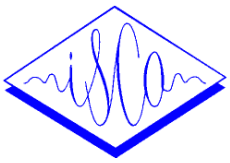

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Visual Feature Analysis for Automatic Speechreading

Patricia Scanlon, Richard Reilly, Philip de Chazal

Dept. of Electronic and Electrical Engineering
University College Dublin, Belfield, Dublin 4, Ireland
{patricias,richard.reilly,philip}@ee.ucd.ie

Abstract

This paper proposes a novel method of visual feature extraction for automatic speechreading. While current methods of extracting delta or difference features involves computing the difference between adjacent frames, this method proposed provides information on how the visual features evolve over a time period longer than the time period between adjacent frames, the time period being relative to the length of the utterance. These new features provide a visual memory capability for improved system performance. Good visual discrimination is achieved by maintaining a base level of detail in the features. A frame rate of 30 frames per second provides rapid visual recognition of speech. The combination of the novel visual memory features, good visual discrimination and rapid visual recognition of speech movements is shown to improve visual speech recognition. Using this method an isolated word accuracy of 28.1% for a vocabulary 78 words over a database of 10 speakers was achieved.

1. Introduction

Speechreading may be defined as "the ability to understand a speaker's thoughts by watching the movements of the face and body and using information provided by the situation and language" [3]. It is well known that visual information from the face of a speaker provides related speech information that enhances intelligibility of speech utterances under difficult listening conditions. Even under noiseless conditions, speechreading is known to improve the accuracy of speech perception in people with normal hearing [1]. Audio-Visual Automatic Speech Recognition (AVASR) systems use visual information to enhance Automatic Speech Recognition (ASR) systems [4 -10].

Auditory and visual speech provides two independent sources of information but they can also be considered complementary. Certain characteristics which may be strong in one modality may be weak in the other i.e. some phonemes which are difficult to understand acoustically are easily distinguished visually and vice versa. Visemes are defined as visually distinguishable speech units. Most visemes cannot be uniquely associated with a single phoneme, therefore there is often a many-to-one mapping of phonemes to visemes. It is this complementary nature of speech that drives the use of visual information in ASR.

As little is known on exactly how humans perceive speech, it is beneficial to look at human speech perception in order to obtain clues on methods to optimise speech recognition by machines. It is clear from the McGurk effect [2], that humans integrate the audio and visual modalities, but at what stage is not so apparent. Consequently, significant research has been carried out focusing on the integration of the audio and visual modalities [6], [7], another active area of research is in determining which features should be extracted for classification of speech [8-10].

The skills required for speechreading are visual recognition, visual discrimination, visual memory and flexibility. Visual recognition is the ability to recognize different speech movements and to be able to do this rapidly. Visual discrimination is more concerned with the ability to see fine differences in lips, teeth, and jaw positions in order to distinguish between different speech movements. Another important skill is visual memory, which is the ability to recall previous visual patterns of speech. As the speechreader may need to change their interpretation of the utterance as the conversation unfolds, a capacity to remember the previous visual patterns of speech is vital. Flexibility is a further skill that is also required allowing the speechreader to make quick changes in the perception of the utterance as it is being spoken.

In the next section visual feature extraction methods are discussed and in section 3 the implementation of the proposed method is described. Section 4 describes the results where the level of detail required for visual speech recognition was investigated. Also more traditional methods of computing delta features are compared with the proposed method of computing delta features over longer time periods.

2. Visual feature extraction methods

The objective of video processing for AVASR is to pre-process facial images into a set of meaningful parameters for subsequent classification into words. A region of interest (ROI), consisting of the mouth area and perhaps the jaw, is then located in the acquired facial images. Image pre-processing is employed within the ROI to minimize the effects of variable lighting conditions, using techniques such as histogram flattening and balancing of the left-to-right brightness distribution. The ROI images are also often downsampled at this stage, to reduce their dimensionality and therefore the complexity of the feature extraction process. The downsampling step also transforms the image into a square matrix of pixel values of dimension *NxN* to allow ease of manipulation. Downsampling the ROI image reduces the resolution (number of pixels) and as a result can reduce the level of detail in the image.

