Block-Based Motion Estimation Analysis for Lip Reading User Authentication Systems

KHALED ALGHATHBAR HANAN A. MAHMOUD

Centre of Excellence in Information Assurance,
King Saud University, Riyadh,
Kingdom of Saudi Arabia

ghathbar@coeia.edu.sa hanan.hosni@coeia.edu.sa

Abstract: - This paper proposes a lip reading technique for speech recognition by using motion estimation analysis. The method described in this paper represents a sub-system of the Silent Pass project. Silent Pass is a lip reading password entry system for security applications. It presents a user authentication system based on password lip reading. Motion estimation is done for lip movement image sequences representing speech. In this methodology, the motion estimation is computed without extracting the speaker's lip contours and location. This leads to obtaining robust visual features for lip movements representing utterances. Our methodology comprises of two phases, a training phase and a recognition phase. In both phases an n x n video frame of the image sequence for an utterance (can be an alphanumeric character, word or a sentence in more complicated analysis) is divided into m x m blocks. Our method calculates and fits eight curves for each frame. Each curve represents motion estimation of this frame in a specific direction. These eight curves are representing set of features of a specific frame and are extracted in an unsupervised manner. The feature set consists of the integral values of the motion estimation. These features are expected to be extremely effective in the training phase. The feature sets are used to characterize specific utterances with no additional acoustic feature set. A corpus of utterances and their motion estimation features are built in the training phase. The recognition phase is accomplished by extracting the feature set, from the new image sequence of lip movement of an utterance, and compare it to the corpus using the mean square error metric for recognition.

Key-Words: - Lip reading, Speech recognition, Motion estimation, User authentication, Feature Extraction.

1 Introduction

Automatic Speech Recognition (ASR) systems are playing successful roles in an recognizing speech with high accuracy rates [1]. 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 speech recognition system which combines both auditory and visual information has been demonstrated to outperform the audio-only system. Most of the multi-modal speech recognition methods use visual features, typically lip information, in addition to the acoustic features [3]. Visual information aids in distinguishing acoustically similar sounds, such as nasal sounds: /n/, /m/, and /ng/ [7, 10]. Lip-reading has become a hot topic for human computer interaction (HCI) and audio-visual speech recognition (AVSR). Lip reading systems can be utilized in many application such as hearing impaired aid and for noisy environment, where speech is highly unrecognizable and lately as password entry scheme as suggested by our research listed in [2]. This paper concentrates on the visual-only lip-reading system which has also attracted significant interest. Feature extraction is a crucial part for a lip-reading system. Various visual features have been proposed in the literature. 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. A typical method to extract pixel based features are image transforms such as Discrete Cosine Transform (DCT) [7, 8, 10-12], Principal Component Analysis (PCA) [5-9,13] , Discrete Wavelet Transform (DWT) [7] and Linear Discriminate Analysis (LDA) [8] have been

employed for lip-reading. Other feature extraction methods utilize motion analysis of image sequences representing lip movement while uttering some speech are previously done. Mase et al reported their lip-reading system for recognizing connected English digits using an optical-flow analysis [10]. Various advantages exist in using the optical-flow analysis for audio-visual bimodal speech recognition. The optical-flow vectors are calculated without using any prior knowledge about the shape of the object [11], [12]. Thus the visual features can be detected without extracting the lip locations and contours [13], [14]. But such method suffers from main disadvantages of the optical flow analysis methodology [15],[16].

In this paper, we propose using the 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. This paper is organized as follows: Previous work is shown in Section 2. The principle of the motion estimation method is explained in section 3. The proposed lip-reading speech recognition system is described in Section 4. Experimental setup and results are shown in Section 5. Finally, conclusion and future works will follow in Section 6.

