Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

A Comparison of Visual Features for Audio-Visual Automatic Speech Recognition

Nasir Ahmad, Sekharjit Datta, David Mulvaney and Omar Farooq

Loughborough Univ, LE11 3TU Leicestershire, UK
n.ahmad@lboro.ac.uk

Abstract

The use of visual information from speaker's mouth region has shown to improve the performance of Automatic Speech Recognition (ASR) systems. This is particularly useful in presence of noise, which even in moderate form severely degrades the speech recognition performance of systems using only audio information. Various sets of features extracted from speaker's mouth region have been used to improve upon the performance of an ASR system in such challenging conditions and have met many successes. To the best of authors knowledge, the effect of using these techniques on recognition performance on the basis of phonemes have not been investigated yet. This paper presents a comparison of phoneme recognition performance using visual features extracted from mouth region-of-interest using discrete cosine transform (DCT) and discrete wavelet transform (DWT). New DCT and DWT features have also been extracted and compared with the previously used one. These features were used along with audio features based on Mel-Frequency Cepstral Coefficients (MFCC). This work will help in selecting suitable features for different application and identify the limitations of these methods in recognition of individual phonemes.

1 Introduction

Automatic Speech Recognition (ASR) has attracted a lot of research during the past few decades to make the man-machine interaction more natural. Although ASR capabilities are reported to have advanced to near humans level of recognition, this is generally only achievable under ideal conditions and the performance deteriorates significantly in the presence of audio noise [1]. To overcome this drawback a number of approaches have been proposed which are broadly based on the following techniques:

- extraction of features that are robust to noise [2, 3];
- noise compensation techniques [4];
- noise reduction or speech/spectral enhancement front ends [21];
- audio visual feature extraction [5, 6].

To improve the capability of current ASR systems and make it robust to noise, visual speech is a natural candidate. It has long been known that visual information from speaker's mouth region improves speech recognition by humans in presence of noise [7]. However the use of both audio and visual modalities for ASR, known as Audio-Visual Automatic Speech Recognition (AVASR), was first reported in [8]. In this work geometric parameters were extracted from black and white images of the mouth region of the speaker. This was followed by a number of investigations into other informative features from speaker's mouth region for AVASR [9, 10].

AVASR system consist of two channels of information i.e. audio and video channels. Incorporation of visual information is associated with new tasks which can be subdivided into (1) face tracking and mouth region of interest (ROI) extraction (2) visual features extraction and (3) audio-visual integration [11].

Several algorithms have been proposed for face tracking and mouth extraction including colour based segmentation [12], facial features tracking [13] and lip templates [14]. However, these techniques in general perform sub optimally in varying lighting condition. AVASR task rather requires a more accurate estimate of lip parameters for getting an acceptable performance [15]. For this reason, speakers lips are coloured or some marking is placed on it to help accurate tracking and estimation of lip contour [20].

For the integration of audio and visual streams of information; there are three main methods namely, early integration performed on the features level, late integration carried out at the decision level and methods that fall between these two extremes.

The performance of AVASR systems is greatly dependent on extraction of visual features that retain as much information as possible about the original images that is relevant to speech recognition. Geometric based visual features such as mouth opening/closing, mouth height, width and area etc have been used. Most of the techniques used in this kind of features uses coloured lips or other marking [20], however this approach is far from the real world situations. Techniques for automatic lip contour extraction have been proposed by several authors, but to date these have met with limited success. An alternative approach is to adopt appearance based features extraction methods which apply suitable transformation of mouth ROI followed by dimensionality reduction techniques such as linear discriminant analysis (LDA) or principle component analysis (PCA). This paper reports on a comparison of visual features extracted using an appearance approach. New features are extracted from bands of spatial frequencies in discrete cosine transform (DCT) and discrete wavelet transform (DWT) domains and their performance compared with existing approach of using whole range of spatial frequencies for features extraction.

