# ‘Automatic speech reading by oral motion tracking for user authentication system’

Vibhanshu Gupta,
M.E. (EXTC, IV SEM)
vibhanshu.gupta@rediffmail.com,
V.E.S.Institute of Technology
Mumbai University, India,

Sharmila Sengupta
M.E. (EXTC.)
sharmilase@yahoo.com
V.E.S.Institute of Technology
Mumbai University, India

**Abstract-** Automatic speech recognition (ASR) systems are used in recognizing speech with high accuracy rates. Visual information is important for human machine interface. It not only increases the accuracy of an Automatic Speech Recognition (ASR) but also improves its robustness. This paper presents an overview of different approaches used for speech recognition and concentrates on visual only lip reading system. Lip reading can be utilized in many applications such as hearing impaired aid and for noisy environment where speech is highly unrecognizable and as password entry system. The visual feature extraction methods are pixel based such as discrete cosine transform (DCT), discrete wavelet transform (DWT)etc. Other feature extraction methods utilize motion analysis of image sequences representing lip movement. This paper is a survey paper explaining comparisons, pros and cons, analysis of various techniques and methods for speech recognition by lip motion tracking.

**Keywords**-Authentication, Automatic, motion, oral, reading, speech, tracking

## INTRODUCTION

Automatic Speech Reading is a method or a system of recognizing the spoken word and comprehends its meaning. This paper is about relative merits, demerits of various approaches to particular speech recognition problems. The speech recognition is a natural alternative interface to computers for people with limited mobility in their arms and hands, or those with sight limitations. For some aspects of computer applications, speech may be a more natural interface than a keyboard. The performance of speech signal degrades drastically in a noisy environment. The ASR is quite useful where speech is highly unrecognizable. It is a hands-free operation, in many situations hands are not available to issue commands to a device. Using a car phone and controlling the microscope position in an operating room are some applications for which limited vocabulary systems exist. The password entry scheme using a visual lip reading is another application of speech security [2].There are many reasons why speech recognition is often quite difficult. The natural speech is continuous, it often does not have pauses between the words, and this makes it difficult to determine wherethe word boundaries are. Among other things, speakers change their mind in mid sentence about what they want to say and utter filled pauses (uh, um etc.) while thinking next message. Large vocabularies are often confusable. A 20,000 word vocabulary is more likely to have more words that sound like each other than a 10 word vocabulary. There is also the issue of out of vocabulary words that have not been seen before. How to model these unknown words is an important unsolved problem. Although high recognition accuracy can be obtained for clean speech using the state-of-the-art technology even if the vocabulary size is large, the accuracy largely decreases in noisy environments. Increasing the robustness to noisy environments is one of the most important issues of ASR. The recorded speech is variable over room acoustics, channel characteristics, Microphone Characteristics and background noise. Background noise and acoustics in the environment that has a speaker is in, will also have tangible effects on the signal. All of these factors can change the characteristics of the speech signal, a difference that humans can often compensate for, but the current recognition systems often cannot.

The recognition performance of VCV syllables is analyzed; reflective markers are placed on the speaker’s mouth, and used these to extract 14 distances, sampled at 30 frames a second. Both mean squared error distance, and an optimized weighted mean squared error distance are considered. The pixel values of the mouth image are fed to a multi-layer network with no feature extraction for the mouth height or the mouth width performed.

Other method is to combine the visual features, either geometric parameters such as the mouth height and width or non-geometric parameters such as the wavelet transform of the mouth images to form a joint feature vector [8]. It was also tried to convert mouth movements into spoken speech directly.

A system called "image-input microphone"[6] takes the mouth image as input, analyze the lip features such as mouth width and height, and derive the corresponding vocal-tract transfer function. The transfer function is then used to synthesize the speech waveform. The advantage of the image-input microphone is that it is not affected by acoustic noise, and therefore is more appropriate for a noisy environment.