2 Previous Work

There has been much research in recent years in building automatic visual speech recognition systems (automated lip-reading) . Many systems have been designed to show that speech recognition is possible using only the visual information. Some researchers have done comparisons on a number of visual feature sets in attempts to find those features that yield the best recognition performance. Such systems are either speaker dependent or speaker independent systems. In literature experimentations are done by examining isolated vowels, CVC syllables (consonant-vowel-consonant) , isolated words, connected digits and continuous speech [1-3]. The recognition engine takes many forms, some recognition systems were based on dynamic time warping. Others use neural network architectures. An increasing number of systems that rely on hidden Markov models (HMM) are built. Petajan developed one of the first audio-visual recognition systems [6]. In this system, a camera captures the mouth image and thresholds it into two levels, mouth images are analyzed to derive the mouth open area, the perimeter, the height and the width. In this system, the speech is processed by the acoustic recognizer to produce the first few candidate words, and then these words are passed on to the visual recognizer for final decision. Later, the system was modified by using dynamic time warping, where a number of features of the binary images such as height, width, and perimeter, along with derivatives of these quantities are used as the input to an HMM based visual recognition system. He sought to find the combination of parameters that would lead to the best recognition performance. The feature set he settled upon was dominated by derivative information. This showed that the dynamics of the visual feature set is important for speech recognition. The same conclusion had been reached in a study that used optical flow as input for a visual speech recognizer [8]. In [7] it was shown that the physical dimensions of the mouth can provide good recognition performance. They analyzed recognition performance of VCV syllables. They placed reflective markers on the speaker's mouth, and used these to extract 14 distances, sampled at 30 frames a second. They experimented with both a mean squared error distance, and an optimized weighted mean squared error distance. Equal weighting of the parameters led to a 78% viseme recognition rate. In [3], 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 researcher combined 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]. Researchers have also tried to convert mouth movements into spoken speech directly. In [6], a system called "image-input microphone" 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 was 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 Motion Estimation Analysis

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 a reference frame. Motion

estimation tries to identify this best matching region in the reference frame for every block in the current frame.

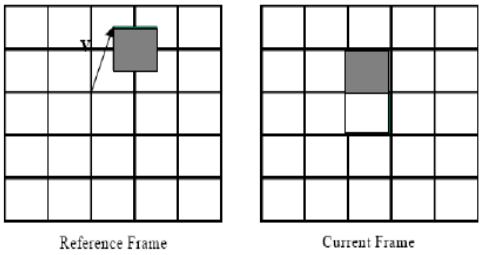

Fig. 1 Block Based Motion Estimation

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, 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 ME 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. In this paper, we are going to employ the different mentioned block based to induce the motion vector feature set. Performance evaluation of different techniques will be studied. Experiments using different training algorithms will be used Recognition error percentages will be presented for different methodologies.

3.1 Full-search block-matching algorithm

Full-search block matching algorithm (FSBM) finds the best match for a reference block in the current frame within a search area *S* in the previous frame. The criterion for best match is the candidate block with the minimum amount of distortion when compared with the reference block. The measure used for calculating distortion is the sum of absolute differences (*SAD)* of intensity values between the two blocks. The SAD for the candidate block of size *N* x *N* at position *(u,v)* can be defined as:

$$SAD(u,v) = \sum_{i=1}^{N}\sum_{j=1}^{N} \left| u(i+u, j+v) - v(i,j) \right| \quad (1)$$

Where v(i,j) and u(i+u, j+v) are intensity values at position (i,j) of the reference block and (i+u, j+v) of the candidate block in search area *S*. The search area is formed by extending the reference block by a search range *w* on each side forming a search area of $(2w+N)^2$ pixels. As a result, there are *(2w +1)* candidate blocks in both horizontal and vertical directions i.e. a total of $(2w+1)^2$ candidate blocks have to be searched corresponding to each reference block. The distortion value is computed for each candidate block and the minimum value SAD_{min} is found. The block matching process generates a motion vector $(u,v)_{min}$ and the corresponding distortion value SAD_{min}.

4 Visual Lip Reading System

In this section we are going to discuss the proposed technique for lip reading. The proposed technique is composed of two phases. The first phase is the training phase which results in feature extraction from image sequences representing different utterance lip movement. The second phase is the recognition phase, where new utterance through lip movement is compared against the output of the training phase and recognized.

4.1 Training phase and feature extraction

An image sequence is captured with the frame rate of 30 frames/sec and the resolution size of 360×240. Block-based motion estimation technique is used. Motion vectors representing motion of block is computed from a pair of consecutive images. We extract the motion vectors of the different blocks. The feature extraction of the training phase is illustrated in the following algorithm. The used

block matching algorithm can be full search or 3SS or FSS or DS. The previously mentioned algorithm will produce motion vectors with values from -3 to +3 in one of the eight geographical directions. This restriction in defining motion vectors is due to the fact that lip motion in utterance is very restricted at the rate of 15 frames per second, motion is very slow.

