A Novel Algorithm for Acoustic and Visual Classifiers Decision Fusion in Audio-Visual Speech Recognition System

R. Rajavel rettyraja@gmail.com
Research Scholar, ECE Department,
National Institute of Technology Calicut,
Kerala-673601, India

P.S. Sathidevi sathi@nitc.ac.in
Professor, ECE Department,
National Institute of Technology Calicut,
Kerala-673601, India

Abstract

Audio-visual speech recognition (AVSR) using acoustic and visual signals of speech has received attention recently because of its robustness in noisy environments. Perceptual studies also support this approach by emphasizing the importance of visual information for speech recognition in humans. An important issue in decision fusion based AVSR system is the determination of the appropriate integration weight for the speech modalities to integrate and ensure the combined AVSR system's performances better than that of the audio-only and visual-only systems under various noise conditions. To solve this issue, we present a genetic algorithm (GA) based optimization scheme to obtain the appropriate integration weight from the relative reliability of each modality. The performance of the proposed GA optimized reliability-ratio based weight estimation scheme is demonstrated via single speaker, mobile functions isolated word recognition experiments. The results show that the proposed scheme improves robust recognition accuracy over the conventional uni-modal systems and the baseline reliability ratio-based AVSR system under various signals to noise ratio conditions.

Key words: Audio-visual speech recognition, side face visual feature extraction, audio visual decision fusion, Reliability-ratio based weight optimization, late integration

1. INTRODUCTION

Many researchers were trying to design automatic speech recognition (ASR) systems which can understand human speech and respond accordingly [16]. However, the performances of the past and current ASR systems are still far behind as compared to human's cognitive ability in perceiving and understanding speech [18]. The weaknesses of most modern ASR systems are their inability to cope robustly with audio corruption which can arise from various sources, for example environment noises such as engine noise or other people speaking, reverberation effects or transmission channel distortion etc. Thus one of the main challenges being faced by the ASR research community is how to develop ASR systems which are more robust to these kinds of corruptions that are typically encountered in real-world situations. One approach to this problem is to introduce another modality to complement the acoustic speech information which will be invariant to these sources of corruptions [18]. Visual speech is one such source, obviously not perturbed by the acoustic environment and noise. Such systems that combine the audio and visual modalities to identify the utterances are known as audio-visual speech recognition systems [18]. The first AVSR system was reported in 1984 by Petajan [19]. During the last decade more than hundred articles have appeared on AVSR [5, 6, 13, 18]. AVSR systems can enhance the

performance of the conventional ASR not only under noisy conditions but also in clean condition when the talking face is visible [20]. The major advantage of utilizing the acoustic and the visual modalities for speech understanding comes from "Complementarity" of the two modalities: The two pronunciations /b/ and /p/ are easily distinguishable with the acoustic signal, but not with the visual signal; on the other hand, the pronunciations /b/ and /g/ can be easily distinguished visually, but not acoustically [21] and, "synergy" : Performance of audio-visual speech perception can outperform those of acoustic-only and visual-only perception for diverse noise conditions [22]. Generally, the AVSR systems work by the following procedures. First, the acoustic and the visual signals of speech are recorded by a microphone and a camera, respectively. Then, each signal is converted into an appropriate form of compact features. Finally, the two modalities are integrated for recognition of the given speech. The integration can take place either before the two information sources are processed by a recognizer (early integration/feature fusion) or after they are classified independently (late integration/decision fusion). Some studies are in favor of early integration [1, 5, 6, 7, 13, 23], and other prefers late integration [2, 3, 4, 5, 7, 23, 24]. Despite of all these studies, which underline the fact that speech reading is part of speech recognition in humans, still it is not well understood when and how the acoustic and visual information are integrated. This paper takes the advantages of late integration on practical implementation issue to construct a robust AVSR system. The integration weight which determines the amount of contribution from each modality in decision fusion AVSR is calculated from the relative reliability measure of the two modalities [32]. In this work, the integration weight calculated from the reliabilities of each modality is optimized against the recognition accuracy using genetic algorithm. The performance of the proposed GA optimized reliability ratio-based weight estimation scheme is demonstrated via single speaker, mobile functions isolated word recognition experiments. An outline of the remainder of the paper is as follows. The following section explains the integration schemes in AVSR and the reason for decision fusion in this work. Section 3 describes our own recorded experimental database, audio and visual feature extraction schemes. How Genetic Algorithm can be used to obtain the appropriate integration weight from the relative reliability of two modalities for decision fusion is explained in section 4. Section 5 discusses the HMM training and recognition results. The discussion, conclusion and future direction of this work are outlined in the last section.

2. APPROACHES FOR INFORMATION FUSION IN AVSR

The primary focus of AVSR is to obtain the recognition performance which is equal to or better than the performance of any individual modality for various SNR conditions. Secondly, the use of audio-visual information for speech recognition is to improve the recognition performance with as high synergy of the modalities as possible [2]. Generally, while combining two modalities, the integrated system should show high synergy effect for a wide range of SNR conditions. On the contrary, when the fusion is not performed appropriately, we cannot expect complementarity and synergy of the two information sources and moreover, the integrated recognition performance may be even inferior to that of any of the uni-modal systems, which is called "attenuating fusion" [25].