2 System description

This section discusses the database used in the current system and the method developed for mouth ROI extraction

2.1 Database

Unlike audio-only speech recognition where standard databases are abound only a small number of databases are available for audio-visual speech recognition. This is due to a number of factors such as the relative complexity of acquiring and processing audio-visual data, the greater storage requirement for video stream and that AVASR research is mostly carried out by individual researchers or small groups of researchers. The databases that are available often suffer from poor video quality, limited number of speakers, and so are not suitable for continuous speech recognition experiments. To the best of author's knowledge there are currently two databases suitable for large vocabulary AVASR tasks namely audio-visual TIMIT (AVTIMIT) [16] and VidTIMIT [17] databases. The VidTIMIT database contains 43 speakers (24 males and 19 females) and a subset of this database having 32 speakers (16 males and 16 females speakers) was used in the work described in this paper. Each speaker utters eight different sentences (the two sentences common to all speakers in VidTIMIT were not used) in front of a camera centred on the face of the speaker. The sentences in the database are all examples of continuous speech taken from the standard TIMIT database and contain a total of 216 utterances and vocabulary of 925 words. The audio is recorded at sample rate of 32kHz and 16 bits depth; the video is recorded at rate of 25 frames per second.

2.2 ROI extraction

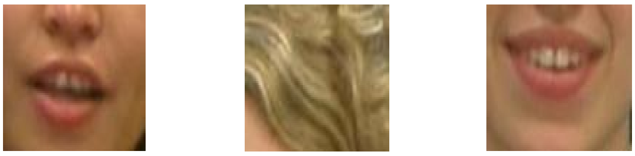
As in our work features are extracted from audio steam 100 times every second, video frames were up-sampled to a rate of 100 frame per second using linear interpolation. Local successive mean quantization transform (SMQT) [18] features were used to find face region in the images. The face region is detected in the first frame of the utterance and the coordinates of the lower half-part are determined. A 100x75 area centred on these coordinates is extracted as the mouth ROI and these same coordinates are used for ROI extraction in the remaining frames of the utterance. This approach was found to work well in general and also reduces the time required for extracting mouth region in every image individually. One such mouth regions thus extracted is shown in Fig. 1 (a). In small number of cases where mouth region was not accurately located by this process, manual correction is done. Fig 1 (b) shows one such missed face while Fig 1 (c) shows the same corrected manually.

3 Visual feature extraction

The extraction of visual features that contain high-quality information suitable for speech recognition purpose is a critical stage in AVASR. The approaches that have been adopted for visual feature extraction, can be grouped into categories of appearance-based, geometric-based and hybrid methods. In this work, appearance-based features have been used.

3.1 Appearance based features

Appearance-based feature extraction approaches consider entire mouth region of the speaker to be informative for speech recognition. Frequency information is often important in signal analysis and suitable transformations of the mouth ROI are often taken to extract such information. The most commonly-used transforms in image compression literature are the DCT and the DWT. The DCT is simply the real component of the Fourier transform with signal analysis being performed at a uniform resolution whereas wavelet analysis perform analysis at a rang of resolution both in time and scale and is therefore called multi-resolution analysis. As is shown in Fig. 2 the DCT and the DWT transforms place spatial-frequency information in increasing order. LL2, HH2 and HH1 contains image at increasing horizontal and vertical details, where index 1 and 2 corresponds to level of decomposition. Here Harr mother wavelet is used at level 1 followed by another single level decomposition of low frequency image LL1 to get LL2, HL2, LH2 and HH2 and similarly for high frequency image HH1.