Using pixel based methods every pixel in the entire raw image is considered a feature. This approach ensures that no information is lost but the size of the feature vector can be very large and may also contain considerable redundancy. Image transformations transform *NxN* pixels into *NxN* transform coefficients. The transform coefficients are ordered according to the importance of their information [11]. The transformation process compacts most of the image's energy in a relatively small number of coefficients thus removing the

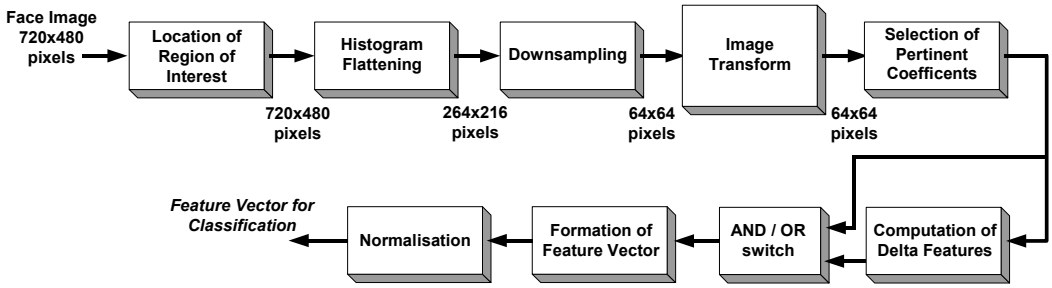

Figure 1 Visual Subsystem

linguistically redundant information. The low frequency coefficients represent the low level detail of the image, with the finer image details in the higher frequency coefficients. A subset of the transform coefficients is used as the features.

The Discrete Cosine Transform (DCT) is the most widely used transformation in transform coding and performs well for highly correlated data and has excellent energy compaction. The Hadamard and Haar transforms are also implemented for image processing application [11]. Unlike the DCT, these transforms are non-trigonometric based transforms. The Hadamard transform basis functions consist of +1's and -1's, therefore requiring no multiplications in the transformation. The Hadamard transform provides good energy compaction but not as efficiently as the DCT [11]. The Haar transform coefficients are the differences along the rows and columns of the local averages of pixels in the image, which results in good edge extraction. The Haar transform provides relatively poor energy compaction [11]. These two transforms have enormous benefit for embedded systems, as the computational requirements are far less than that of the DCT.

The number of transform coefficients used reflects the level of detail in the feature vector. While smaller feature vector dimensions result in better use of the available training data, larger feature vectors result in more discriminating detail being included. What is needed is a realistic feature vector dimension that maintains a good level of discriminatory detail for the application at hand.

In speech, the acoustic and the visual information necessary for the recognition of a word is not present simultaneously. Therefore it is the dynamic nature of speech that provides the information on what was said and therefore information further back in the utterance needs to be taken into account. An investigation into geometric features required for visual speech recognition described in [12] concluded that the most discriminative features were primarily dynamic. Another method that was used to include dynamic information involved concatenating features from a number of neighboring frames and then performing linear discriminate analysis to reduce the number of features used [13]. Another way to include dynamic information is to include the temporal difference between frames. Current methods of extracting delta features involve computing the difference between adjacent frames [14].

However, it can reasoned that since visual speech is generally sampled at around 30 frames per second (fps), twice the minimum requirement for efficient human speech reception [15], the temporal difference between the two adjacent frames will not be great. Therefore this temporal difference between adjacent frames does not provide significant discriminating features for classification. However, taking the difference between the current frame and a frame much further back in the image sequence should provide information on how the visual features evolve over periods of time longer than 1/30sec.

