FEATURE-FUSION BASED AUDIO-VISUAL SPEECH RECOGNITION USING LIP GEOMETRY FEATURES IN NOISY ENVIROMENT

M. Z. Ibrahim[1], D. J. Mulvaney[2] and M. F. Abas[1]
[1]Faculty of Electrical and Electronics Engineering, University Malaysia Pahang, Pahang, Malaysia
[2]School of Electronic, Electrical and Systems Engineering, Loughborough University, United Kingdom
E-Mail: zamri@ump.edu.my

ABSTRACT

Humans are often able to compensate for noise degradation and uncertainty in speech information by augmenting the received audio with visual information. Such bimodal perception generates a rich combination of information that can be used in the recognition of speech. However, due to wide variability in the lip movement involved in articulation, not all speech can be substantially improved by audio-visual integration. This paper describes a feature-fusion audio-visual speech recognition (AVSR) system that extracts lip geometry from the mouth region using a combination of skin color filter, border following and convex hull, and classification using a Hidden Markov Model. The comparison of the new approach with conventional audio-only system is made when operating under simulated ambient noise conditions that affect the spoken phrases. The experimental results demonstrate that, in the presence of audio noise, the audio-visual approach significantly improves speech recognition accuracy compared with audio-only approach.

Keywords: lip geometry, feature fusion, audio-visual speech recognition, OpenCV.

INTRODUCTION

People with hearing impairments successfully used visual lip movements to aid the understanding of speech, promises the potential of being able to improve the robustness of automatic speech recognition in environments where substantial audible ambient noise is present. Studies available in the literature have shown a close correlation between the information present in lip movements and speech signals, and consequently the addition of visual information has been a line of investigation followed by a number of researchers in their efforts to improve machine perception of the spoken word [1].

Whenever two modalities are to be considered jointly, the question arises as to the processing stage at which the modalities' information content should be fused. In the case of speech and visual speech modalities, fusion can take place either at the feature level (often termed early integration) or the decision level (often termed late integration).

Decision fusion

In the decision fusion approach, recognition is performed separately for each of the modalities, with the partial results from each sub-process being combined to produce the final classification [2]. As the models may often deliver different partial classification decision outcomes, the decision-fusion approach must provide a suitable method for their ranking and collation. A major drawback of this approach is that fusion itself normally only takes place after the complete utterance has been recognized, which, compared to the feature-fusion approach, can lead to a delay in generating the classification result and so make interactive sessions appear unnatural. The main advantage of this approach is that each of the classifications performed is specific to that modality, allowing specifically tailored methods to be adopted.

Feature fusion

In this paper, the feature fusion approach is adopted and the features extracted for each modality are combined into a common vector to be used by the recognition system. A main drawback with this approach is that, due to the large number of visual features often acquired, the combined feature vector often becomes considerably longer. The advantage of this type of fusion is the straightforward extension of techniques already developed for audio-only speech recognition to include the visual aspects, although in a practical implementation, the modalities need to be synchronized and interpolation used to correct for the different frame rates. In order to reach satisfactory convergence, a substantial increase is required in both the number of training vectors and the training time of the recognition models. In the literature, this problem is known as 'the curse of dimensionality' [3], [4]. The contribution of this paper is twofold. Firstly, experiments have been carried out to demonstrate that the geometrical features established in this work have information content that is highly relevant for the recognition task when classification is carried out in the presence of acoustic noise. Secondly, the investigation of a propose system applied to digit recognition under noisy conditions is performed and the results showed that the recognition of individual digits exhibits a performance that depends substantially on the magnitude of the movement of the lips required in its articulation.

METHODOLOGY

The software used in this work was developed using Microsoft Visual C# 2010 [5] and utilized both the open source image processing library, OpenCV [6] and the Hidden Markov Model Toolkit (HTK) speech processing library [7]. The components of the geometric approach for AVSR are shown in Figure-1. The system can be divided into three phases, namely feature extraction, integration and classification.

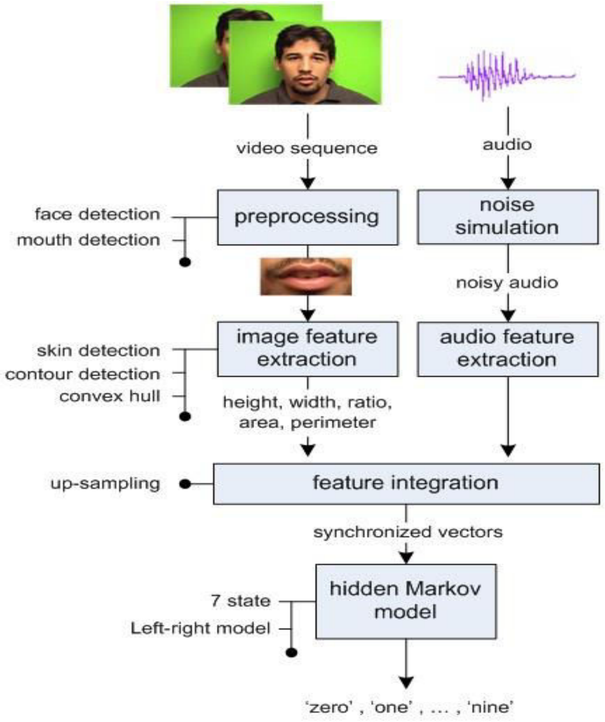


Figure-1. A block diagram for geometrical-based AVSR system.