(a) (b) (c)

Fig 1 Region of interest (ROI) extraction, (a) accurately extracted ROI (b) Missed ROI (c) Manually corrected ROI

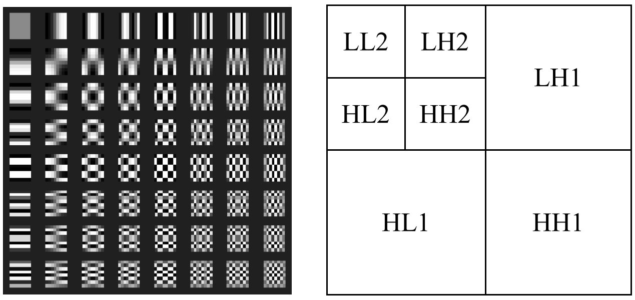


Fig. 2 (a) DCT and (b) DWT based image decomposition

3.2 Linear discriminant analysis (LDA)

A number of techniques are used for classification of data. Two commonly used data analysis techniques are PCA and LDA. PCA transforms data in order of decreasing variance such that maximum of variance lies about the first axis then the second axis and so on. This type of technique is more useful for representing data in a compact set of dimension. However, for data classification, where the prime purpose is to discriminate between different classis, this method is not optimum. LDA on the other hand transforms data as to maximise between class variance and minimizes the within class variance. If within-class scatter matrix is denoted by S_w and between class scatter matrix by S_b then the transformation matrix W is such that,

$$J(W) = \frac{|WS_bW|}{|WS_wW|} \qquad (1)$$

is maximized. The optimum W consists of eigenvectors corresponding to k largest eigen values, where k is the desired dimensionality of the transformed space.

3.3 Feature extraction

In this work appearance based visual features are extracted from the speaker's mouth region and used for speech recognition. Both DCT and DWT based coefficient have been studied.

Most of the features extraction techniques used for audio-visual speech recognition come from data reduction literature where the main goal is to retain information required for image restoration in a compact set of coefficients. Few low frequency DCT and DWT coefficients contain sufficient information for restoring the approximate image. However, this approach does not guarantee if these coefficients also have the most discriminative information required for speech recognition. A number of regions from low frequency edge of DWT transform have been reported in [19]. To the best of author's knowledge a study to evaluate performance of AVASR on different frequency regions of the transformed space has not been carried out yet. In this work the transformed space is subdivided into four regions based on the horizontal and vertical details. These regions are named as R1, R2 corresponds to LL2, HH2 Fig. 2 and similarly R3 and R4 for decomposition of HH1. These regions contains the coefficients of horizontal and vertical frequency components in increasing order. Furthermore, in previous studies [15], high energy coefficients from transformed space have been selected and usually reduced to a lower dimensions using LDA or PCA. Here we use both high energy coefficients (with suffix energy in Fig. 5) and full coefficients set (with suffix full in Fig. 5) for extracting our visual observation vector. The dimensions of the observation vector is reduced using LDA.

4 Experiments

As discussed earlier in section 2.2, the face region is located using SMQT and the mouth region of size 100x75 is extracted around the centre of its lower half. Mouth region thus extracted is resized to size 80x80 (multiple of four to get our four regions). Two Dimensional Discrete Cosine (2D-DCT) and Discrete Wavelet Transform (2D-DWT) are taken and separated into the four frequency region as shown in Fig. 2. In first set of experiments 100 highest energy coefficients are extracted from the four regions both for DCT and DWT coefficients, followed by Linear Discriminant Transform to get our video observation vector of 30 dimensions. In second experiment the 20x20 region is reshaped to a vector of 400 dimensions and followed by LDA to get the final set of 30 dimensions. Class labels for the LDA step are provided by bootstrapping on audio-only HMM developed earlier using forced alignment.

R1			
	R2		
		R3	
			R4

Fig. 3 Regions for features selection

4.1 HMM modeling

Observation vector obtained by our features extraction methods discussed above form a 30 dimensional static feature. Delta and delta-delta features are appended with it to for a dynamic visual feature vector of dimension 90. In our audio-visual experiments 13 MFCC coefficients and their first and second derivatives are extracted and appended to the ninety dimension dynamic visual feature vector thus giving rise to an early integration strategy. A three states hidden Markov Model (HMM) as shown in Fig. 4, is developed for our phoneme set of 46 along with their context dependent tri-phone models using Cambridge University's HTK Toolkit.