In general, the audio-visual information fusion can be categorized into feature fusion (or early integration) and decision fusion (or late integration), which are shown in figure 1. In feature fusion approach the features of two modalities are concatenated before given to the classifier for recognition, where as in decision fusion approach, the features of each modality are used for recognition separately and, then the outputs of the two classifiers are combined for the final recognition result [2]. Each approach has its own advantages and disadvantages. Most of the audio-visual speech recognition systems [1, 5, 6, 7, 13] are based on feature fusion. The main attraction of this approach is its computational tractability, since only a single classifier is used, and that existing procedures for training and testing of HMMs can be applied without significant modification [4, 26]. There are many advantages in implementing a noise-robust AVSR system using decision fusion. First, in the decision fusion approach it is relatively easy to employ an adaptive weighting scheme for controlling the amounts of the contributions of the two modalities to the final recognition [2, 5]. Second, the decision fusion allows flexible modeling of the temporal coherence of the two information streams, whereas the feature fusion assumes a perfect synchrony between the acoustic and the visual feature sequences [2]. It is proved [27] that there exists an asynchronous characteristic between the acoustic and the visual speech: The lips and the tongue sometimes start to

move up to several hundred milliseconds before the acoustic speech. Finally and most importantly, in the feature fusion approach the combination of the acoustic and the visual features results in high dimensional data sets, which makes training HMMs difficult. Since we have very limited training samples, practical implementation of feature fusion is impossible. Hence this work focuses on the decision fusion for AVSR system

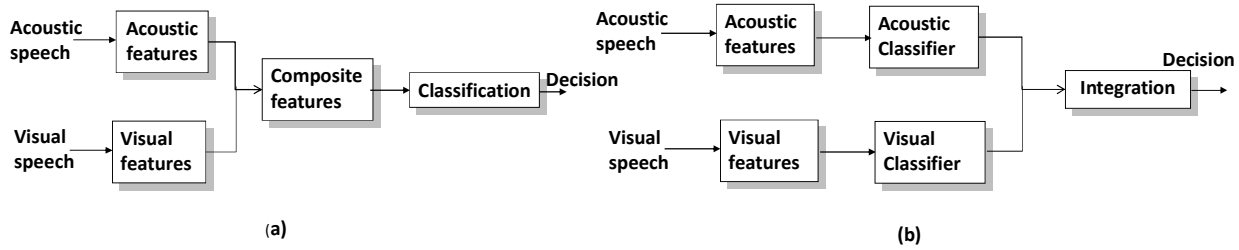

FIGURE 1: AVSR integration schemes. (a) Feature fusion. (b) Decision fusion

3. AUDIO-VISUAL FEATURES EXTRACTION SCHEMES

3.1. Experimental database

This paper focuses on a slightly different type of AVSR system which is mainly useful for mobile applications. Most of the past and current AVSR systems [2, 3, 4, 5, 6, 7] use the mouth information extracted from frontal images of the face, but these systems cannot be used directly for mobile applications because the user needs to hold a handset with a camera in front of their mouth at some distance, which may be unnatural and inconvenient for conversation. As the distance between the mouth and the mobile phone increases, SNR decreases which may worsen the recognition accuracy. If the mouth information can be taken by using a handset held in the usual way for telephone conversation this would greatly improve the usefulness of the system [1]. This paper focuses on this point of view and proposes an audio-visual speech recognition system using side-face images, assuming that a small camera can be installed near the microphone of the mobile device in the future.

Potamianos et al. has demonstrated that using mouth videos captured from cameras attached to wearable headsets produced better results as compared to full face videos [28]. With reference to the above, as well as to make the system more practical in real mobile application, around 70 commonly used mobile functions isolated words were recorded 25 times each by a microphone and web camera located approximately 10-15 cm away from single speaker's right cheek mouth region. Samples of the recorded side-face videos are shown in figure 2. Advantage of this kind of arrangement is that face detection, mouth location estimation and identification of the region of interest etc. are no longer required and thereby reducing the computational complexity [9]. Most of the audiovisual speech databases available are recorded in ideal studio environment with controlled lighting or kept some of the factors like background, illumination, distance between camera and speaker's mouth, view angle of the camera etc. as constant. But in this work, the recording was done purposefully in the office environment on different days with different values for the above factors and also to include natural environment noises such as fan noise, bird's sounds, sometimes other people speaking and shouting sounds.

FIGURE 2: Samples of recorded side-face images

3.2. Acoustic features extraction

Most of the speech recognition systems [1, 2, 3, 4, 5] use the so-called Mel Frequency Cepstral Coefficients (MFCC) and its derivatives as acoustic features for recognition since it shows good recognition performance. This work also adapts, MFCC and its derivatives as acoustic features for recognition. This section briefly reviews the MFCC feature extraction process.

Assume that $s(k)$ represents a speech signal that is multiplied by a hamming window $w(k)$ to obtain a short segment $V_m(k)$ of speech signal defined as:

$$V_m(k) = \begin{cases} s(k).w(k-m) & if \quad k = m,....m+1-N \\ 0 & else \end{cases} \qquad \text{----------} \qquad (1)$$

Where N is the window length and m is the overlapping segment length. [In this work N=256 samples (or 32ms) and m=100 samples (or 12.5ms) with the sampling frequency of fs=8000Hz]. The short speech segment $V_m(k)$ is transformed from time domain to frequency domain by applying an N-point Fast Fourier Transform (FFT). The resulting amplitude spectrum is | V (n) |. For further processing, only power spectrum of the signal is interested, which is computed by taking squares of | V (n) |. Since $V(n)$ is periodic and symmetry, only the values | V (n) |^2 . . . | V (N/2) |^2 are used, giving a total number of $N/2 + 1$ value. Next, the coefficients of the power spectrum | V (n) |^2 are transformed to reflect the frequency resolution of the human ear. A common way to do this is to use K triangle-shaped windows in the spectral domain to build a weighted sum over those power spectrum coefficients | V (n) |^2 which lie within the window. We denote the windowing coefficients as

$$\eta_{kn} \quad ; \quad k = 0,1,....k-1 \quad ; \quad n = 0,1,....N/2 \qquad \text{----------} \qquad (2)$$