Rapidly analysing the dynamics of the speech articulators, faster than the human visual system can detect, should enable a computer-based system to accurately identify all utterances visually. However the confusability of some visemes reduces this accuracy. Also, unless all image pixels are used, some level of detail is sacrificed, which could provide assistance in distinguishing speech articulator dynamics. By using a combination of good level of image detail, fast frame rate and accounting for previous visual patterns in the utterance a good level of accuracy can be achieved.

3. Implementation

The original acquired face image is of size 720x480 pixels. Location of the ROI is the first process and in this study the ROI is centred about the mouth area using a window of size 264x216 pixels [17]. The images are then pre-processed with histogram flattening to minimise the effects of variable lighting conditions. The images are subsequently downsampled to a resolution of 64x64 pixels.

The downsampled ROI images are image transformed and the transform coefficients used as the feature vector for classification. The transforms investigated included the Discrete Cosine, Haar and Hadamard, with subsets of the transform coefficients used to examine the level of detail required for good visual speech recognition. The visual feature vector was formed, by taking 15, 28 or 36 of the highest energy components produced on applying the image transform.

In addition to the transform coefficient features, temporal difference information or *delta* features between frames are also calculated. Given frame *k,* delta features are computed between current frame *n* and *n-k*. The value *k* is relative to the total number of frames N within an utterance e.g. *k = 1, N/3, N/2, N*2/3, N*5/6, N-1*. To account for the different speaking rates of each instance of an utterance, *k* was set relative to the total number of frames, *N*. Therefore, the delta features are more reliably computed between the same segments in each instance of the utterance. With these delta features included in the feature vectors, the feature vector dimension increases in length to 30, 56 and 72 respectively. For frames numbered 1 to *k* the delta features appended to the feature vector are zero.

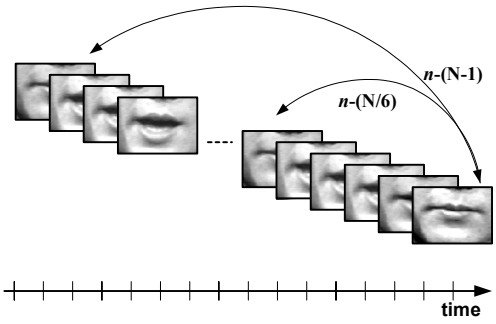

Figure 2 Visual memory features (e.g.N = 30 frames shown).

All features in the feature vector *X* were normalized prior to classification. The normalization procedure for each feature *x,* required calculation of the mean and standard deviation across the training samples and these were used to rescale the feature values. After rescaling each feature across the training sample had a mean of zero and a standard deviation of one. The same normalisation values were used to normalise the test data.

The audio pre-processing was based on extraction of mel-frequency Cepstral coefficients of order 12, in conjunction with log energy of the speech frame. The first and second derivatives are also included in the feature vector, which has a dimension of 39 [16].

Hidden Markov Models (HMMs) were used for classification of both audio and visual isolated word recognition. The Hidden Markov Model Toolkit, HTK 3.0 was used to implement the HMM topologies [16]. Both the audio and visual HMMs contained 10 states with one mixture per state.

The audio and visual modalities were integrated following separate classification. The fusion scheme chosen is multiplicative using probabilistic rules. The scheme initially selects the candidate that maximises the cross product of the N-best output probabilities of the audio and visual modalities; N was set to 10 in this study. This process is known as late integration [7].