Visual feature extraction

Speaker images acquired from the video files of the database are cropped to the mouth region by applying a face detection process followed by a mouth detection process using the Viola-Jones object recognizer [8], as shown in Figure-2(a). This approach forms a product of Haar-like operators at each image location and at several scales and the results are sent to train a weak classifier using AdaBoost. Calculation time can effectively be halved by assuming that the mouth is located in the lower half of the face, which also reduces the risk of false detection that can follow from the inadvertent classification of an eye as the mouth.

A skin detection technique is then used to segment the lip and non-lip areas in the mouth region as shown in Figure-2(b). This work has adopted the HSV color model for segmentation as this model comes closest to mimicking how humans perceive skin color [9]. The segmented image containing the lip region is then converted to a binary format and contour extraction is achieved using border following [10]. This generates a collection of contours with a range of sizes and shapes distributed along the contour of the binary image, as shown in Figure-2(c). Finally, the lip geometry features, height, width, ratio (of height to width), area and the perimeter, are extracted by applying the convex hull method as shown in Figure-2(d).

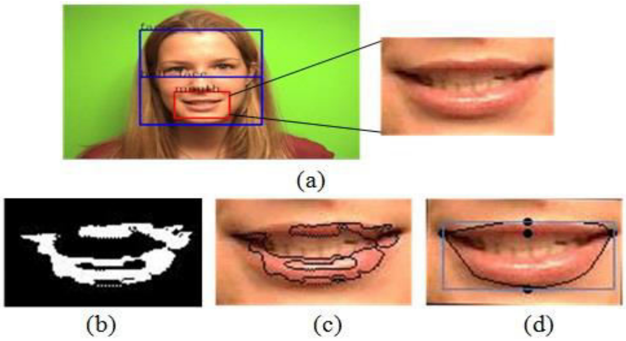


Figure-2. Visual feature extraction showing (a) face and mouth detection, (b) skin detection based on HSV color filter, (c) series of contours generated based on the border following method, (d) convex polygon generated.

Audio feature extraction

Several results have been reported in the literature regarding the audio feature extraction techniques [11]. Mel-frequency cepstral coefficients (MFCCs) [12] and linear prediction coefficients (LPCs) [13] represent the most commonly used audio features in last few decades. There are still on-going researches in the field of robust audio features and such features will not be considered in this work.

MFCCs is very popular and has been shown to outperform others feature extraction techniques as revealed in [12]. MFCCs are derived from a Mel-frequency where this frequency axis is warped according to the Mel-scale, which approximate the human auditory system's response. The dynamic features which are first (delta-MFCCs) and second time-derivatives (delta-delta-MFCCs) of cepstral coefficients is now commonly employed to improve speech recognition performance [14], [15].

Audio-visual integration

The HTK library was used to extract the MFCC features and their first and second derivatives, resulting in a feature vector of 39 dimensions. To achieve feature integration, the visual and audio feature extraction rates must be equalized. In the current work, the visual feature rate is the video frame rate of 29.97Hz, whereas the audio MFCC feature rate is 100Hz. Equalization involved linear interpolation of the visual features to match the audio frame rate. Finally the audio and visual features are combined and the resulting synchronized feature vector of dimensionality 44 (39 audio features and five visual features) is used for training and testing. This process is shown in Figure-3.

Classification using hidden markov models (HMMs)

HMMs are a popular and successful approach to the statistical modelling of audible speech [16]. In this work, 7-state left-to-right models were used with each Markov state being modelled using solely Gaussian functions with diagonal covariance. The models are trained using the Baum-Welch algorithm [11], [17] with recognition performed using the Viterbi algorithm [11], [17].