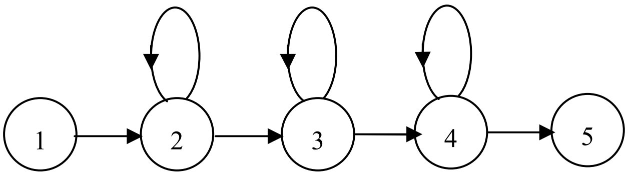


Fig. 4 HMM with tree emitting states

5 Results

Although the experiments have been performed on both audio-visual and video-only tasks but here the results on video-only experiments are reported. A reason for this is, that the language model is only based on a phoneme basis and incorporating language model may affect the results which will not be exactly a measure of performance of its video components. Here we compare the results across a number of factors. We compare DCT based coefficients with DWT based counterparts. As is clear from Fig. 5 the DCT based features outperform their counterpart in DWT based set. Again a comparison of using the energy based features with using the whole feature set is performed. It is obvious that using the whole set of coefficients for features extraction using LDA in general gives better results than using high energy coefficients. A possible argument in support of using energy based coefficients may be its lower dimensions but as the training is done offline and the effects of using whole features set will have very minimal effect on the recognition time. Again comparison of different frequency regions provides that intermediate frequencies are more informative for speech recognition than low energy features.

6 Discussion and Conclusion

This paper investigates the performance of various visual features for automatic speech recognition. Due to use of different databases by audio-visual ASR community and non-existence of standard face and mouth extraction techniques our results can not be compared directly with any other work on this subject. In this work VidTIMIT database is used for a continuous speech recognition task which consist of relatively large number of subjects. We compare our results with the techniques reported in audio-visual ASR literature, on using our own experimental set up. Proposed regions based features are new and are compared with features extracted from whole set of coefficients in transformed space indicated in results as full image. The results are reported here on visual only speech recognition with no use of language model. This provides a direct comparison of visual features with no other factor involved. Our results show that DCT based features in general gives superior performance compared to wavelet transform based features. Again we see that low frequency coefficients although gives best performance for image restoration but for the purpose of speech recognition intermediate frequencies gives better performance. This might be the reason that intermediate level features contains more information about lip moment which is required for speech recognition. Using the whole feature set for training instead of high energy coefficients also add to performance of speech recognition system. Use of whole feature set although increase the dimensionality and increases the training time but its effect on speech recognition is minimal as the training process is performed offline.

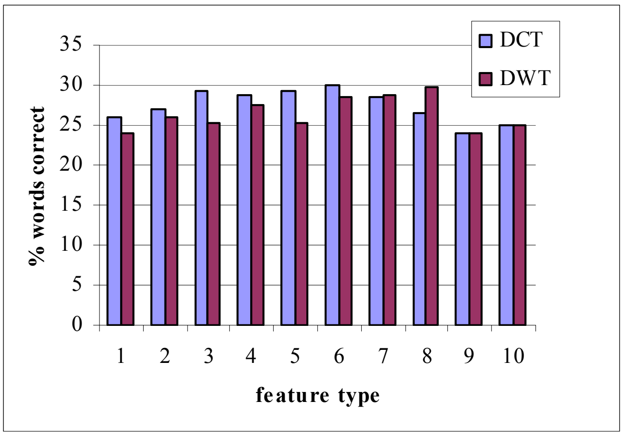


Fig. 5 Results for DCT and DWT based features

1. R1energy 2. R1whole 3. R2energy 4. R2whole 5. R3energy

6. R3whole 7. R4energy 8. R4whole 9. fullenergy 10. fullwhole

References

[1] G. Potamianos, G. Gravier, A Garg, A. W. Senior, C. Neti, “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, *In Proc. Of the IEEE,* Vol. 91, No. 9, pp. 1306-1326 (2003)

[2] H. Hermansky, N. Morgan, “RASTA processing of speech”, *IEEE Transactions on Speech and Audio Processing,* Vol. 2, No. 4, pp. 578-589 (1994)