### 3. DIFFERENT METHODS OF FEATURE EXTRACTION

Feature extraction is a crucial part of speech reading system by tracking the lip movements. Visual information aids in distinguishing acoustically similar sounds, such as nasal sounds: /n/, /m/, and /ng/ [7, 10].Various visual features have been proposed. In general, feature extraction methods are pixel based where features are employed directly from the image or lip contour based, in which a detection model is used to extract the mouth area or some combinations of the two methods DCT & PCA, DWT & LDA .Motion and color based Tracking proceeds in two ways after user selects the object by drawing a rectangle around it. Within a GOP, motion based tracking, using forward motion vectors of both P and B frames, is performed .The advantage is that the object is tracked in both P and B frame of the GOP. GOP boundaries are quite comfortably crossed by using backward motion vectors of the last B frame of each GOP. This is a more reliable method than use of reversed P frame motion vectors ([9] and faster than block matching ([4] and [5]), besides being done completely in the compressed domain. However, due to the limitations of motion vectors, in the absence of any verification for the object being tracked, errors can get introduced in motion vector based tracking. These prove to be cumulative in nature. So at each I frame, color based tracking is performed, which involves identifying the best image area that matches the original object marked out by the user. For this purpose chrominance DCT values of Cr and Cb in the I frames are used, unlike in [3] and [9] where Y DCT coefficients are used.

#### 3.1 Pixel based image transforms technique, DCT (discrete cosine transforms)

This method of visual feature extraction is based on a Discrete Cosine Transform (DCT). Different strategies are evaluated to choose those DCT coefficients which are best suited for the recognition process and their efficiency is analyzed in a video only and in different audio-visual recognition scenarios. Furthermore a comparison can be drawn between these DCT features and geometric lip features with different noise types at different SNR levels. The comparison is based on two similar data sets, for the first dataset we use the DCT to extract the relevant visual features and in the second the lips of the speaker are colored which allows to precisely extracting the geometric lip parameters. The recognition tests aiming to assess the performance of the DCT features for audio-visual speech recognition are carried out with an ANN/HMM hybrid model. The extraction of the video features is performed with the Discrete Cosine Transform (DCT) [7]. The reasons for the widespread use of the DCT as well in image compression [8] as feature extraction [1] are the high compaction of the energy of the input signal onto a few DCT coefficients and the availability of a fast implementation of the transform, similar to the Fast Fourier Transform (FFT) [7]. Since the DCT is not shift invariant, the performance depends on a precise tracking of the ROI.As features for our recognition experiments, DCT coefficients are selected following three different strategies: Energy features :( features with the highest energy Variance features :( features with highest variance Relative variance features: features with the highest variance after normalization to their mean value.

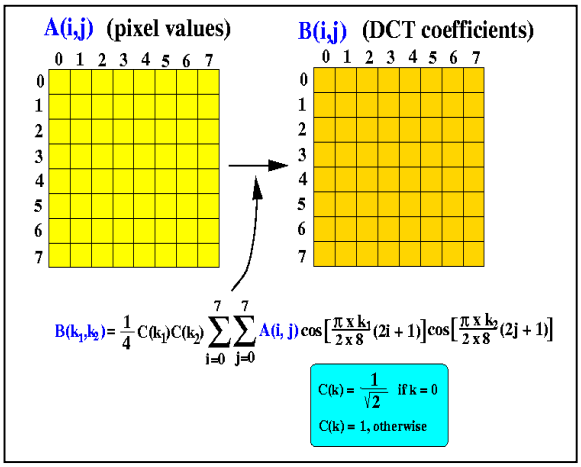


**Fig.1 Pixel based Image transform (DCT Technique)**

The number of features are extracted from each image frame are varied between and the necessary mean values and variances were calculated over the complete training set. Synchronization between the audio and video stream was obtained via an interpolation of the DCT coefficients to the audio feature rate.