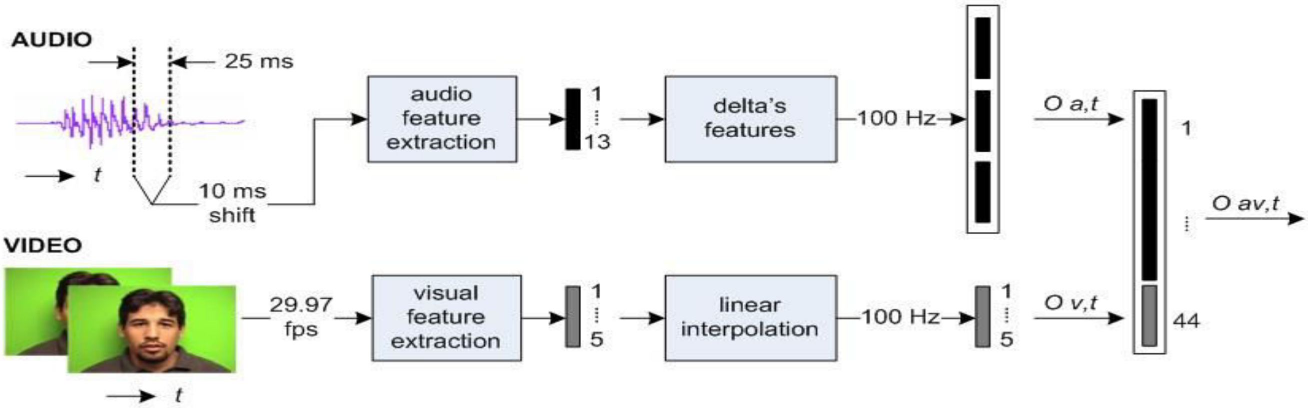


Figure-3. Block diagram of the feature fusion for AVSR. The algorithm generates time-synchronous 39-dimensional audio $O_{a,t}$ and 5-dimensional visual feature $O_{v,t}$ vectors at 100 Hz rate.

RESULTS AND DISCUSSION

This section presents the results obtained for the geometrical-based AVSR system and their comparison with audio-only approach. Both systems are also exposed to simulated variations in environmental conditions that arise from the introduction of audible noise.

Data corpus

In order to investigate the effectiveness of the proposed approaches in practical applications, the CUAVE corpus database was used to provide examples of speech and video sequences of the speakers [18]. The database consists of 7000 utterances of connected and isolated digits spoken by 36 individuals, where 19 speakers are male and the remainder are female. The speakers also have a range of skin and lip tones as well as a variety of face and lip shapes and a number of the subjects wore additional visual items such as glasses, facial hair, and hats. Lighting was controlled and a green background was employed to allow custom video backgrounds to be added using chroma-keying if required. The video sequences were recorded at a resolution of 720 x 480 in MPEG-2 format at 29.97 frames/s and encoded at a data rate of 5,000 kbit/s.

The CUAVE database consists of five sessions, in each of which the subject speaks the words 'zero' to 'nine'. In the investigations, data from sessions 1, 2 and 3 (30 samples) were employed for training and the data from sessions 4 and 5 (20 samples) were used for testing. All 36 speakers in the database were investigated in this work, making a total of 1800 samples for use in demonstrating the utility of the new approach.

Performance under noise conditions

This section presents results obtained for an AVSR system that utilizes lip geometry information in an attempt to improve the speech recognition rate in noisy environments. The experiments conducted under the influence of 'babble', 'factory1', 'factory2' and 'white' noise. These types of noise are part of NOISEX-92 dataset [19] and have been added to the speech signals obtained from CUAVE data corpus, such that specific signal-to-noise ratios (SNRs) are attained.

Figure-4 shows the performance of the geometrical-based AVSR system in terms of word accuracy rate as a function of SNR. The audio signal was disturbed by adding 'babble noise' from the NOISEX-92 database (100 people speaking in a canteen) provided at a range of intensities such that the SNR lies in the interval 25 to -10dB. It can clearly be seen that as the noise level increases (here specifically at SNRs below 10dB), using combined information from the audio and visual modalities gives the best classification performance of the tests executed. For instance, at an SNR of 0dB the improvement in performance is more than 25% compared with the corresponding audio-only figure.

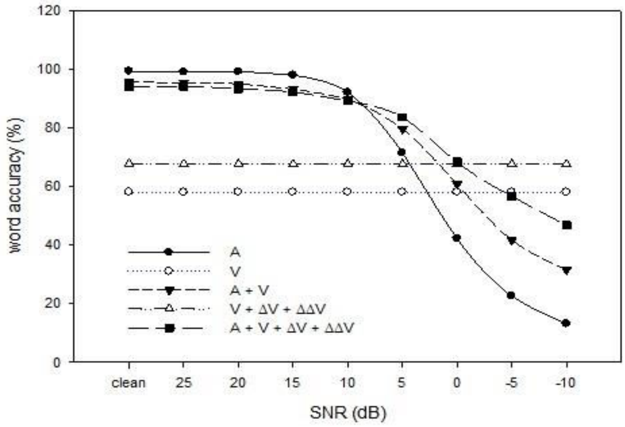


Figure-4. AVSR system performance using geometrical features when 'babble noise' is applied. Shown are the noisy audio (A), the visual only information (V), dynamic visual information with delta and delta-delta features (V + ΔV + ΔΔV), the combination of audio and visual (A + V) features and the combination audio and visual with delta and delta-delta features (A + V + ΔV + ΔΔV).