In this work, the window coefficients are computed with fs=8000Hz, N=256, and K=22. This gives a new set of coefficients $G(k)$; k = 0, 1, ...K – 1 the so-called mel spectral coefficients

$$G(k) = \sum_{n=0}^{N/2} \eta_{kn} . \left| V(n) \right|^2 \quad ; \quad k = 0,1,....K-1 \qquad \text{----------} \qquad (3)$$

After this, a discrete cosine transform (DCT) is applied to log of mel spectral coefficients. Thus, the Mel frequency cepstral coefficients for frame m can be expressed as

$$c_m(q) = \sum_{k=0}^{K-1} \log(G(k)).\cos\left[\frac{\pi q(2k+1)}{2K}\right] \quad ; \quad q = 0,1,.....Q-1 \qquad \text{----------} \qquad (4)$$

Where $0 \le q \le Q-1$ and Q=12 is the desired number of cepstral features.

The segmented speech signal's energy is also considered as one of the features in this work, which is computed as

$$e = \sum_{n=0}^{N-1} s^2(n) \qquad \text{----------} \qquad (5)$$

In order to better reflect the dynamic changes of the MFCC in time, usually the first and second derivatives in time are also computed, i.e. by computing the difference of two coefficients lying τ times indices in the past and in the future of the time index. The first derivative is computed as:

$$\Delta c_m(q) = c_{m+\tau}(q) - c_{m-\tau}(q) \quad ; \quad q = 0,1,....Q-1 \qquad \text{----------} \qquad (6)$$

The second derivative is computed from the difference of the first derivatives:

$$\Delta\Delta c_m(q) = \Delta c_{m+\tau}(q) - \Delta c_{m-\tau}(q) \quad ; \quad q = 0,1,....Q-1 \qquad \text{----------} \qquad (7)$$

The time interval τ is taken as 4.

3.3. Visual features extraction

Visual features proposed in the literature of AVSR can be categorized into shape-based, pixel-based and motion-based features [29]. Pixel-based and shape based features are extracted from static frames and hence viewed as static features. Motion-based features are features that directly utilize the dynamics of speech [11, 12]. Dynamic features are better in representing distinct facial movements and static features are better in representing oral cavity that cannot be captured either by lip contour or motion-based features. This work focuses on the relative benefits of both static and dynamic features for improved AVSR recognition.

3.3.1. DCT based static feature extraction

G. Potamianos et al. [13] reported that intensity based features using discrete cosine transform (DCT) outperform model-based features. Hence DCT is employed in this work to represent static features. Each side-face mouth region video is recorded with a frame rate of 30 frames/sec and [240 x 320] pixel resolutions. Prior to the image transform the recorded video frames $\{V_t(a, b, c); 1 \le t \le 60; 1 \le a \le 240; 1 \le b \le 320; 1 \le c \le 3\}$ are converted to equivalent RGB image. This RGB image is converted to the YUV color space and only the luminance part (Y) of the image is kept as such since it retains the image data least affected by the video compression [14]. The resultant Y- image was sub sampled to [16 x 16] and then passed as the input $\{A_t(m, n); 1 \le t \le 60; 1 \le m \le 16; 1 \le n \le 16\}$ to the DCT. The images of [16 x 16] pixels provided slightly better performance than [32 x 32] pixel images [14], and hence in this work [16 x 16] pixel images are taken as input to the DCT.

The DCT has the property that, most of the visually significant information about the image is concentrated in just a few coefficients of the DCT. The two dimensional DCT of an m-by-n image sequence $\{A_t(m, n); 1 \le t \le 60; 1 \le m \le 16; 1 \le n \le 16\}$ is defined as:

$$B_t(p,q) = \left\{ \frac{1}{\sqrt{2N}} C(p)\, C(q) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_t(m,n) \cos\left[\frac{(2m+1)p\pi}{2M} \right] \cos\left[\frac{(2n+1)q\pi}{2N} \right] \right\} \quad \text{---------} \quad (8)$$

Where, $M = N = 16$; $0 \le p \le M-1$; $0 \le q \le N-1$; and

$$C(x) = \begin{cases} 1/\sqrt{2} & if \quad x = 0 \\ 1 & if \quad x > 0 \end{cases}$$

The DCT returns a 2D matrix $B_t(p, q)$ of coefficients and moreover, the triangle region feature selection outperforms the square region feature selection, as those include more of the coefficients corresponding to low frequencies [14]. Hence in this work, [6 x 6] triangle region DCT coefficients without the DC component are considered as 20 static features of a frame.

3.3.2. Motion segmentation based dynamic feature extraction

In this work, dynamic visual speech features which show the side-face mouth region movements of the speaker are segmented from the video using an approach called motion history images (MHI) [11]. MHI is a gray scale image that shows where and when movements of speech articulators occur in the image sequence.

Let $\{A_t(m, n); 1 \le t \le 60; 1 \le m \le 16; 1 \le n \le 16\}$ be a luminance part (Y) image sequence, the difference of frames is defined as

$$DIF_t(m,n) = \left| A_t(m,n) - A_{t-1}(m,n) \right| \quad \text{----------} \quad (9)$$

Where $A_t(m, n)$ is the intensity of each pixel at location (*m, n*) in the t^{th} frame and $DIF_t(m, n)$ is the difference of consecutive frames representing region of motion. Binarization of the difference image $DIF_t(m, n)$ over a threshold τ is

$$DOF_t(m,n) = \begin{cases} 1 & if \quad DIF_t(m,n) \ge \tau \\ 0 & otherwise \end{cases} \quad \text{----------} \quad (10)$$

The value of the threshold τ is optimized through experimentation. Finally MHI (m, n) is defined as

$$MHI(m,n) = Max \bigcup_{t=1}^{N-1} DOF_t(m,n) \times t \quad \text{----------} \quad (11)$$

Where N represents the number of frames used to capture the side-face mouth region motion. In equation (11), to show the recent movements with brighter value, the binarized version of the *DOF* is multiplied with a ramp of time and integrated temporally [11]. Next, DCT was applied to MHI (m, n) and the transformed coefficients are obtained. Similar to static feature extraction, only [6 x 6] triangle region DCT coefficients without DC component are considered as the dynamic features. Finally, the static and dynamic features are concatenated to represent visual speech.

