image features. Isc. 10th Int. Symp. Commun. Inf. Technol (2010). 1–6.
[39] Petridis, S. & Pantic, M. Deep complementary bottleneck features for visual speech recognition. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc (2016), 2304–2308.
[40] Frisky, A. Z. K., Wang, C.-Y., Santoso, A. & Wang, J.-C. Lip-based visual speech recognition system. Secur. Technol. (ICCST), 2015 Int. Carnahan Conf (2015). 315–319.
[41] Kocaguneli, E. & Menzies, T. Software effort models should be assessed via leave-one-out validation. J. Syst. Softw. 86 (2013), 1879–1890.
[42] Lucey, S., Chen, T., Sridharan, S. & Chandran, V. Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition. IEEE Trans. Multimed. 7 (2005), 495–506.
[43] Stewart, D., Seymour, R., Pass, A. & Ming, J. Robust audio-visual speech recognition under noisy audio-video conditions. IEEE Trans. Cybern. 44 (2014), 175–184.
[44] Pawar, G. S. Realization of Hidden Markov Model for English Digit Recognition. 98 (2014), 98–101.
[45] Huang, J. & Kingsbury, B. Audio-visual deep learning for noise robust speech recognition. IEEE Int. Conf. Acoust. Speech Signal Process (2013). 7596–7599.
[46] Shah, D., Han, K. J. & Narayanan, S. S. Robust Multimodal Person Recognition Using Low-Complexity Audio-Visual Feature Fusion Approaches. Int. J. Semant. Comput. 4 (2010), 155–179.
[47] Ahmed Hussen Abdelaziz, Steffen Zeiler, D. K. Twin-HMM-based audio-visual speech enhancement. Digit. Signal Process (2013), 3726–3730.
[48] Receveur, S., Scheler, D. & Fingscheidt, T. A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition (2014), 179–192.
[49] Ibrahim, Z. A novel lip geometry approach for audio-visual speech recognition (2014).
[50] Tantithamthavorn, C., Mcintosh, S., Hassan, A. E. & Matsumoto, K. An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. IEEE Trans. Softw. Eng. 5589 (2016), 1–16.