The classification performance provided by the visual modality was found to be improved if motion information is incorporated by the addition of the first-order (delta) and second-order difference (delta-delta) geometrical features. Such features are commonly used in audio speech recognition and Figure-4 shows the improvements in both audio and visual recognition that results from the inclusion of the difference features; for example at an SNR of -5dB the performance can be seen to have been improved by more than 30% with respect to audio-only recognition.

Further tests of the geometrical-based AVSR system were carried out using 'factory1', 'factory2' and 'white noise' datasets from NOISEX-92. Factory1 noise was recorded near to plate-cutting and electrical welding equipment, while factory2 noise was recorded in a car production hall. White noise was acquired by sampling a high-quality analog noise generator operating in the range 0 Hz to 16 kHz. The results obtained for these three types of noise were similar to those of the babble noise, where the combined audio-visual performance at a SNR of -5dB improved by more than 30%.

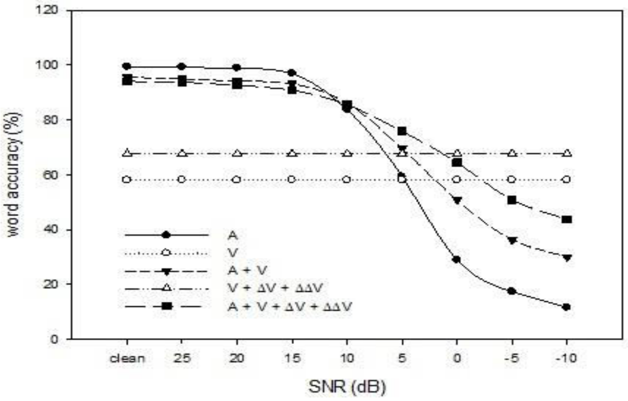


Figure-5. AVSR system performance using geometrical features when 'factory1 noise' is applied.

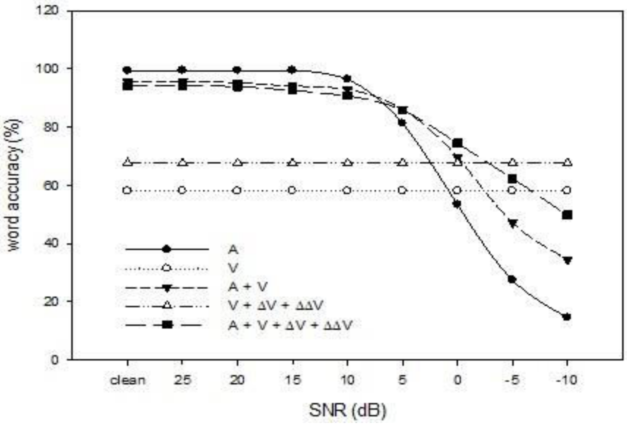


Figure-6. AVSR system performance using geometrical features when 'factory2 noise' is applied.

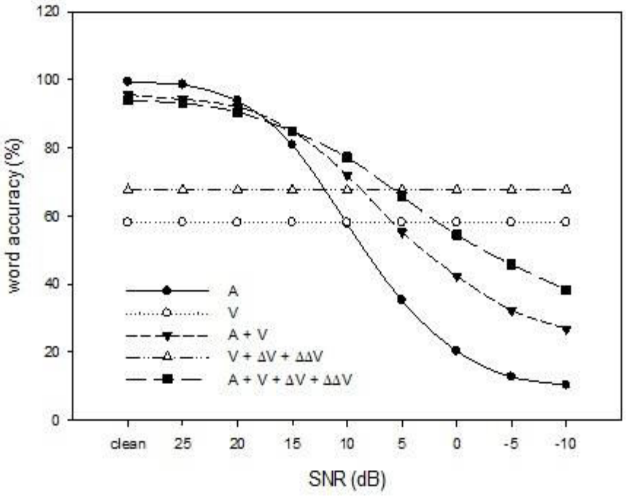


Figure-7. AVSR system performance using geometrical features when 'white noise' is applied.

Digit performance analysis

Due to wide variability in the lip movement involved in articulation, not all English digit recognition can be substantially improved by audio-visual integration. For example, when pronouncing the word 'six', only small movements of the lips are involved, while in the production of the word 'seven' considerable lip movements are required. In general, the greater the lip movements required to generate the word, the better an AVSR system is likely to perform. Figure-8 and Figure-9 show the recognition performance of the new system in identifying the digit 'seven' when simulated under 'white' and 'babble' noise. The graphs show that the performance when using only visual information is 75% and the combination of the audio-visual information improved the performance by more than 40% relative the audio-only results at SNRs below 0dB.