4. DECISION FUSION WITH GA OPTIMIZED RELIABILITY RATIO-BASED INTEGRATION

The main focus of this work is the estimation of optimal integration weight for the modalities in the decision fusion. After the acoustic and visual subsystems perform recognition separately, their outputs are combined by a weighted sum rule to produce the final decision. For a given audio-visual speech test datum of O_A and O_V the recognized utterance C^* is given by [5],

$$C^* = \arg\max_i \left\{ \gamma \log P(O_A / \lambda_A^i) + (1-\gamma) \log P(O_V / \lambda_V^i) \right\} \quad \text{----------} \quad (12)$$

Where λ_A^i and λ_V^i are the acoustic and the visual HMMs for the i^{th} utterance class, respectively, and $\log P(O_A / \lambda_A^i)$ & $\log P(O_V / \lambda_V^i)$ are their outputs. The weighting factor γ ($0 \le \gamma \le 1$) determines how much each modality contributes to the final decision. If it is not estimated appropriately we cannot expect complementarity and synergy of the two information sources and moreover, the combined recognition performance may be even inferior to that of any uni-modal systems [25].

One simple solution to this problem is assigning a constant weight value over various SNR conditions or manual determination of the weight [30]. In some other work, the weight is determined from SNR by assuming that SNR of the acoustic signal is known which is not always a feasible assumption [4]. Indeed, some researchers determine the weight by using an additional adaptation data [31]. Finally, the most popular approach among such schemes is the reliability ratio (RR)-based method in which the integration weight is calculated from the relative reliability measures of the two modalities [32]. This work proposes a Genetic Algorithm based optimization scheme to determine appropriate integration weight from the relative reliability measures of the two modalities, which ensures complementarity and synergy of AVSR without a priori knowledge of the SNR or additional adaptation data. The following subsections briefly explain the baseline reliability ratio - based integration method [32] and the proposed GA optimized reliability ratio - based integration procedure to determine the appropriate integration weight from the reliability measures of acoustic and visual classifiers.

4.1. Baseline reliability ratio - based integration

The reliability of each modality can be measured from the outputs of the corresponding HMMs. When the acoustic speech is not corrupted by any noise, there are large differences between the acoustic HMMs output or else the differences become small. Considering this observation, the reliability of a modality is defined by the most appropriate and best in performance [2]

$$S_m = \frac{1}{N_c - 1} \sum_{i=1}^{N} (\max_j \log P(O / \lambda^j) - \log P(O / \lambda^i)) \quad \text{----------} \quad (13)$$

Which means the average differences between the maximum log-likelihood and the other ones and N_c is the number of classes being considered to measure the reliability of each modality m $\in$ {A, V}. In this case, N_c is 70 i.e. all class recognition hypotheses are considered to measure the reliability. Then, the integration weight γ can be calculated by [32]

$$\gamma = \frac{S_A}{S_A + S_V} \quad \text{----------} \quad (14)$$

Where S_A and S_V are the reliability measure of the outputs of the acoustic and visual subsystems, respectively.

4.2. Proposed GA optimized reliability ratio - based integration

The audio-visual integration method proposed in sections 4.1 can improve the recognition accuracy as compared to the audio-only over certain SNR conditions. This may not be the optimal method of integration weight estimation since that did not show performance improvement at all SNRs. This was experimentally proved in this work for the noisy speech data. To overcome this problem and ensure performance improvement at all SNR conditions, this work proposes a method which optimizes the integration weight estimated in section 4.1 by using genetic algorithm.

The genetic algorithm is a method for solving both constrained and unconstrained optimization problems. It is built on the principles of evaluation via natural selection: an initial population of individual is created and by iterative application of the genetic operators (selection, crossover, mutation) an optimal solution is reached according to the defined fitness function. The procedure of the proposed algorithm is as follows:

Step 1: Initialization: Generate a random initial population of size 20.

Step 2: Fitness evaluation: Fitness of all the solutions in the populations is evaluated. The steps for evaluating the fitness of a solution are given below:

Step 2a: Assume the matrix P of size [$N_c \times N_c$] with all zero values. Where N_c is the Number of utterance class.

Step 2b: class = 1 : No of class (N_c = 70).

Step 2c: test datum = 1 : No of test datum (*Nts* = 5).

Step 2d: Get the acoustic and visual subsystems log likelihood $\log P(O_A / \lambda_A^i)$ and $\log P(O_V / \lambda_V^i)$; respectively, for the class and test datum given in steps 2*b* & 2*c*.

Step 2e: Find the maximum value of acoustic log likelihood
i.e., *amax* = *max* (*sort* (log P (O_A/λ_A^i))*, decend*) for the class and test datum given in steps 2*b* & 2*c*.

Step 2f: Find the maximum value of visual log likelihood
i.e., *vmax* = *max* (*sort* (log *P* (O_V/λ_V^i))*, decend*) for the class and test datum given in steps 2*b* & 2*c*.