[3] K. H. You, H. C. Wang, “Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences”, *Speech Communication*, Vol. 28, pp. 13-24 (1999)

[4] L. Neumeyer, M. Weintraub, “Probabilistic optimum filtering for robust speech recognition”, *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASP)*, Vol. 1, pp. I-417-I-420 (1994)

[5] P. Duchnowski, U Meier, A. Waibel, “See Me, Hear Me: Integrating Automatic Speech Recognition and Lip-Reading”, *In Proc. of ICSLP* pp. 547-550 (1994)

[6] X Li, C. Kwan, “Geometric Feature Extraction for Robust Speech Recognition”, *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, pp. 558-562 (2005)

[7] W. H. Sumby, I. Pollack, “Visual Contribution to Speech Intelligibility in Noise”, *Journal of the Acoustical Society of America*, Vol. 26, No. 2, pp. 212-215 (1954)

[8] E. D. Petajan, “Automatic Lipreading to Enhance Speech Recognition”, *Proceedings of the IEEE Communication Society Global Telecommunications Conference* (1984)

[9] M. S. Gray, J. R. Movellan, T. J. Sejnowski, “Dynamic features for visual speech-reading: A systematic comparison”, *In Advances in Neural Information Processing System 9*, M. C. Mozer, M. I. Jordan, and T. Petsche, Ed., MIT Press, Cambridge, pp. 751-757 (1997)

[10] K. Saenko, T. Darrell, J. R. Glass, “Articulatory features for robust visual speech recognition”, *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 152-158 (2004)

[11] G. Potamianos, C. Neti, J. Luettin, I. Matthews “Audio-Visual Automatic Speech Recognition: an Overview”, *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds., MIT Press, (2004)

[12] T. W. Lewis, D. M. W. Power, “Lip Feature Extraction Using Red Exclusion and Neural Networks”, *25th Australasian Computer Science International Conference,* pp. 61-67 (2002)

[13] R. Steifelhagen, J. Yang, U. Meier, “Real Time Lip Tracking for Lipreading”, *Proc. Eurospeech, ’97* (1997)

[14] D. Chandramohan, P. L. Silsbee, “A Multiple Deformable Template Approach for Visual Speech Recognition”, *Proc of Fourth International Conference on Spoken Language,* Vol. 1, pp. 50–53 (1996)

[15] G. Iyengar, A. Verma., T. Faruquie, G. Potamianos, C. Neti, “Robust detection of visual ROI for automatic speechreading” , *Proc of IEEE fourth Workshop on multimedia signal processing,* pp. 79-84 (2001)

[16] T. J. Hazen, K. Saenko, C. H. La, J. Glass, “A segment based audio-visual speech recognizer, data collection development and initial experiments”, *In Pro. ICMI,* pp. 235-242 (2004)

[17] C. Sanderson, K. K. Paliwal, "Polynomial Features for Robust Face Authentication", *Proc. IEEE Inter. Conf. on Image Processing*, Vol. 3, pp. 997-1000 (2002)

[18] M. Nilsson, J. Nordberg I. Claesson, “Face Detection Using Local SMQT Features and Split-up Snow Classifier”, *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Vol. 2, pp. II-589-II-592 (2007)

[19] G. Potamianos, H. P. Garf and E. Cosallo, “An Image Transform Approach for HMM Based Automatic Lipreading” , Proc. Of International Conference on Image Processing, Vol. 3, pp. 173-177 (1998)

[20] M. Hckmann, F. Berthommier, K. Kroschel, “A hybrid ANN/HMM audio-visual speech recognition system*”, Proc. International Conference on Auditory-Visual Speech Processing*, pp. 190-195 (2001)

[21] W. Yifang, Z. Jingjie, Y. Kaisheng, C. Zaigang, M. Zhengxin, “Robust recognition of noisy speech using speech enhancement”, *Pro of 5th International conference on signal processing .WCCC-ICSP,* Vol. 2, pp. 734-737 (2000)