The audio and visual N-best output probabilities are subsequently weighted according to the dispersion or variances of their N-best output probabilities. These adaptive weights account for the confusability of phonemes visually and also the confusability of phonemes acoustically for varying levels of SNR. This weighting indicates the reliability of the modalities [6].

$$\lambda = \frac{\sigma_v}{\sigma_v + \sigma_a} \qquad (1)$$

The weighting is carried out using Equation 1, where σ_a and σ_v are the variances of the audio and visual modality's N-best output probabilities, respectively. The visual N-best output probabilities are weighted using λ and the audio N-best output probabilities using $1-\lambda$. However when one of the modalities becomes corrupt, it can mask reliable output recognition from the other modality. Placing thresholds on the integration process can avert this corruption. When λ falls below a threshold, *t*, the candidate that maximizes the visual output probabilities is selected, also when λ is greater than *1-t,* the candidate that maximizes the audio output probabilities is selected. Through empirical testing $t = 0.3$ was found to maximizes audio-visual recognition.

4. Results

The audio-visual database employed in this study consisted of 10 speakers (7 male and 3 female) [20]. For each speaker ten full frontal recordings were taken for a vocabulary of 78 words. The audio signals were contaminated with white Gaussian noise giving audio recognition across varying SNR's e.g. -12, -6, 0, 6, 12, 18 and 24dB.

N-fold cross validation was used in all results to maximise the use of the data available. The data was divided into *n* subsets of equal size, where *n* is the number of speakers and each subset contains all the recordings of one speaker. The system is trained and tested *n* times, each time leaving out one subset from training and using the omitted subset for testing. The results are obtained using 4 recordings of the 10 speakers uttering the entire 78 word vocabulary.

The usefulness of the delta features is first compared to the static features only and to the static plus delta features in Table 1, where 28 static and 28 delta features were used alone or in combination.

Features Used	Recognition accuracy (%)
Static only (28 features)	7.2
Delta only (28 features)	28.1
Static + Delta (56 features)	20.2

Table 1 Comparing recognition results using, Static, Delta and Static+Delta features

The effects of varying *k* as part of the delta(*n-k*) feature set, for visual speech recognition was first considered. Static transform coefficients of dimension 15, 28 or 36 were first extracted. Then delta features were computed as the difference between the current frame and the kth previous frame, where frame and the feature vector dimension was increased to 30, 56 or 72.

Figure 3 shows the variation of visual recognition for different feature vector dimensions. Visual recognition was found to be maximized for delta(*n-k*) where *k* is *N/6* over all feature vector dimensions, i.e. the delta or difference features are computed between the current frame *n* and the frame (n-*N/6).* Feature vector dimension of 30 was found to provide a higher accuracy across all *k* than that of 56 and 72. The best visual recognition rate was 20.2%, which occurred at *k = N/6* and with a feature vector dimension of 56.

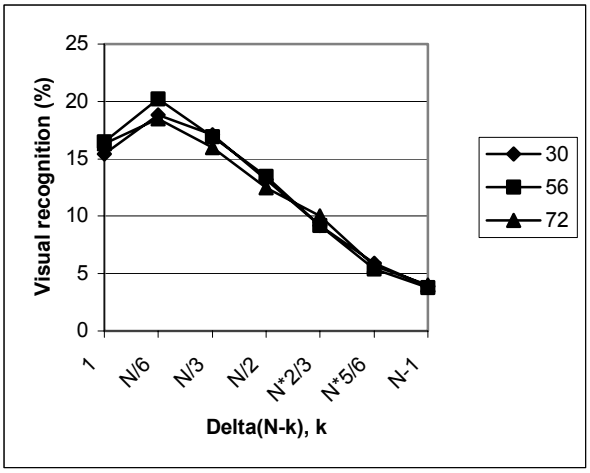

Figure 3 Variation of visual recognition as a function of k for delta(n-k), for different feature vector dimensions. Feature vector includes DCT coefficients and delta features

In Figure 4 we examine the use of feature vectors composed of no transform coefficient features but only delta (*n-k*) features. The exception is for the case of *No Delta*, where only 15, 28 and 36 transform coefficients were used as features. Results are again found to be maximized for $k = N/6$, for all feature vector dimensions. The maximum visual recognition rate obtained was 28.1%, which occurred at $k = N/6$ and with a feature vector dimension of 28. Table 2 indicates the most frequently confused words for $k = 1, N/2, N-1$.