The diagram in fig. 2 illustrates block 1 in the training phase algorithm: Algorithm Training (W). Each video frame, of the utterance lip movement sequence, is fed to the frame division module into 8X8 blocks. Each of these blocks ($block_1$ to $block_n$) are fed to the motion estimation module to for motion analysis and production of the motion vector of such block. a set of n motion vectors $\{mv(block_1), \ldots. mv(block_n)\}$ are produced.

Many videos for the same utterance are fed iteratively to Block 1 to calculate average motion vectors for each block in a frame of these videos. The set of the average motion vectors are fed to Block 2. Where eight curves are built from the average motion vectors for each frame. Each curve is the value of the average motion vector of a block versus its particular location in a video frame. Each curve represents the average motion of blocks in the video frame in a certain direction (we restricted the directions to the eight geographical directions) and this curve is fitted to be a continuous curve. Such curve represents a motion feature of the utterance video. The motion feature will be represented by the area under this curve by taking the integration value of this curve, which is done by the area calculation module, the area is calculated as in equation 1. Number of motion features for an utterance video is equal to **8f** ,where **f** equals to the number of frames in the utterance video. Each utterance and its motion features are stored in the utterance database.

Algorithm Training(Word: W)

- **Begin**
- For word W repeat the following steps j times by different speakers

{

1. Record a video of lip reading the word W by a speaker S_j;
2. Divide the video into n frames;
3. Do for (k= 0, k=k+2, k< n)
   {
   a. Frame k is divided into m blocks each of size 8X8 pixels;
   b. Motion vectors, M, of all the m blocks are calculated between frame k and frame k+1;M = set of motion vectors = $\{mv_i , i= 1 \text{ to } 8\}$, i is one of the eight principle geographical directions;
   c. for i= 1 to 8 do
   { 1. Draw and fit a discreet graph: $DG_i(k)$ between mv_i and the location of the block, location of the block is numbered in a spiral fashion starting from the center of each 64X64 block;

$$2.\ Area_i(k) = \oint DG_i(k) \qquad (2);$$

   }
   }
4. feature $set_j = \{Area_i^j(k),\ i = 1 \text{ to } 8 , \forall k\}$

}

- Training-feature-set (W) = $\{average\,(Area_i^j(k)),\ i = 1 \text{ to } 8 , \forall k\}$
- **End**

4.2 Explanation of the training algorithm

1. The video sequence of the lip read word has n frames, each frame is divided into blocks of 8X8 pixels for the motion vectors calculations, as we assume that an 8X8 block moves translations motion as a one unit.
2. To draw the graphs $DG_i(k)$, a frame is divided into blocks each of 64X64 pixels, or 8X8 blocks of size 8X8 pixels. Location of the block is numbered in a spiral fashion starting from the center of each 64X64 block.
3. There are 8 curves representing the motion vectors of the video sequence.
4. To calculate the integration of each curve, area under the curve is calculated by an approximate method.

4.3 Lip Reading Recognition Phase

The lip reading phase is similar to the training phase, it starts with the unknown utterance lip movement video, these utterance should be recognized. This video is fed to Block 1, where motion vectors for this video are calculated and fitted into 8f curves. Integral values of these curves are the areas under these curves and are fed to the comparison module. The comparison module compares the motion features of the input video

against the utterance stored in the database by using the mean square error function. MSE is calculated between the input utterance features and every utterance features stored in the database. MSE is calculated as shown in equation 2. MMSE is computed to choose the candidate utterance that is most similar to the input utterance. MMSE is calculated as in equation 3.

Algorithm Recognize (WORD: video)

Begin
Divide the WORD video into n frames;
1. Do for (k= 0, k=k+2, k< n)
{
a. Frame k is divided into m blocks each of size 8X8 pixels;
b. Motion vectors, M, of all the m blocks are calculated between frame k and frame k+1; M = set of motion vectors = $\{mv_i, i= 1 \text{ to } 8\}$, i is one of the eight principle geographical directions;
c. for i= 1 to 8 do{
1. Draw and fit a discreet graph: $DG_i(k)$ between mv_i and the location of the block, location of the block is numbered in a spiral fashion starting from the center of each 64X64 block;

2. $Area_i(k) = \oint DG_i(k)$;
}
}
2. feature (WORD)= $\{fArea_i(k), i = 1 \text{ to } 8, \forall k\}$
3. Calculate mean square error MSE_j for the input utterance and utterance j as follows:

$$MSE_j = \frac{1}{x} \sum_{i=1} \sum_{k=1} (fArea_i(k) - Area_i(k))^2 \qquad (3)$$