### 3.2 Pixel based Image transforms technique DWT (discrete wavelet transforms)

The use of the DWT and our implementation of the PCA, based on the correlation matrix of three-dimensional data, are new. In all cases, the extracted visual features are appropriately post processed and used in a hidden Markov model (HMM) [14] based automatic lip reading system. The effects of different types of video degradation on lip-reading accuracy are investigated in conjunction with the use of image transform based features. All three, field rate decimation, noise, and compression artifacts, arise in practical video capturing and transmission. Where the effects of rate decimation are studied on a simple human lip reading task, and [16], where possible effects of video compression on person authentication accuracy are discussed. In its current implementation, for each video field, two channels of processing are used, a combination of shape

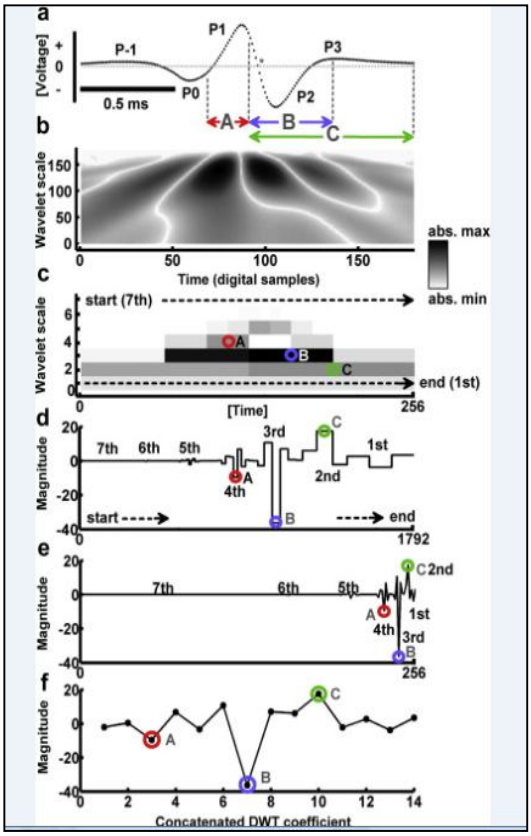


**Fig.2 Pixel based Image Transform (DWT)**

and texture analysis, and a color segmentation, to first locate the mouth and then the precise lip shape.

### 4. LIP CONTOUR BASED FEATURES

For a single speaker, part of the outer lip contour is missed in less than 0.25% of the processed images. However, inner lip and multi-speaker contour estimation are less robust. The video frame is read from the active input buffer and the left and right corners of the mouth are located using the
Contrast information for the lower half of the face. The top and bottom of the mouth are found by tracing along the lip border. The design can be broken into two main blocks: one to search for the left and right corners of the mouth and the other one to search for the top and bottom of the mouth. Both blocks use
a common block to compute the contrast of the image, which uses several adders. The left–right search block accepts the starting point and search direction as its input. The top–bottom search block
takes the search direction (up or down) and the starting and ending points, which were found previously using the left–right search block. The result from the lip tracking block can be either displayed on a monitor or sent to a host.

### 5. BLOCK BASED MOTION ESTIMATION ANALYSIS

The other method is motion estimation analysis for robust speech recognition from lip reading alone. Visual features are extracted from the image sequences and are used for model training and recognition. Block based motion estimation techniques are used to extract visual features blindly without any prior knowledge of lip location Motion estimation removes temporal redundancies among video frames and is a computation intensive operation in the video encoding process. Block based schemes assume that each block of the current frame is obtained from the translation of some corresponding region in reference frame. The motion estimation tries to identify this best matching region
in the reference frame for every block in the current frame. In fig. 1, the gray block on the right corresponds to the current block and the gray area on the left represents the best match found for the current block, in the reference frame. The displacement is called the motion vector. The search range specified by the baseline H.263 standard allows motion vectors to range between –15 pixels and 16 pixels in either dimension. The size of the Search window is of size 32 X32 about the search centre.
Block matching algorithm (BMA) for motion estimation (ME) has been widely adopted by the current video compression standards, such as H.261,