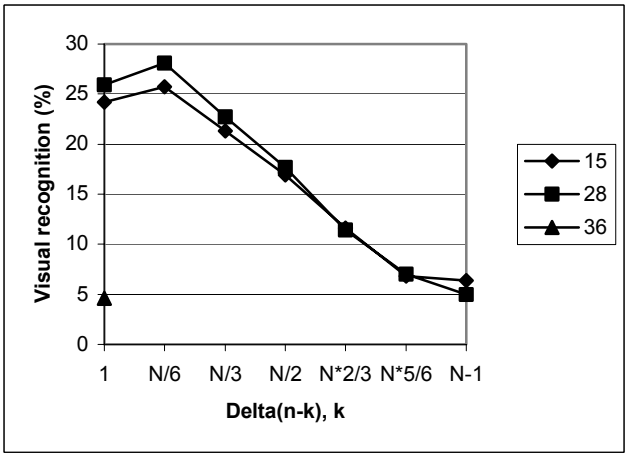

Figure 4 Variation of visual recognition as a function of k for delta(n-k), for different feature vector dimensions. The feature vector includes only delta features

In Figure 5 the different image transforms implemented, Discrete Cosine, Haar and Hadamard Transforms are compared. Again, the results are obtained using 4 recordings of the 10 speakers uttering the entire 78 word vocabulary. The feature vector consisted of 28 delta features. Again, as in Figure 4 and Figure 5, visual recognition was seen to be maximized at $k = N/6$, for all three image transforms.

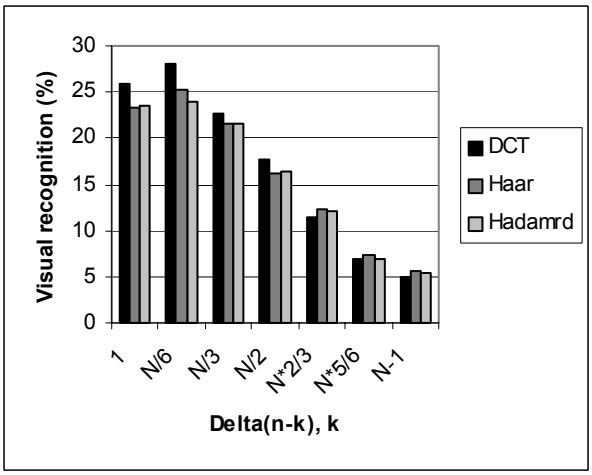

Figure 5 Variation of visual recognition as a function of k for delta(n-k), for different transformations. Feature vector dimension is 28 and includes only delta features

The results of AVASR across all SNR values are shown in Figure 6. The visual recognition was performed on feature vectors composed of 28 delta features.

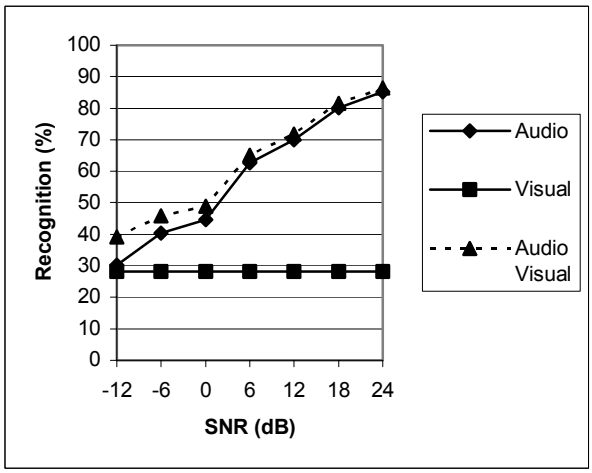

Figure 6 Variation of audio, visual and audio-visual recognition as a function of audio SNR

5. Discussion

The results in Table 1 show that the visual delta features provide a much more consistent description of the isolated utterance than the static features and the combined static and delta feature vector. Also choosing a value for *k*, as in *delta(n-k),* relative to the number of frames in the utterance, *N*, consistently provides difference features between the same segments of the utterance, for all instances of the utterance.