Calculate minimum mean square error MMSE as follows:

$$MMSE = \min(MSE_j) \qquad j = 1 \ to \ x \qquad (4)$$

4. If MMSE > threshold then the WORD is unrecognizable otherwise the WORD = W_d (W_d is a word in the corpus corresponds to the minimum MSE)
End

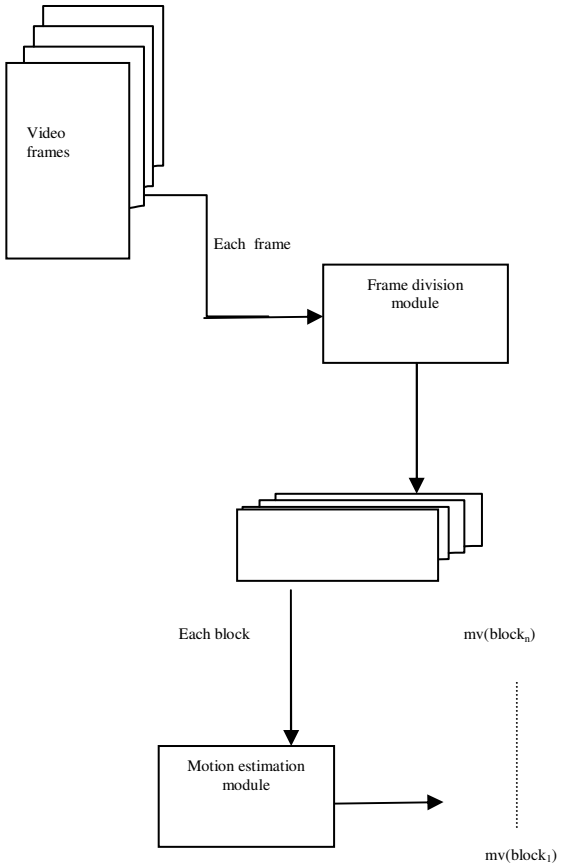

Fig. 2.a Block1 (Motion estimation and motion vector extraction for one video sequence)

