J Sign Process Syst (2012) 68:83–93
DOI 10.1007/s11265-011-0578-x

Adaptive Reliability Measure and Optimum Integration Weight for Decision Fusion Audio-visual Speech Recognition

R. Rajavel · P. S. Sathidevi

Received: 9 October 2009 / Revised: 12 January 2011 / Accepted: 14 January 2011 / Published online: 2 February 2011

Abstract Audio-visual speech recognition (AVSR) using acoustic and visual signals of speech has received attention recently because of its robustness in noisy environments. An important issue in decision fusion based AVSR system is the determination of appropriate integration weight for the speech modalities to integrate and ensure better performance under various SNR conditions. Generally, the integration weight is calculated from the relative reliability of two modalities. This paper investigates the effect of reliability measure on integration weight estimation and proposes a genetic algorithm (GA) based reliability measure which uses optimum number of best recognition hypotheses rather than N best recognition hypotheses to determine an appropriate integration weight. Further improvement in recognition accuracy is achieved by optimizing the above measured integration weight by genetic algorithm. The performance of the proposed integration weight estimation scheme is demonstrated for isolated word recognition (incorporating commonly used functions in mobile phones) via multi-speaker database experiment. The results show that the proposed schemes improve robust recognition accuracy over the conventional unimodal systems, and a couple of related existing bimodal systems, namely, the baseline reliability ratio-based system and N best recognition hypotheses reliability ratio-based system under various SNR conditions.

R. Rajavel (✉) · P. S. Sathidevi
ECE Department, National Institute of Technology Calicut, Calicut 673601, India
e-mail: rettyraja@gmail.com

P. S. Sathidevi
e-mail: sathi@nitc.ac.in

Keywords Audio-visual speech recognition · Side face visual feature extraction · Audio-visual decision fusion · Reliability-ratio based weight optimization · GA based reliability measure

1 Introduction

Human's speech perception is bimodal in nature: human combine audio and visual information in deciding what the others speak. The first AVSR system was reported in 1984 by Petajan [18]. During the last decade more than hundred articles have appeared on AVSR [5, 6, 8, 9, 13, 17, 23, 25]. AVSR systems can enhance the performance of the conventional ASR not only under noisy conditions but also in clean conditions when the talking face is visible [20, 26]. The major advantage of utilizing the acoustic and the visual modalities for speech understanding comes from "Complementarity" [21] of the two modalities and, "Synergy": Performance of audio-visual speech perception can outperform those of acoustic-only and visual-only perception for diverse noise conditions [22]. Generally, in AVSR systems, 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, 6, 7, 13], and others prefer late integration [2–5, 19, 24]. Despite 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.

Commonly, the integration weight which determines the amount of contribution from each modality in decision fusion based AVSR system is calculated from the relative reliability of the two modalities [31]. The method of reliability measure proposed in [3, 32] use all the classes of recognition hypotheses, where as the method proposed in [5, 31] uses only N (i.e $N = 4$) best recognition hypotheses. But, both of these methods did not show performance improvements practically at very low SNR conditions. To solve this issue, this work proposes a genetic algorithm based reliability measure which uses optimum number of best recognition hypotheses rather than N best recognition hypotheses to determine an appropriate integration weight. Further improvement in recognition accuracy is achieved by optimizing the above measured integration weight by genetic algorithm. The performance of the proposed integration weight estimation scheme using GA based reliability measure is demonstrated for isolated word recognition (incorporating commonly used functions in mobile phones) via multi-speaker database experiment. After the recognition tasks were carried out over the common audio-visual side face speech database, the performance of proposed system is compared with the audio-only, visual-only unimodal systems and some existing bimodal AVSR systems namely, the baseline reliability ratio-based system and N best recognition hypotheses reliability ratio-based system under various SNR conditions. An outline of the remainder of the paper is as follows. The following section explains some of the existing methods to find the integration weight based on reliability measure of the modalities. How Genetic Algorithm can be used to measure the correct reliabilities of each modality and optimize the integration weight is explained in Section 3. Section 4 discusses the database, audio, and visual features. Section 5 discusses the Hidden Markov Model (HMM) training and recognition results. The discussion, conclusion and future direction of this work are outlined in the last section.