A number of factors contribute to the rest of the visual speech recognition results obtained in Section 4. Firstly, the effects on recognition of the level of detail included in the feature vector were examined. Figure 3 illustrates the need for a base level of detail to be contained in the feature vector. A realistic feature vector dimension that maintains a good level of discriminatory detail is required. Increasing the number of transform coefficients increases the visual recognition accuracy. However, the size of training data available limits the possible feature vector dimensions for

good recognition. It is observed in Figure 3 that a feature vector dimension of 56 outperforms that of 30 and 72. There are insignificant differences in the recognition accuracies of feature vector lengths of 30 and 72. Therefore 56 is a realistic feature vector dimension that maintains good accuracy for this database size.

The addition of the new delta features, shows an increase in visual speech recognition for all feature vector dimensions. The number of frames per utterance in the database varies considerably. Therefore the difference or delta features delta(n-k), are being taken between different positions within the utterances. What was required was a method of choosing the delta frames, which is relative to the number of frames per utterance. Figures 3 and 4 shows the results for choosing the value of k relative to the number of frames N, e.g. $k = 1, N/3, N/2, N*2/3, N*5/6, N-1$. Visual recognition is observed to be maximum at $k = N/6$ i.e. for an utterance 30 frames of length, $k = 5$.

As speech is analysed rapidly at 30 fps, the difference in features between two consecutive frames does not provide sufficient discriminatory features. Thirty fps corresponds to a difference in the features occurring 0.03 seconds apart. However, from Figure 3, visual recognition is maximised where k is N/6. At 30 fps this corresponds to the generating a feature vector based on two frames 0.2 seconds apart, for an utterance 30 frames long.

The different transforms implemented in Figure 5, show highly comparable results. The DCT while being the most widely implemented image transform, performs slightly better than the Haar and Hadamard transforms. This indicates that it is not necessarily the type of transform chosen that provides useful visual information, rather it is how the features evolve over time, through analysis with delta features that provide good discriminatory information. It is useful to note that while the Haar and Hadamard transforms are much simpler to implement than the DCT, they are providing highly comparable visual recognition results to those of the DCT.

Figure 4 examines the effects of using delta features only. The transform coefficient features are not included in the feature vector and this reduces the feature vector dimension in half. It can be seen that it is the delta features that provide robust visual features for recognition, as the visual recognition results in Figure 4 are significantly higher than in Figure 3 e.g. maximum recognition for feature vector including both transform coefficient features and delta features is 20.2%, while from Figure 4 maximum recognition for feature vectors composed of delta features only is 28.1%.

The audio and visual modalities were integrated using thresholded adaptive weighting. Introducing a threshold at t =0.3, the audio-visual recognition is greater than both the audio and visual recognition alone, across all SNR's. In highly noisy conditions, –12dB, the audio-visual recognition is 9% higher than audio-only recognition. Even in noiseless conditions, 24dB, the audio-visual integrated results shows an improvement of 1.5% over audio-only recognition. In fact, the audio-visual recognition accuracy is greater than audio-only or visual-only recognition across all SNR.

The audio-visual database employed in this study while containing a wide vocabulary does not contain many visually similar words. Further investigation is required into what values of k for delta(n-k) should be used for more robust visual speech recognition to distinguish between utterances with visually similar phonemes. A feature vector including a combination of delta feature sets of different k could provide more robust recognition for visually similar words.