Step 2g: Compute the acoustic reliability S_A as:

$$S_A = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} (a\max - \log P(O_A / \lambda_A{}^i))$$

Where N_c is the number of classes being considered.

Step 2h: Compute the visual reliability S_V as:

$$S_V = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} (v\max - \log P(O_V / \lambda_V{}^i))$$

Step 2i: Estimate the integration weight γ as:

$$\gamma = x \times \left[\frac{S_A}{S_A + S_V} \right]$$

According to the solution *x*.

Step 2j: Integrate the log likelihoods as follows

$$C = \arg\max_{i} \left\{ \gamma \log P(O_A / \lambda_A^i) + (1-\gamma) \log P(O_V / \lambda_V^i \right\}$$

Using the estimated integration weight value γ.

Step 2k: Find the maximum value and its corresponding index in *C*.

Step 2l: Increment the value of matrix *P* according to the class and index of *C* as follows

P (class, index) = P (class, index) + 1

Step 2m: Go to step 2*c* until all the test datum are over.

Step 2n: Go to step 2*b* until all the classes are over.

Step 2o: The recognition accuracy or fitness value is defined as

$$\text{Recognition Accuracy} = \frac{\sum diag(P)}{\sum\sum(P)} \times 100$$

Step 3: Updating Population: Two best solutions in the current population are forwarded to the next generation parents without any changes, the remaining solutions in the new population are generated using crossover and mutation.

Step 4: Termination: Repeat steps 2 to 3 until the algorithm reaches the maximum number of iterations.

The final best fitness value gives the maximum recognition accuracy and its corresponding solution gives best integration weight multiplier to obtain the appropriate integration weight for decision fusion.

5. HMM TRAINING AND RECOGNITION RESULTS

The bimodal decision fusion speech recognition system using side-face mouth region image is shown in figure 3. Both speech and side-face mouth region images are simultaneously recorded using low cost microphone and web camera. Audio signals are sampled at 8 kHz with 16-bit resolution. A single frame contains 32 milliseconds speech samples and the frame window proceeds by 12.5 milliseconds. The 12th-order MFCCs, the normalized energy and their delta terms are used for the acoustic features. Further, the cepstral mean subtraction (CMS) technique was applied to remove the channel distortion contained in the speech samples. Visual signals focusing side-face mouth region images are recorded with a frame rate of 30 frames/sec and [240 x 320] pixel resolutions. This work involves decision fusion and hence there is no frame rate synchronization problem between the acoustic and visual speech features. The static visual speech features are computed via DCT image transform approach and dynamic visual speech features are computed via MHI approach. The dynamic features are computed for the whole word not for individual frames. Finally, the static and dynamic features are concatenated to represent visual speech.

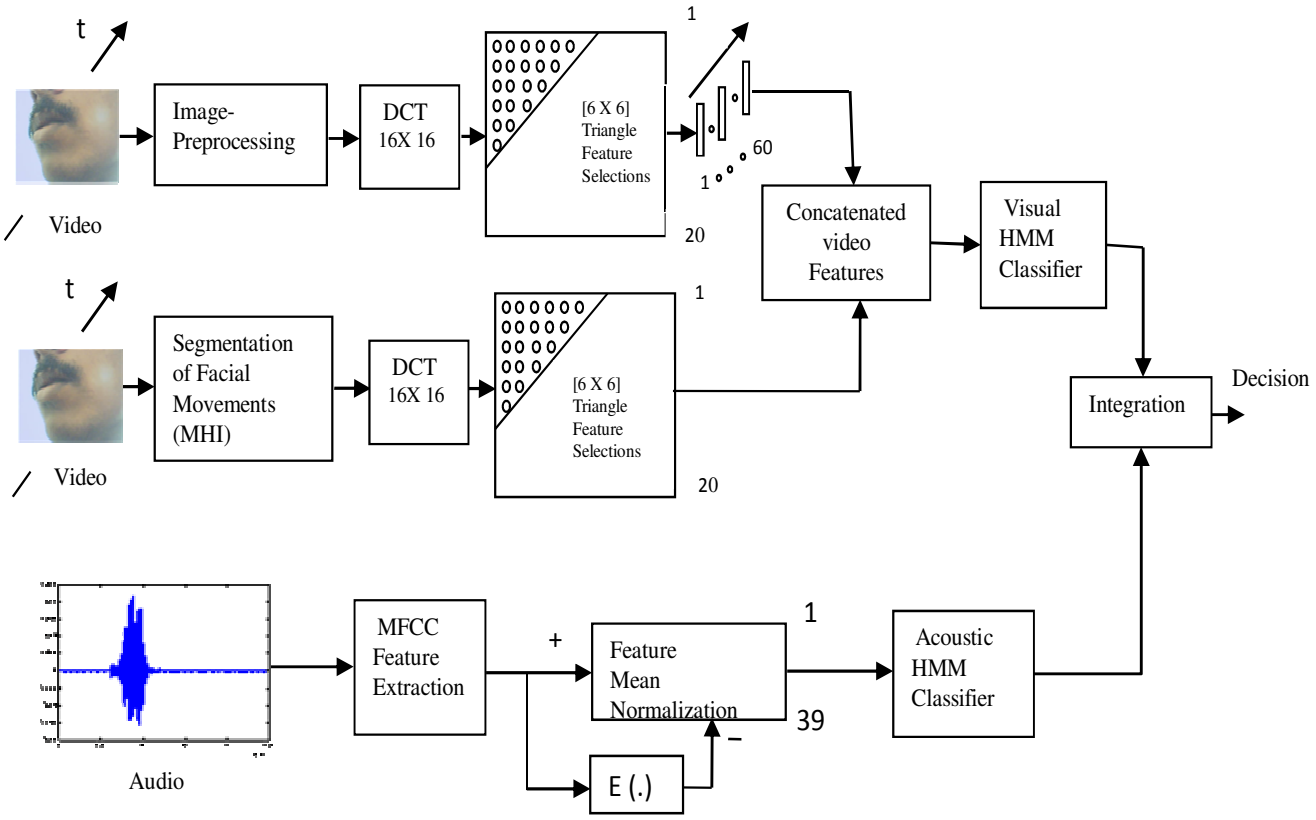

FIGURE 3: Audio-Visual decision fusion speech recognition system using mouth region side-face images