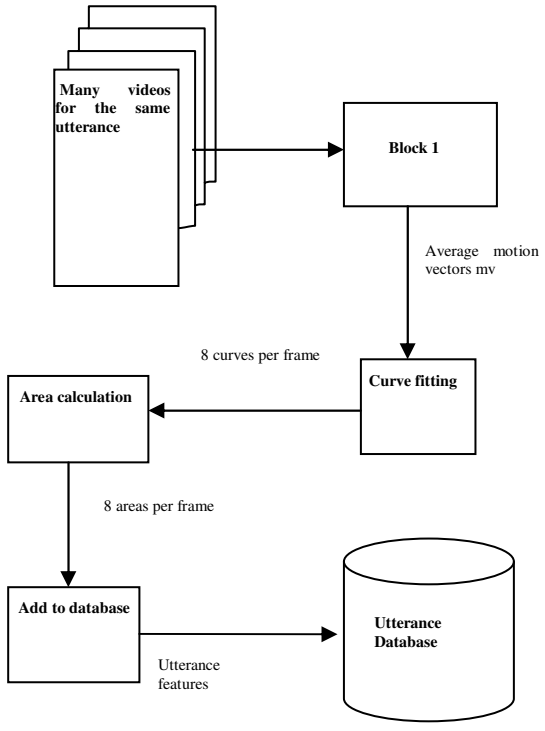

Fig. 2.b Populating the utterance database

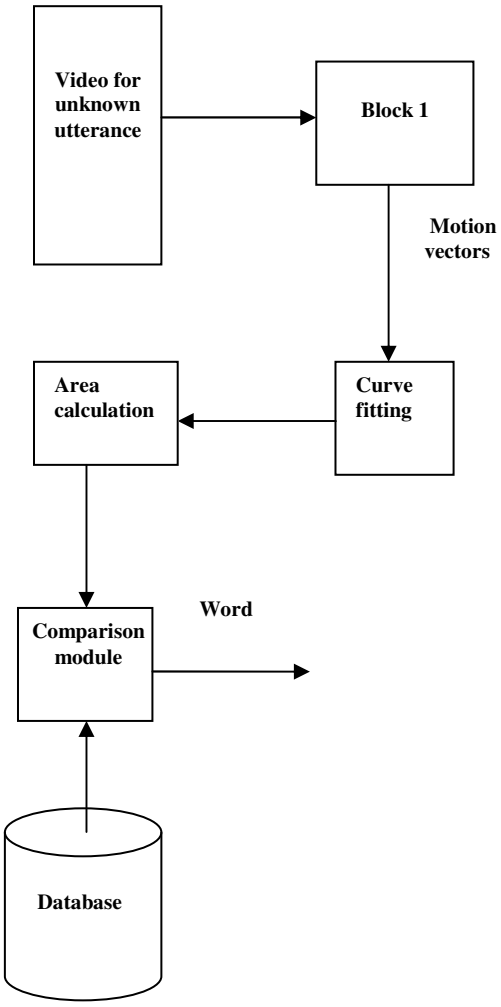

Fig. 3 Lip Reading Recognition Phase

4. 4 Limitations

Data corpora are an important part of any visual speech recognition research such as lip reading research. However, because the field is still new, and time to record a visual or bimodal data corpus can be overwhelming, the number of existing visual data corpus for lip reading systems is small. Having a good data corpus might be of great help for the researchers in this field. Data corpora are also developed to be shared between different researchers in order to have the means for comparison of their results, so a greater level of reusability is required. There are a number of limitations that a visual dataset has, such as:

• The recordings contain only a small number of respondents. This greatly reduces the generality of the results, since it generally generates highly undertrained systems.

• The pool of utterances is usually very small.

• They usually contain only isolated words or digits or even only the letters of the alphabet rather than continuous speech.

• They have a poor coverage of the set of visemes in the language. Using utterances that are rich in visemes should be a strong requirement especially for the cases when the dataset is intended for lip reading.

For this research we used the database Tulips 1.0 which is a small audiovisual database of 12 subjects saying the first 4 digits in English. The database was compiled at Javier R. Movellan's laboratory at the Department of Cognitive Science, UCSD. Tulips1.0 was used in this research only for lip reading, therefore only visual data is used.

5 Experimental Results

Due to the novelty of the approach we were not able to perform tests on a large database. Instead we used the audio-visual Tulips 1 database [7]which was recorded for speech reading research. It consists of 96 grey-level image sequences of 12 speakers (9 male, 3 female) each uttering the first four digits in English twice. The images contain only the mouth area of the speakers and are digitized at 30 frames/sec, 360 x 240 pixels and 8 bits/pixel. We used the first utterance of each word and each speaker as the training set for the motion vector feature set extraction training phase. The second instances were used as the test set for testing the recognition phase. In the end, a total of 8 features/block, with a total of 8 curves/utterance were fed to the recognizer in the form of areas. G is the number of blocks in each video and is calculated equation 5 as follows.

$$G = ((S / b) \times t \times f) \quad (5)$$

Where,

S is the size of the frame,

b is the size of the block,

t is the duration t ime of the video of the utterance,

f is the number of frames per second,

In fig.4, the eight curves for the digit “2” is shown. These curves are representing the Average motion vector feature set (MVFS). The curve is the motion vector value versus the block location after fitting the discreet curve into a continuous curve. These curves are built using several training videos for the same utterance.

The experiments are designed as follows, the training phase is carried using full search block-

based motion estimation techniques (FS), since they are done offline and doesn't have timing constraints. The recognition phase is done for each digit using full search (FS), three step search (3SS), four step search 4SS, and diamond search (DS). Error rates are calculated for each case. Error rates are displayed in fig. 5 and 6.

Fig. 5.a, Fig. 5.b, Fig. 5.c and Fig. 5.e are scaled using logarithmic scale for clarity. Fig. 5.d uses "6" as the highest mark on the Y axis.