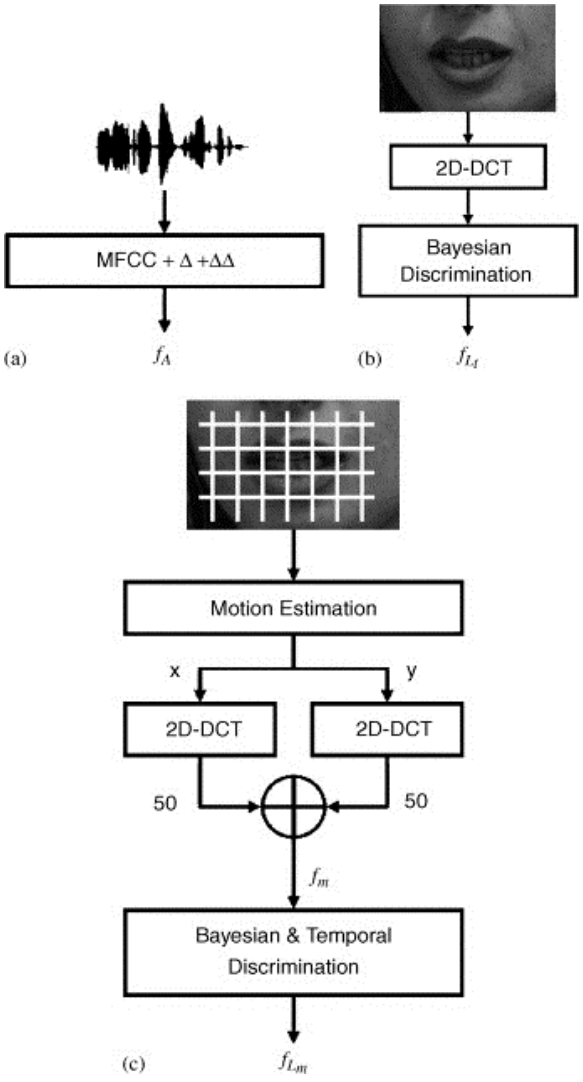


fig.3 Block Based Motion Estimation Analysis

H.263, MPEG-1, MPEG-2, MPEG-4 and H.264 [1] due to its effectiveness and simple implementation. The most straightforward BMA is the full search (FS), which exhaustively evaluates all the possible Candidate blocks within the search window. However, this method is very computationally intensive, and can consume up to 80% of the computational power of the encoder. This limitation makes Motion Estimation the main bottleneck in real-time video coding applications including lip reading systems. Consequently, fast BMAs are used to decrease the computational cost with the expense of less accuracy in determining the correct motion vectors. Many fast BMAs were proposed, such as three-step search (TSS), four-step search (4SS), block-based diamond search (DS) algorithms, etc.

**6. SUMMARY**

Both lip contour and image transform based visual features are considered for HMM based automatic lip reading. The latter are shown to perform significantly better, while being robust to video degradations, and they result in high visual-only recognition performance on single speaker, multi-speaker, and speaker-independent tasks. The superiority of image transform based features is not surprising, since significant speech reading information lies within the oral cavity that cannot be captured by the lip contours. In addition, lip contour estimation errors compromise the recognition accuracy. As is clear from our experiments, a number of image transform based features resulting similar lip reading performance. Given the fact that PCA requires an intensive training phase and is not amenable to fast implementation, the use of DWT or DCT based features is recommended. Finally, the demonstrated robustness of image transform features to video degradations is promising for low cost lip reading system implementation. In the first approach, the speaker's lip contours are extracted from the image sequence. A parametric [2], [9] or statistical [10] lip contour model is then obtained, and the model parameters are used as visual features. Alternatively, lip contour geometric features are used [4], [5]. In the second approach, the entire image containing the speaker's mouth is considered as informative for lip reading, and appropriate transformations of its pixel values are used as visual features [1], [3], [5]-[10]. feature approaches. We can define a number of lip contours based visual features that our experiments have demonstrated to be successful in lip reading.