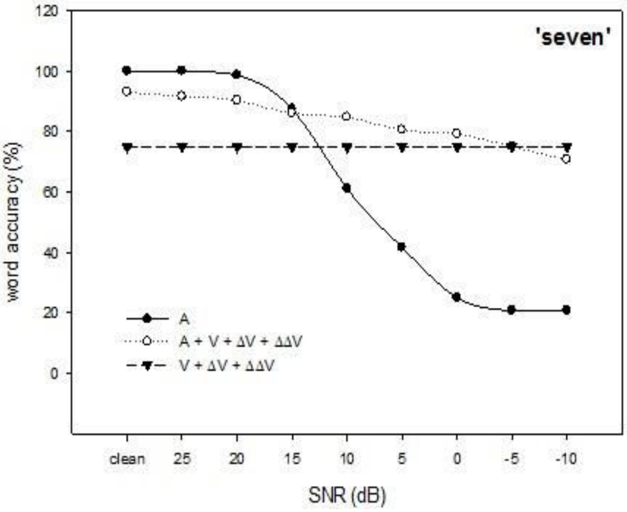


Figure-8. AVSR system performance for digit 'seven' using geometrical features when 'white noise' is applied.

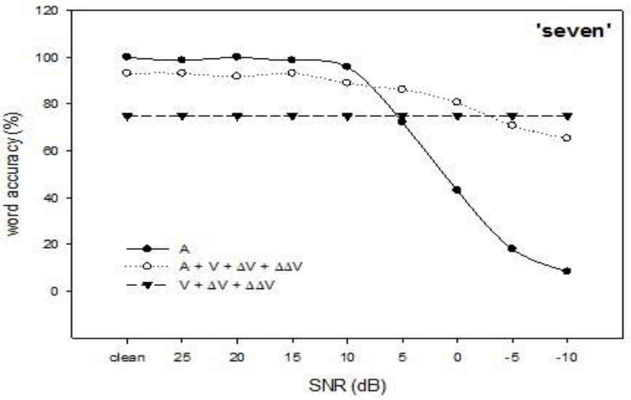


Figure-9. AVSR system performance for digit ‘seven’ using geometrical features when ‘babble noise’ is applied.

The contrast in performance for the audio-visual recognition of the digit ‘six’ can be seen in Figure-10 and Figure-11, again when ‘white’ and ‘babble’ noise are added. The recognition performance using video can be seen to be 50%, such poor performance being due to the minimal movement of the lip when ‘six’ is pronounced. The audio-visual results are also adversely affected, with only up to a 30% improvement compared to audio-only recognition for SNRs less than 0dB.

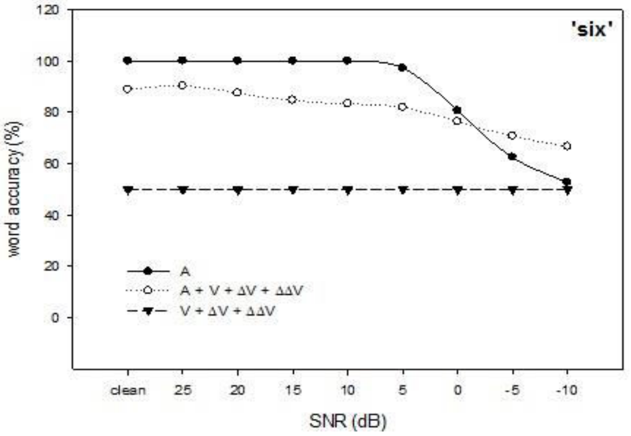


Figure-10. AVSR system performance for digit ‘six’ using geometrical features when ‘white noise’ is applied.

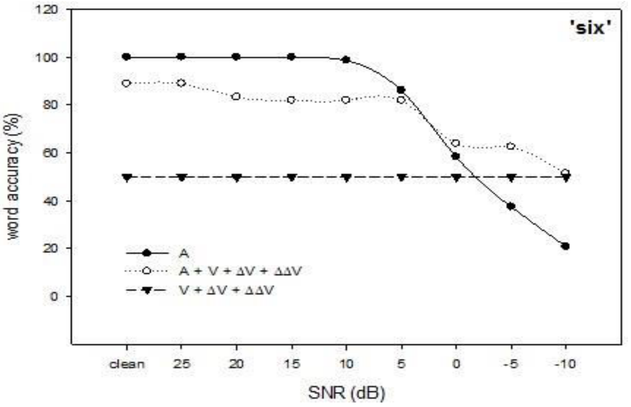


Figure-11. AVSR system performance for digit ‘six’ using geometrical features when ‘babble noise’ is applied.

Digit confusion analysis

To understand further the digit recognition of the new geometrical-based AVSR system, a confusion analysis is now described in order to identify under what circumstances misclassification occurs.

Table-1 shows the confusion matrix for the audio-only speech recognition and Table-2 for the audio-visual system, including lip geometry features and its derivative at 0dB SNR with white noise added. The white noise was again obtained from the NOISEX-92 database. As seen in confusion matrix in Table-1, the similarity between white noise and digit six will bias the recognizer in favour of this digit. This is because digit six begins and ends with sibilant sounds. The accuracy of the digit six classifying correctly is the highest among other achieving up to 80%.