5.1. HMM Recognizer

HMM is a finite state network based on stochastic process. The left-right HMM is a commonly used classifier in speech recognition, since it has the desirable property that it can readily model the time-varying speech signal [16]. This work also adopts left-right continuous HMMs having Gaussian mixture models (GMMs) in each state. The whole- word model which is a standard approach for small vocabulary speech recognition task was used. The number of states in each HMM and number of Gaussian functions in each GMM are set to 10 and 6 respectively, which are determined experimentally. The initial parameters of the HMMs are obtained by uniform segmentation of the training data onto the states of the HMMs and iterative application of the segmental k-means algorithm and the Viterbi alignment. For training the HMMs, the standard Baum-Welch algorithm was used [16]. The training was terminated when the relative change of the log-likelihood value is less than 0.001 or maximum number of iteration is reached, which is set to 25.

SNR	Audio only (%)	Visual only (%)	AV Baseline-RR (%)	AV GA Optimized-RR (%)	Optimum Weight γ
20 dB	94.86	54	98	98.29	0.91
10 dB	50	54	68.57	78	0.25
5 dB	13.71	54	32	66.57	0.07
0 dB	2.86	54	14.57	58	0.04
-5 dB	1.43	54	10	56	0.02
-10 dB	1.43	54	8	54.86	0.01
Average (%)					
(-10dB ~20 dB)	27.38	54	38.52	68.82	
(-10dB ~ 5 dB)	4.86	54	16.14	58.86	

TABLE 1: Audio - only, visual-only, audio-visual speech recognition accuracies

5.2. Results

The proposed GA optimized reliability ratio-based integration algorithm has been tested on single speaker, seventy mobile functions isolated word. The dataset was recorded in an office environment with background noise. Each word was recorded 25 times, 80% of which have been used for training and 20% for testing. The recorded noisy acoustic signal is again artificially degraded with additive white Gaussian noise at SNRs of 20, 10, 5, 0, -5, and -10dB. As mentioned earlier, the main focus of this work is estimating the optimal integration weight for the modalities and in turn maximizing the synergy effect.

Table 1 shows recognition accuracies obtained by the audio-only, visual-only, audio-visual baseline reliability ratio, and the proposed bimodal system at various SNR conditions. Similarly figure 4 compares the recognition performance of all the systems. From the results, the following observations were made,

1. The audio-only recognition system shows nearly 95% for the recorded real time noisy speech at 20dB SNR but, as the speech becomes more noisy, its performance is degraded sharply; the recognition accuracy is even less than 2% at -5 and -10dB SNR conditions.

2. The visual-only system shows 54%, recognition accuracy at all SNRs.

3. The baseline reliability ratio-based method shows synergy effect only at 20 and 10dB SNR conditions but, in the remaining SNR conditions (i.e., -10dB 5dB) their performances are inferior to that of visual-only system i.e. they show attenuation fusion at these SNR conditions.

4. But, the proposed GA optimized reliability ratio-based bimodal system shows synergy effect at all SNR conditions. The amount of synergy at all SNRs is plotted in figure 5. The maximum synergy of 24% occurs at 10dB SNR.

5. Compared to the acoustic-only system, relative recognition performance by the proposed bimodal system is 41.44% on average at all SNR conditions. Under high-noise conditions (i.e., -10dB 5dB), relative recognition performance is 54%.

6. Similarly, compared to the baseline reliability ratio-based system, relative recognition performance by the proposed bimodal system is 30.3% on average at all SNR conditions. Under high-noise conditions (i.e., -10dB 5dB), relative recognition performance is 42.72%, which demonstrates that the noise robustness of recognition is achieved by the proposed system.