We have considered various linear image transforms for feature extraction, among which the discrete wavelet transform (DWT)[11], the discrete cosine transform [12], and a principal component analysis (PCA) based projection [13] performed the best.

**REFERENCES**

[1]. Hanan Mahmoud "Reducing Shoulder-surfing by Using Silent Speech Password Entry "Technical report, KSU, Center of Excellence in Information Assurance, November 2008.

[2]. C. Miyajima, K. Tokuda, and T. Kitamura,"Audio-Visual Speech Recognition Using MCE-Based HMMs and Model-Dependent Stream weights," in Proc. ICSLP2000, vol. 2, 2000, pp. 1023–1026.

[3]. K. Iwano S. Furui T. Yoshinaga, S. Tamura,"Audio-visual speech recognition using lip movement extracted from side-face images,"Proc. Auditory Visual Speech Processing (AVSP), pp. 117–120, 2003.

[4]. G. Potamianos P. Lucey, "Lip reading using profile versus frontal views," IEEE Multimedia Signal Processing Workshop, pp. 24–28, October 2006.

[5]. T. Chen, "Audiovisual speech processing. lip reading and lip synchronization," IEEE Signal Processing Mag., vol. 18, pp. 9–21, January 2001.

[6]. Khaled Alghathbar, Hanan A. Mahmoud ISSN: 1790-0832 837 Issue 5, Volume 6, May 2009,Mase, K., and Pentland, A., "Automatic lipreading by optical flow analysis," Systems and Computers in Japan, vol. 22, no. 6, pp. 67-75,1991.

[7]. X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic speech reading with applications to human-computer interfaces," EURASIP J. Appl. Signal Process.,pp. 1228–1247, 2002.

[8]. J. F. G. Perez, A. F. Frangi, E. L. Solano, and K. Lukas, "Lip reading for robust speech recognition on embedded devices," in Proc. Int.Conf. Acoustics, Speech and Signal Processing,
2005, vol. I, pp. 473–476.

[9]. Matthews, T.F. Cootes, J.A. Bangham, S.Cox, and Harvey, "Extraction of visual features for lip reading," IEEE Trans. PatternAnal. Mach. Intell., vol. 24, no. 2, pp. 198–13 , Feb. 2002.

[10]. K. Iwano, S. Tamura, and S. Furui, "Bimodal Speech Recognition Using Lip Movement Measured by Optical-Flow Analysis," in Proc.HSC2001, 2001, pp. 187–190.

[11]. Kulkarni, H. Gunturu, And S. Datla,"Association-Based Image Retrieval," WSEAS Trans. On Signal Processing, Issue 4, Volume 4, April 2008, pp. 183–189.

[12] M.Tun, K.K loo and J. Cosmas, "Semi Hierarchical Based Motion Estimation Algorithm for the Dirac Video Encoder," WSEAS Trans. On Signal Processing, Issue 5,Volume 4, May 2008, pp. 261–270.

[13] H. Nam, S. Lim, "A New Motion Estimation Scheme Using a Fast and Robust Block Matching Algorithm" WSEAS ans. On Information Science & Applications, Issue 11, Volume 3, November 2006, pp. 2292-2299.

[14] Y. Shi, C. Yi and Z. Cai, "Multi-Direction Cross-Hexagonal Search Algorithms for Fast Block Motion Estimation," WSEAS Trans. On Computers, Issue 6, Volume 6, June 2007, pp.959-963.

[15] C. L. Lin, J. J. Leou, "An Adaptive Fast Search Motion Estimation Algorithm for H.264"WSEAS Trans. on Communications, Issue 7,Volume 4, July 2005, pp. 396-406. WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS

[16] Khaled Alghathbar, Hanan A.Mahmoud ISSN: G. Potamianos, H. P. Graf, and E. Cosatto,"An image transform approach for hmm based automatic lip reading," in Proc. of the Int. Conf. on Image Proc. (ICIP), Chicago, 1998.