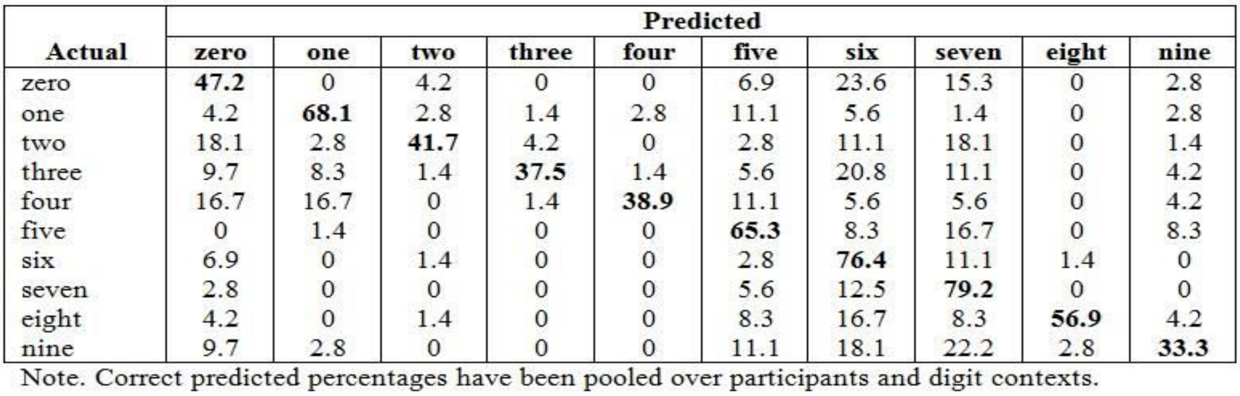
By combining the visual and speech information in noisy environments, the new system achieved better performance than audio alone as shown in Table-2. It can be seen that the recognition of all digits improves considerably, except digit six which worsens slightly by 4.2%. The substantial differences in the results for digits ‘one’ and ‘seven’ are probably due to the greater movement of the lips required in the articulation; in contrast digit ‘six’ and ‘nine’ that require only small lip movements in their articulation, showed only small changes in recognition performance.

Table-1. Confusion matrix for audio only recognition at 0 dB SNR using white noise.

	Predicted									
Actual	**zero**	**one**	**two**	**three**	**four**	**five**	**six**	**seven**	**eight**	**nine**
zero	**20.8**	2.8	0	0	2.8	4.2	45.8	18.1	5.6	0
one	15.3	**4.2**	1.4	0	5.6	9.7	26.4	19.4	8.3	9.7
two	8.3	2.8	**5.6**	1.4	0	1.4	59.7	13.9	5.6	1.4
three	8.3	2.8	2.8	**2.8**	0	4.2	59.7	11.1	5.6	2.8
four	18.1	2.8	0	0	**4.2**	5.6	44.4	16.7	6.9	1.4
five	8.3	4.2	0	0	4.2	**25.0**	26.4	20.8	5.6	5.6
six	2.8	0	0	0	0	0	**80.6**	12.5	4.2	0
seven	5.6	1.4	0	1.4	1.4	8.3	51.4	**25.0**	5.6	0
eight	4.2	1.4	6.9	2.8	0	0	62.5	8.3	**12.5**	1.4
nine	9.7	2.8	0	0	4.2	8.3	25.0	20.8	6.9	**22.2**

Note. Correct predicted percentages have been pooled over participants and digit contexts.

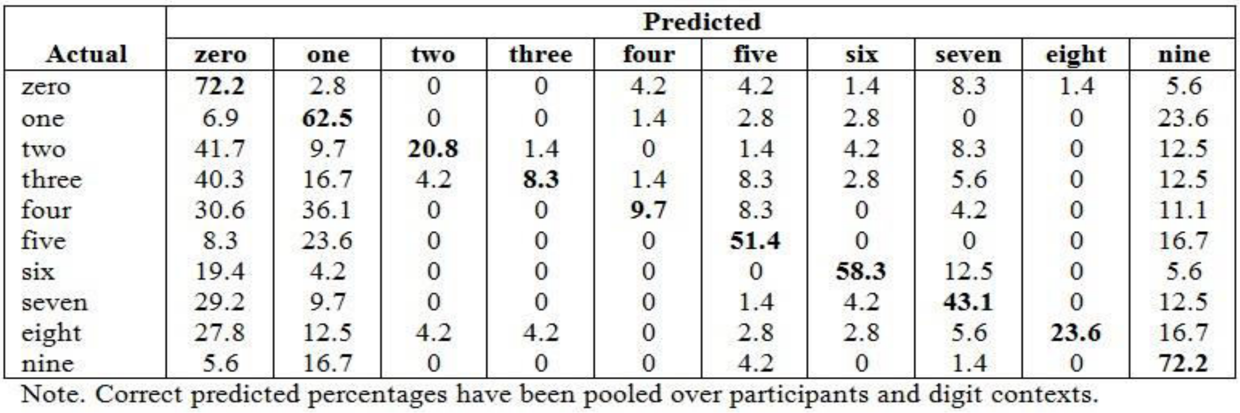
Table-2. Confusion matrix for audio-visual recognition at 0dB SNR using white noise.