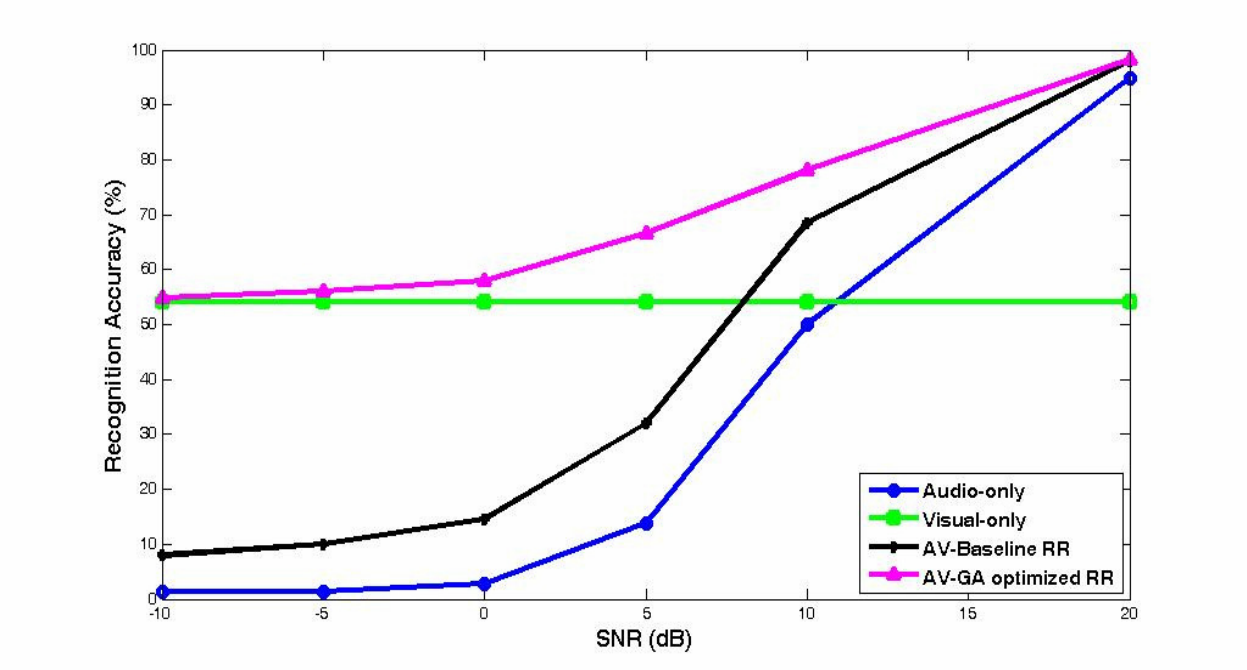

FIGURE 4: Recognition performance of the uni-modal and bimodal systems

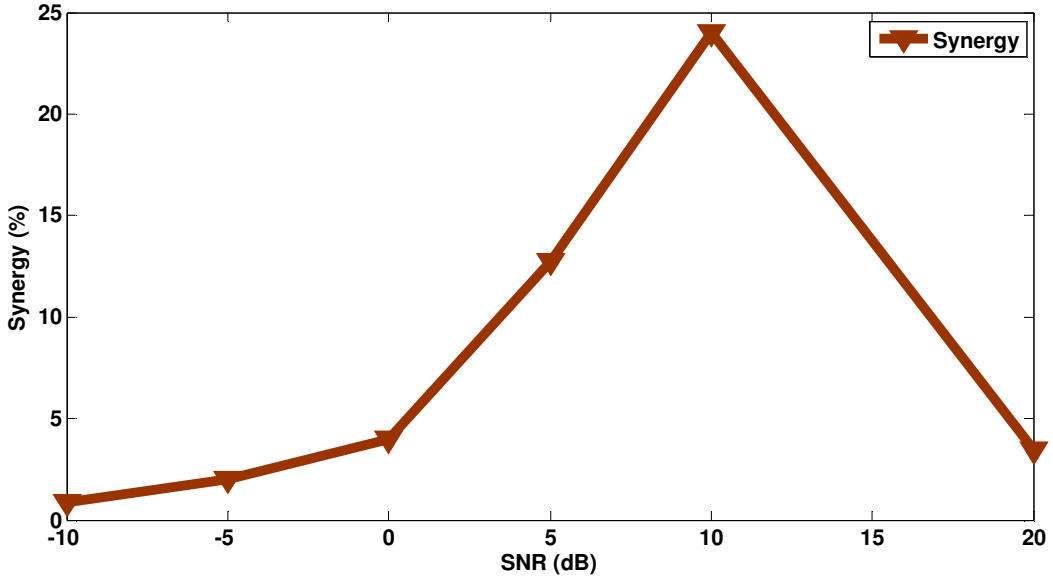

FIGURE 5: Synergy effect of proposed GA optimized RR-based system on various SNR

6. DISCUSSION AND CONCLUSION

In this paper, a GA optimized reliability ratio-based integration weight estimation scheme for decision fusion AVSR system is proposed. The proposed system uses an audio-visual speech data base developed by us, which extracts visual features from the side-face mouth region images rather than frontal face images to focus on mobile applications. Generally, the dynamic visual speech features are obtained by derivatives of static features [14], but in this work the dynamic features are obtained via MHI approach and concatenated with static features to represent the visual speech. For evaluating the proposed method, the recognition accuracy is compared with the related method called baseline reliability ratio-based method in section 5.2. Results show that the proposed method significantly improves the recognition accuracy at all SNR conditions as compared to the baseline reliability ratio-based method. At low SNR, baseline reliability ratio-based method shows very poor recognition accuracy. But the proposed method solves this issue and improves the recognition accuracy considerably. Our future work needs to address the following issues:

1. The baseline reliability ratio-based system show "attenuating fusion" on high-noise conditions (i.e., -10dB 5dB). Therefore an effective denoising algorithm is to be developed to improve the performance further.

2. Moreover, this work was done on a single speaker, small vocabulary mobile function isolated words recognition task. In practice to cover all the recent mobile applications this work needs to be extended to multi speaker, medium size vocabulary, and continuous word recognition task.

7. REFERENCES

1. K. Iwano, T. Yoshinaga, S. Tamura, S. Furui. "*Audio-visual speech recognition using lip information extracted from side-face images*". EURASIP Journal on Audio, Speech, and Music Processing, (2007): 9 pages, Article ID 64506, 2007

2. J.S. Lee, C. H. Park. "*Adaptive Decision Fusion for Audio-Visual Speech Recognition*"'. In: F. Mihelic, J. Zibert (Eds.), Speech Recognition, Technologies and Applications, pp. 550 (2008)

3. J.S. Lee, C. H. Park. "*Robust audio-visual speech recognition based on late integration*"'. IEEE Transaction on Multimedia, 10: 767-779, 2008

4. G. F. Meyer, J. B.Mulligan, S. M.Wuerger. "*Continuous audiovisual digit recognition using N-best decision fusion*". Information Fusion. 5: 91-101, 2004