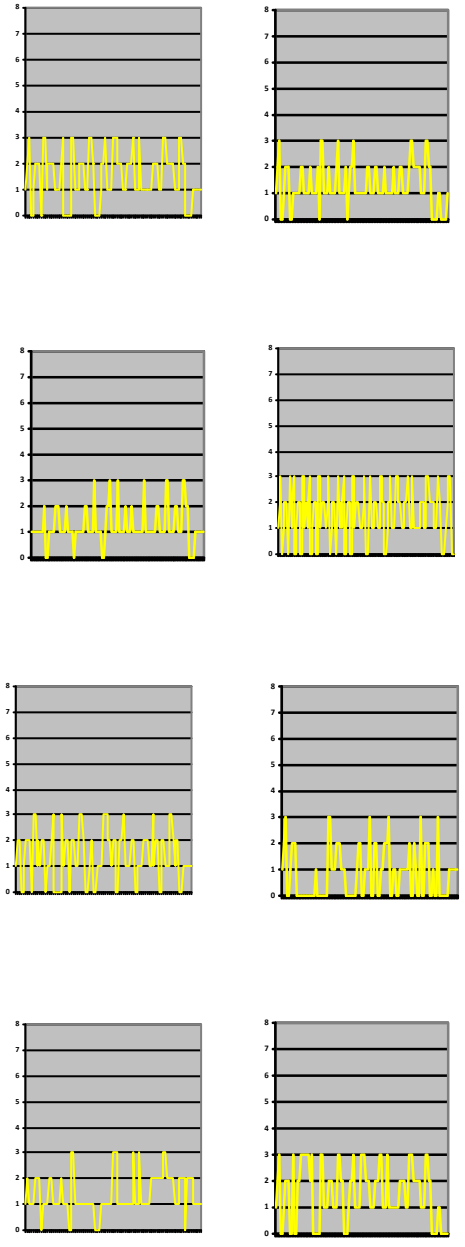

Fig. 4 The eight curves for the digit "2 of the average motion vector feature set (MVFS)

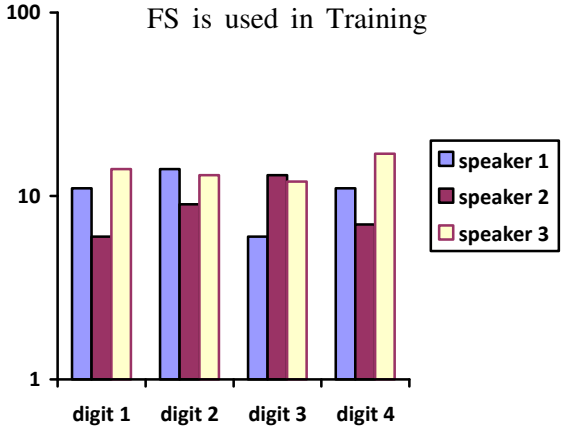

Fig. 5.a The average number of errors in digit recognition within 100 utterances per speaker using 3SS

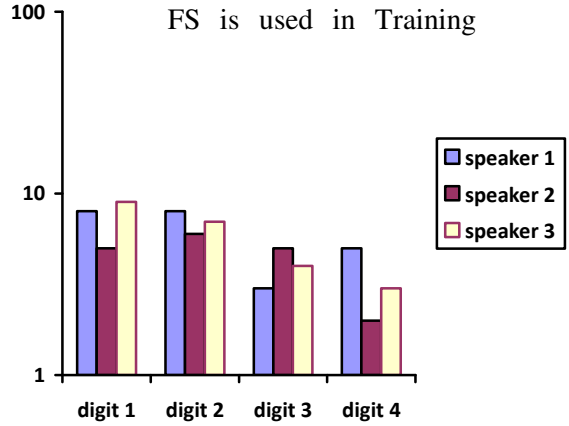

Fig. 5.b The average number of errors in digit recognition within 100 utterances per speaker using 4SS

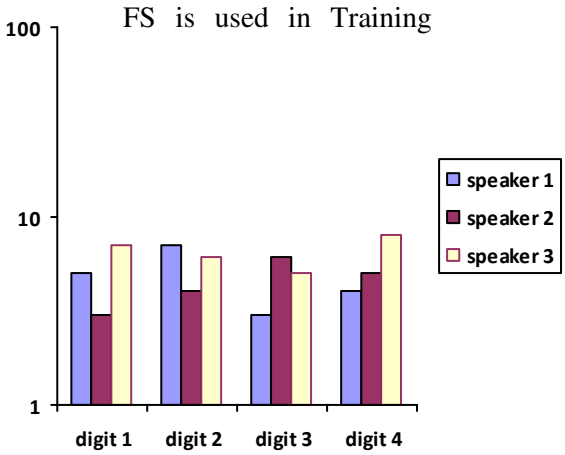

Fig. 5.c The average number of errors in digit recognition within 100 utterances per speaker using DS