Actual	Predicted									
	zero	**one**	**two**	**three**	**four**	**five**	**six**	**seven**	**eight**	**nine**
zero	**47.2**	0	4.2	0	0	6.9	23.6	15.3	0	2.8
one	4.2	**68.1**	2.8	1.4	2.8	11.1	5.6	1.4	0	2.8
two	18.1	2.8	**41.7**	4.2	0	2.8	11.1	18.1	0	1.4
three	9.7	8.3	1.4	**37.5**	1.4	5.6	20.8	11.1	0	4.2
four	16.7	16.7	0	1.4	**38.9**	11.1	5.6	5.6	0	4.2
five	0	1.4	0	0	0	**65.3**	8.3	16.7	0	8.3
six	6.9	0	1.4	0	0	2.8	**76.4**	11.1	1.4	0
seven	2.8	0	0	0	0	5.6	12.5	**79.2**	0	0
eight	4.2	0	1.4	0	0	8.3	16.7	8.3	**56.9**	4.2
nine	9.7	2.8	0	0	0	11.1	18.1	22.2	2.8	**33.3**

Note. Correct predicted percentages have been pooled over participants and digit contexts.

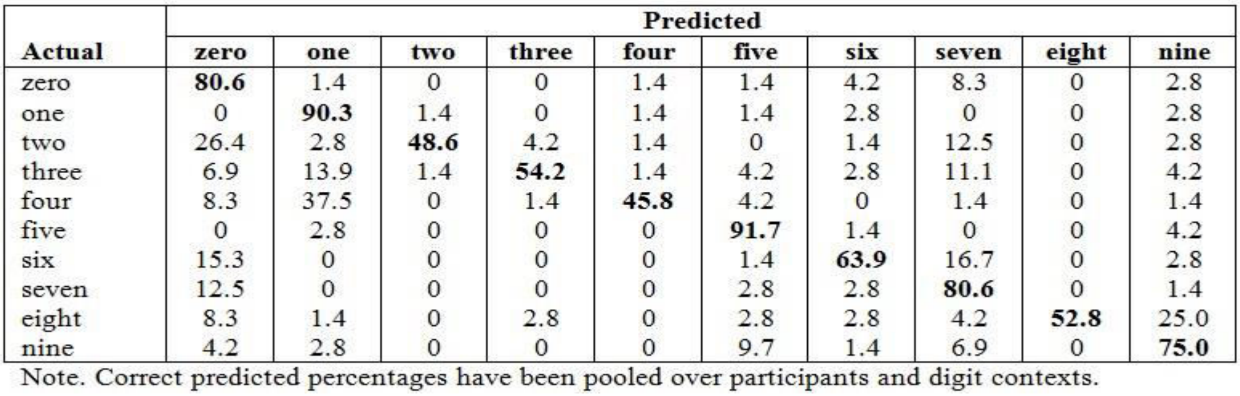
Table-3. Confusion matrix for audio only recognition at 0dB SNR using babble noise.



Actual	Predicted									
	zero	**one**	**two**	**three**	**four**	**five**	**six**	**seven**	**eight**	**nine**
zero	**72.2**	2.8	0	0	4.2	4.2	1.4	8.3	1.4	5.6
one	6.9	**62.5**	0	0	1.4	2.8	2.8	0	0	23.6
two	41.7	9.7	**20.8**	1.4	0	1.4	4.2	8.3	0	12.5
three	40.3	16.7	4.2	**8.3**	1.4	8.3	2.8	5.6	0	12.5
four	30.6	36.1	0	0	**9.7**	8.3	0	4.2	0	11.1
five	8.3	23.6	0	0	0	**51.4**	0	0	0	16.7
six	19.4	4.2	0	0	0	0	**58.3**	12.5	0	5.6
seven	29.2	9.7	0	0	0	1.4	4.2	**43.1**	0	12.5
eight	27.8	12.5	4.2	4.2	0	2.8	2.8	5.6	**23.6**	16.7
nine	5.6	16.7	0	0	0	4.2	0	1.4	0	**72.2**

Note. Correct predicted percentages have been pooled over participants and digit contexts.

Table-4. Confusion matrix for combination of audio and visual at 0 dB SNR using babble noise.



Actual	Predicted									
	zero	**one**	**two**	**three**	**four**	**five**	**six**	**seven**	**eight**	**nine**
zero	**80.6**	1.4	0	0	1.4	1.4	4.2	8.3	0	2.8
one	0	**90.3**	1.4	0	1.4	1.4	2.8	0	0	2.8
two	26.4	2.8	**48.6**	4.2	1.4	0	1.4	12.5	0	2.8
three	6.9	13.9	1.4	**54.2**	1.4	4.2	2.8	11.1	0	4.2
four	8.3	37.5	0	1.4	**45.8**	4.2	0	1.4	0	1.4
five	0	2.8	0	0	0	**91.7**	1.4	0	0	4.2
six	15.3	0	0	0	0	1.4	**63.9**	16.7	0	2.8
seven	12.5	0	0	0	0	2.8	2.8	**80.6**	0	1.4
eight	8.3	1.4	0	2.8	0	2.8	2.8	4.2	**52.8**	25.0
nine	4.2	2.8	0	0	0	9.7	1.4	6.9	0	**75.0**

Note. Correct predicted percentages have been pooled over participants and digit contexts.