6. Conclusion

Skills required for speechreading in humans have been applied to automatic speechreading. The combination of visual memory, good visual discrimination and rapid visual recognition of speech movements was shown to improve visual speech recognition. The novel delta features, that provide the visual memory, were shown to improve visual recognition over static features only and also over static plus delta features computed over consecutive frames. The maximum recognition for a vocabulary of 78 words, for feature vectors including both transform coefficient features and delta features, is 20.2%. While the maximum recognition for feature vectors composed of delta features only was found to be 28.1%. Also, the different transforms implemented, show highly comparable results. This indicates that it is not necessarily the type of transform chosen that provides useful visual information, rather the new delta features that improve recognition.

The delta features are computed based on the number of frames in the utterance, while this works well for isolated word recognition it does not translate well to the task of continuous speech recognition. However, these results strongly indicate the importance of delta features for visual speech recognition, further investigation is required comparing results using different delta/derivative kernels. Also, future work will involve testing these features on a larger database and on continuous speech recognition experiments.

7. Acknowledgements

The authors are grateful to Professor Tsuhan Chen, Carnegie Mellon University, Pittsburgh, PA, for providing the audio-visual database used in this study. The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

8. References

[1] Summerfield, Q. "Lipreading and audio visual perception", Phil. Trans. R. Soc. London 1972.

[2] McGurk, H., MacDonald, J "Hearing lips and seeing voices", Nature, Vol. 264, 1976.

[3] Kaplan, H., Bally, S.J., Garretson, C., "Speechreading: A way to improve understanding", Revised 2nd Edition, Gallaudet University Press, 1985.

[4] Petajan, E.D. "Automatic lipreading to enhance speech recognition", Proc. of the IEEE Comm Society Global Telecommunications Conference, Atlanta, Georgia, 1984.

[5] Hennecke, M., Stork, D., Prasad, K. "Visionary Speech: Looking Ahead to Practical Speechreading Systems", Speechreading by Humans and Machine, Springer 1996.

[6] Adjoudani, A., Benoit, C. "On the Integration of Auditory and Visual Parameters in an HMM-based ASR", Speechreading by Humans and machine, Springer 1996.

[7] Teissier, P., Robert_Ribes, J., Schwartz, L., Guérin_Dugué, A. "Comparing models for audio-visual fusion in a noisy-vowel recognition task", IEEE Trans. on Speech and Audio Processing, vol. 7, no. 6, 1999.

[8] Potamianos, G., Cosatto, E., Graf, H.P., Roe, D.B. "Speaker independent audio-visual database for bimodal ASR", Proc. European. Tut. Work. Audio-Visual Speech Proc., Rhodes, pp. 65-68, 1997.

[9] Potamianos, G., Graf, H.P Cosatto, E. "An Image Transform Approach for HMM based Automatic Lipreading", Proc. Int. Conf on Image Processing, Rhodes, Greece, p.p. 173-177, 1998.

[10] Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., Harvey, R. "Extraction of Visual Features for Lipreading." IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2):198-213, February 2002.

[11] Jain, A. "Fundamentals of Digital Image Processing", Prentice Hall, 1989.

[12] Goldschen, A., Garcia, O., Petajan, E. "Continuous optical automatic speech recognition by lipreading" 28th Asimolar conf. on Signals, Systems and Computers, 1994.

[13] Potamianos, G. and Neti, C. "Improved ROI and within frame discriminant features for lipreading" Proc. Int. Conf. Image Processing, Thessaloniki, Greece, 2001.

[14] Gray, M. S., Movellan, J. R., and Sejnowski, T. J., "Dynamic features for visual speechreading: A systematic comparison" Advances in Neural Information Processing Systems Volume 9, 751-757, 1997.

[15] Frowein, H.W., et al., "Improved speech recognition through video telephony: experiments with the hard of hearing." IEEE Journal on Selected Areas in Communication, vol. 9, no. 4, 1991.

[16] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P. "The HTK Book (for HTK Version 3.0) ", Microsoft, 2000.

[17] Audio-Visual data corpus, Advanced Multimedia Processing lab, Carnegie Mellon University, Pittsburgh, PA, USA.