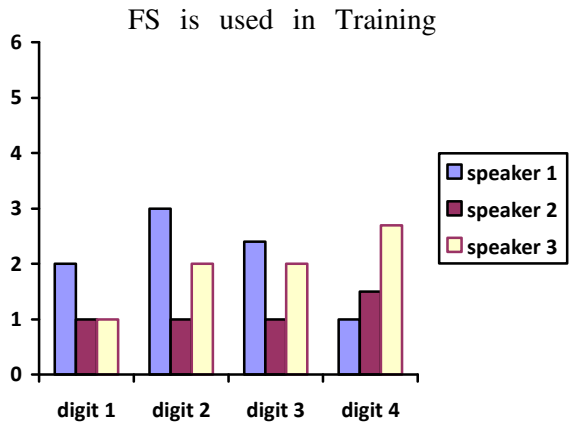

Fig. 5.d The average number of errors in digit recognition within 100 utterances per speaker using FS

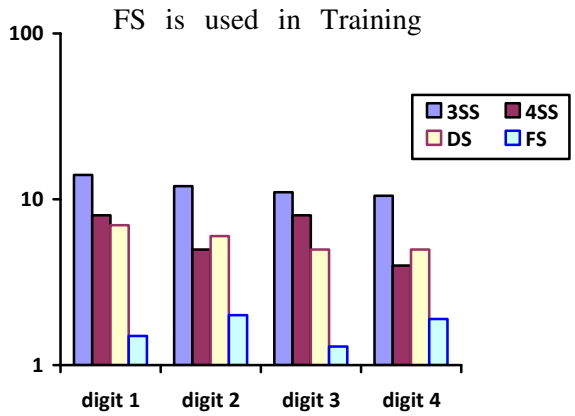

Fig. 5.e Comparison of the average number of errors for the three speakers with 100 utterances per speaker using 3SS, 4SS, DS and FS

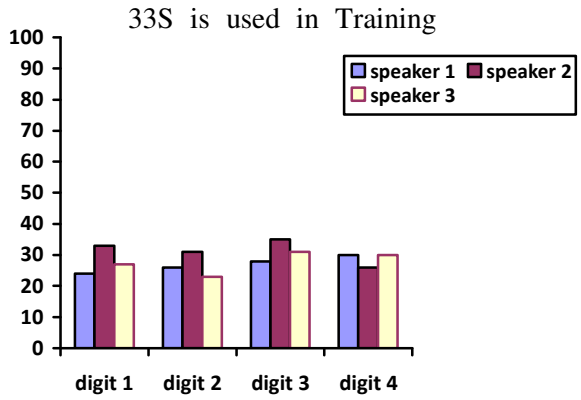

Fig. 6.a The average number of errors in digit recognition within 100 utterances per speaker using 3SS

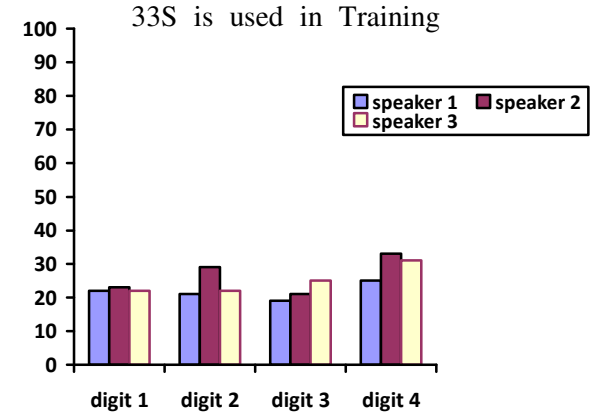

Fig. 6.b The average number of errors in digit recognition within 100 utterances per speaker using 4SS

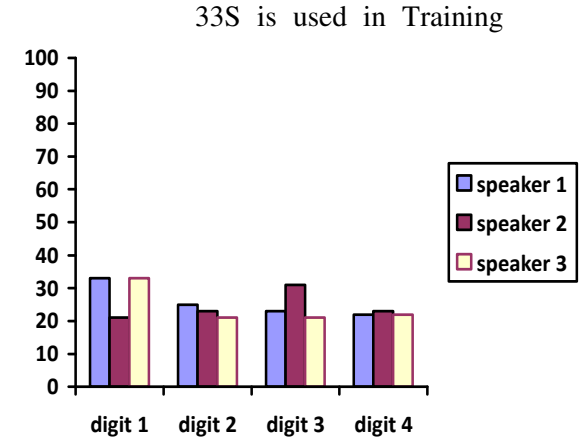

Fig. 6.c The average number of errors in digit recognition within 100 utterances per speaker using DS