5. A. Rogozan, P. Delglise. "*Adaptive fusion of acoustic and visual sources for automatic speech recognition*". Speech Communication. 26: 149-161, 1998

6. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. "*Recent advances in the automatic recognition of audio-visual speech*". In Proceedings of IEEE, 91(9), 2003

7. S. Dupont, J. Luettin. "*Audio-visual speech modeling for continuous speech recognition*". IEEE Transanction on Multimedia, 2: 141-151, 2000

8. G. Potamianos, H. P. Graf, and E. Cosatto. "*An image transform approach for HMM based automatic lipreading*". In Proceedings of International Conference on Image Processing. Chicago, 1998

9. R. Rajavel, P. S. Sathidevi. "*Static and dynamic features for improved HMM based visual speech recognition*". In Proceedings of 1st International Conference on Intelligent Human Computer Interaction, Allahabad, India, 2009

10. G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu. "*A cascade image transform for speaker independent automatic speechreading*". In Proceedings of IEEE International Conference on Multimedia and Expo. New York, 2000

11. W. C. Yau, D. K. Kumar, S. P. Arjunan. "*Voiceless speech recognition using dynamic visual speech features*". In Proceedings of HCSNet Workshop on the Use of Vision in HCI. Canberra, Australia, 2006

12. W. C. Yau, D. K. Kumar, H. Weghorn. "*Visual speech recognition using motion features and Hidden Markov models*". In: M. Kampel, A. Hanbury (Eds.), LNCS, Springer, Heidelberg, pp. 832-839 (2007)

13. G. Potamianos, C. Neti, J. Luettin, and I. Matthews. "*Audio-visual automatic speech recognition: An overview*". In: G. Baily, E. Vatikiotis-Bateson, P. Perrier (Eds.), Issues in visual and audio-visual speech processing, MIT Press, (2004)

14. R. Seymour, D. Stewart, J. Ming. "*Comparison of image transformbased features for visual speech recognition in clean and corrupted videos*". EURASIP Journal on Image and Video Processing. (2008), doi:10.1155/2008/810362, 2008

15. B. Plannerer. "*An introduction to speech recognition: A tutorial* ". Germany, 2003

16. L. Rabiner, B.H. Juang. "*Fundamentals of Speech Recognition*"'. Prentice Hall, Englewood Cliffs (1993)

17. B. Nasersharif, A. Akbari. "*SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features*". Pattern Recognition Letters, 28:1320-1326, 2007

18. T. Chen. "*Audiovisual speech processing. Lip reading and lip synchronization*". IEEE Signal Processing Magazine, 18: 9-21, 2001

19. E. D. Petajan. "*Automatic lipreading to enhance speech recognition*". In Proceedings of Global Telecommunications Conference. Atlanta, 1984

20. P. Arnold, F. Hill. "*Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact*". Brit. J. Psychol., 92: 339-355, 2001

21. A. Q. Summerfield. "*Some preliminaries to a comprehensive account of audio-visual speech perception*". In: B. Dodd, R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-reading. Lawrence Erlbarum, London, pp. 3-51 (1987)

22. C. Benoit, T. Mohamadi, S. D. Kandel. "*Effects of phonetic context on audio-visual intelligibility of French*". Journal of Speech and Hearing Research. 37: 1195-1203, 1994

23. C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. "*Audio visual speech recognition, Final Workshop 2000 Report"*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2000

24. P. Teissier, J. Robert-Ribes, J. L. Schwartz. "*Comparing models for audiovisual fusion in a noisy-vowel recognition task*". IEEE Transaction on Speech Audio Processing, 7: 629-642, 1999

25. C. C. Chibelushi, F. Deravi, J. S. D. Mason. "*A review of speech-based bimodal recognition*". IEEE Transactions on Multimedia, 4(1): 23-37, 2002

26. P.L. Silsbee. "*Sensory integration in audiovisual automatic speech recognition*". In Proceedings of the 28th Annual Asilomar Conference on Signals, Systems, and Computers, 1: 561-565, 1994

27. C. Benot. "*The intrinsic bimodality of speech communication and the synthesis of talking faces*". In: M. M. Taylor, F. Nel, D. Bouwhuis (Eds.), The Structure of Multimodal Dialogue II. Amsterdam, Netherlands, pp. 485-502 (2000)

28. G. Potamianos, C. Neti, J. Huang, J.H. Connell, S. Chu, V. Libal, E.Marcheret, N. Hass, J. Jiang. "*Towards practical development of audiovisual speech recognition*". In Proceedings of IEEE International Conf. on Acoustic, Speech, and Signal Processing. Canada, 2004

29. S.W.Foo, L. Dong. "*Recognition of Visual Speech Elements Using Hidden Markov Models*". In: Y. C. Chen, L.W. Chang, C.T. Hsu (Eds.), Advances in Multimedia Information Processing-PCM02, LNCS2532. Springer-Verlag Berlin Heidelberg, pp.607-614 (2002)

30. A. Verma, T. Faruquie, C. Neti, S. Basu. "*Late integration in audiovisual continuous speech recognition*". In Proceedings of Workshop on Automatic Speech Recognition and Understanding. Keystone, 1999

31. S. Tamura, K. Iwano, S. Furui. "*A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization*". In Proceedings of ICASSP. Philadelphia, 2005

32. A. Adjoudani, C. Benot. "*On the integration of auditory and visual parameters in an HMM-based ASR*". In: D. G. Stork and M. E. Hennecke (Eds.), Speech reading by Humans and Machines: Models, Systems, and Speech Recognition, Technologies and Applications, Springer, Berlin, Germany, pp. 461-472 (1996)