Further tests of the shape-based AVSR system were carried out using the 'babble' noise datasets from NOISEX-92. Table-3 shows the confusion matrix for the audio-only speech recognition system when simulated using babble noise (100 people speaking in a canteen) at a SNR of 0dB. It can be seen that digits 'zero', 'one' and 'nine' are highly effected by babble noise compared to the remaining digits.

The performance of the AVSR system was able to improve recognition performance compared with the audio-only system as shown in Table-4. Recognition accuracy of all digits improved, especially digits 'three' and 'five' by 45.8% and 40.3% respectively. Those digits showing only a small improvement in performance were digits 'six' and 'nine', a similar result to that obtained using white noise. This demonstrates that the improvement offered by AVSR depends largely on the quality of the visual information presented to the system.

CONCLUSIONS

A range of experiments has been carried out to demonstrate that the geometrical features established in this work have information content that is highly relevant for the recognition task. The results were compared to those obtained using conventional audio-only system, when operating under a range of different signal to noise ratio conditions. Experimental results show that the new geometrical-based features outperform audio-only system in terms of recognition accuracy by integrating dynamic visual information with delta and delta-delta features.

This work has also investigated the performance of a geometrical-based AVSR system applied to digit recognition under noisy conditions and the results presented using confusion matrices. The experiments showed that the recognition of individual digits exhibits a performance that depends substantially on the magnitude of the movement of the lips required in its articulation. The potential exists to further enhance the current AVSR system by using a decision-fusion approach where recognition is performed separately for each of the modalities, with the partial results from each sub-process being combined to produce the final classification.

REFERENCES

[1] M. Z. Ibrahim and D. J. Mulvaney. 2015. Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping. J. Vis. Commun. Image Represent. Vol. 30, pp. 219–233.

[2] J. Lee and C. H. Park. 2008. Robust Audio-Visual Speech Recognition Based on Late Integration. IEEE Trans. Multimed. Vol. 10, No. 5, pp. 767–779.

[3] C. Chatfield and A. J. Collins. 1991. Introduction to Multivariate Analysis. London, United Kingdom: Chapman and Hall.

[4] M. Z. Ibrahim and D. J. Mulvaney. 2014. A lip geometry approach for feature-fusion based audio-visual speech recognition. IEEE 6th International Symposium on Communications, Control and Signal Processing (ISCCSP). pp. 644–647.

[5] J. Sharp. 2010. Microsoft Visual C# 2010 Step by Step. Redmond, Washington: Microsoft Press.

[6] G. Bradski and A. Kaehler. 2008. Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media.

[7] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2006. The HTK Book (for HTK Version 3.4). Cambridge University. Vol. 2, No. 2. Cambridge University Engineering Department.

[8] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. IEEE Conference on Computer Vision and Pattern Recognition. Vol. 1, pp. 511–518.

[9] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. 2007. A survey of skin-color modeling and detection methods. Pattern Recognition. Vol. 40, No. 3, pp. 1106–1122.

[10] H. Li and M. Greenspan. 2011. Model-based segmentation and recognition of dynamic gestures in continuous video streams. Pattern Recognition. Vol. 44, No. 8, pp. 1614–1628.

[11] L. Rabiner and B.-H. Juang. 1993. Fundamentals of Speech Processing. Prentice Hall Signal Processing Series.

[12] S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Vol. 28, No. 4.

[13] B. S. Atal and S. L. Hanauer. 1971. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. J. Acoust. Soc. Am., Vol. 50, pp. 637–655.

[14] S. Furui. 1986. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. IEEE Trans. Acoust. vol. ASSP-34, No. 1, pp. 52–59.

[15] J. G. Wilpon, C.-H. Lee, and L. R. Rabiner. 1991. Improvements in connected digit recognition using higher order spectral and energy features. Int. Conf. Acoust. Speech, Signal Process.

[16] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas. 2011. The subspace Gaussian mixture model - A structured model for speech recognition. Comput. Speech Lang. Vol. 25, No. 2, pp. 404–439.

[17] L. R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, Vol. 77, No. 2, pp. 257–286.

[18] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. 2002. Moving-Talker, Speaker-Independent Feature Study, and Baseline Results Using the CUAVE Multimodal Speech Corpus. EURASIP J. Adv. Signal Process. No. 11, pp. 1189–1201.

[19] A. Varga and H. J. M. Steeneken. 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. Vol. 12, No. 3, pp. 247–251.