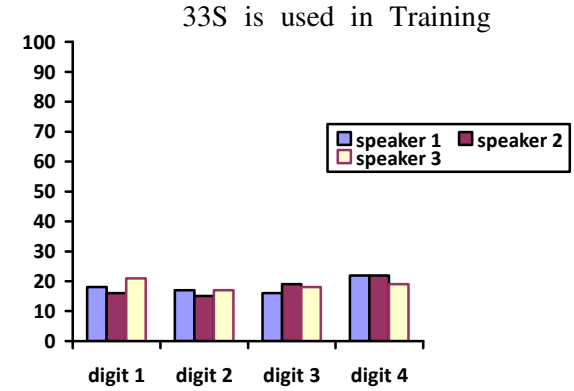

Fig. 6.d The average number of errors in digit recognition within 100 utterances per speaker using FS

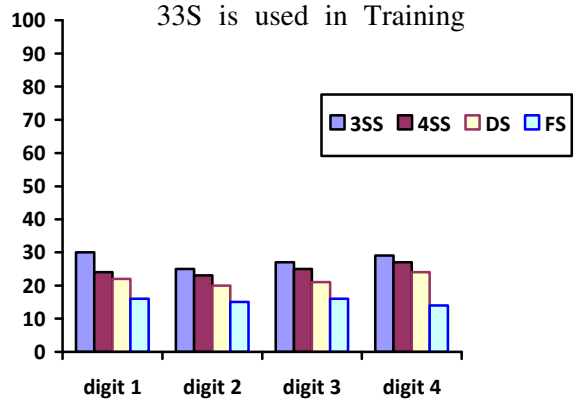

Fig. 6.e Comparison of the average number of errors for the three speakers with 100 utterances per speaker using 3SS, 4SS, DS and FS

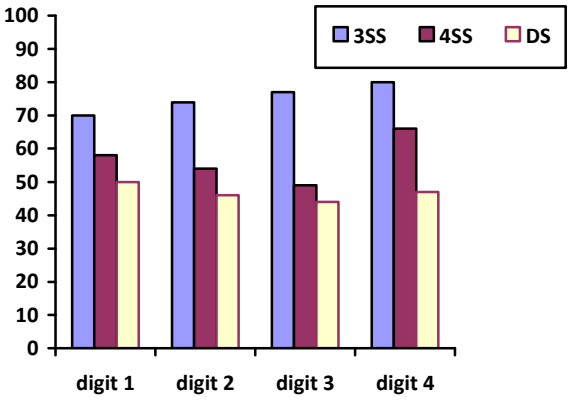

Fig. 7 Comparison of the average speedup of utterances' recognition using 3SS, 4SS, DS as a percentile of the speed of the FS

6 Conclusions

The method described in this paper represents a sub-system of the Silent Pass project [2]. This is a lip reading password entry system and person authentication project for security applications. The aim is to provide a complete solution of secured access to ATM and internet services in multi-media environments. The methodology is expected to improve the performance of authentication systems and to reduce the acceptance of impostors and shoulder surfing threats. The following issues are addressed speaker recognition and lip motion analysis. A database was recorded for this project which contains few utterances and their motion vector extracted features. The database contains motion vector extracted features using three step search and full search block based motion estimation techniques.

In this paper we proposed a visual speech recognition scheme (lip reading system) using motion estimation analysis of lip movements. Visual feature sets were extracted blindly without having to detect or track the lips. The visual features were used in both training and testing. The experimental results show that the proposed method achieves approximately 20% improvement when using the integral information of motion vectors extracted using full search block based motion estimation technique over using the 3SS in the training phase. Our future works include: (1) investigation of more robust and informative visual parameters, such as features including direction and amount of lip movements, (2) investigation of different block sizes with the performance of the recognition system, maybe the usage of deformable motion estimation technique would prove better performance, and (3) collecting a new database with more utterances and continues lip reading for long sentences and detect silence to separate words.

References:

[1] S. Furui, "Speech Recognition Technology in the Ubiquitous/ Wearable Computing Environment," in *Proc. ICASSP2000*, vol. 6, 2000, pp. 3735–3738.

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

[3] 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.

[4] 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.

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

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

[7] G. Gravier A. Garg G. Potamianos, C. Neti and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. Of the IEEE*, vol. 91, 2003.

[8] Mase, K., and Pentland, A., "Automatic lip reading by optical flow analysis," *Systems and Computers in Japan*, vol. 22, no. 6, pp. 67-75, 1991.

[9] 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.

[10] 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.

[11] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lip reading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[12] 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.

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

[14] 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.

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

[16] 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.

[17] 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.