2 Review of Existing Integration Weight Estimation Schemes

The main focus of this work is on the estimation of appropriate integration weight based on the correct reliability measure of audio and visual modalities. 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],

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

where λ_A^i and λ_V^i are the acoustic and the visual HMMs for the ith ($1 \leq i \leq N$) utterance class, respectively, N is the number of utterance classes being used in the recognition experiment, and $\log P(O_A/\lambda_A^i)$ and $\log P(O_V/\lambda_V^i)$ are their log likelihood against the ith class. The weighting factor $\gamma (0 \leq \gamma \leq 1)$ determines the contribution of each modality to the final decision. If it is not estimated appropriately we cannot expect complementarity [21] and synergy [22] of the two information sources and moreover, the combined recognition performance may be even inferior to that of any unimodal systems, which is called "attenuating fusion" [25]. One simple solution to this problem is assigning a constant weight value over various SNR conditions or manual determination of the weight [29]. 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 additional adaptation data [30]. Finally, the most popular approach among such schemes is the reliability ratio(RR) based method in which the integration weight is determined from the relative reliability of the two modalities [31]. Hence, in this section we briefly, review this baseline reliability ratio(RR)-based integration method and another related method called N-best recognition hypotheses reliability ratio-based integration method [5, 31].

2.1 Audio-Visual Decision Fusion Based on Baseline Reliability Ratio Method

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]

$$S_m = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} \left(\max_j \log P\left(O/\lambda_m^j\right) - \log P\left(O/\lambda_m^i\right)\right) \tag{2}$$

which means the average difference between the maximum log-likelihood and the other ones and N_c is the number of utterance classes being considered to measure the reliability of each modality $m \in \{A, V\}$. In this method, all the utterance class recognition hypotheses are used to measure the reliability. Then, the integration weight γ can be calculated by [31]

$$\gamma = \frac{S_A}{S_A + S_V} \tag{3}$$

where S_A and S_V are the reliability measure of the outputs of the acoustic and visual HMMs, respectively.

2.2 Audio-visual Decision Fusion Based on N-best Recognition Hypotheses Reliability Ratio Method

Adjoudani and Benoit [31], measured the reliability of each modality $m \in \{A, V\}$ over N best recognition hypotheses allowing for satisfactory evaluation of certainty versus uncertainty in conformity [5]. Accordingly, the reliability of a modality is defined as

$$S_m = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| \log P\left(O/\lambda_m^i\right) - \log P\left(O/\lambda_m^j\right) \right| \tag{4}$$

which means the average absolute differences of log-likelihood. In this method, only four best recognition hypotheses are used to measure the reliability of each modality and then, the integration weight γ is calculated as in Eq. 3.

3 Audio-Visual Decision Fusion Based on Proposed Integration Weight Estimation Schemes

In this section, we explain our novel integration weight estimation scheme which uses optimum best recognition hypotheses to measure the correct reliability of each modality and in turn appropriate integration weight. In the next subsection, we present a genetic algorithm based optimization scheme to further optimize the integration weight from the above measured reliabilities.

3.1 Audio-visual Decision Fusion Based on GA Adaptive Reliability Measure Method (Proposal 1)

The method of reliability measure proposed in [3, 32] use all the classes of recognition hypotheses, where as the method proposed in [31] uses only N (i.e. $N = 4$) best recognition hypotheses. The estimated integration weight based on these measures shows "attenuating fusion" [25] for noisy speech data on certain SNR conditions. To solve this issue, this work proposes a GA based scheme to select optimum number of best recognition hypotheses to measure the correct reliability of each modality so as to, increase the recognition accuracy at all SNR conditions.

The genetic algorithm is a method for solving both constrained and unconstrained optimization problems. It is built on the principles of evolution via natural selection: an initial population of individuals 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 GA is used in this work to obtain the correct reliabilities of each modality and in turn maximize the recognition accuracy according to the defined fitness function. The problem is formulated as follows:

The optimum number of acoustic recognition hypotheses to measure the correct reliability (S_A) is obtained by solving

$$S_A = \arg\max_{N_A} \left\{ \frac{1}{(N_A - 1)} \sum_{i=1}^{N_A} \max_j \log P\left(O_A/\lambda_A^j\right) - \log P\left(O_A/\lambda_A^i\right) \right\} \tag{5}$$

similarly, the correct visual reliability (S_V) is obtained by solving

$$S_V = \arg\max_{N_V} \left\{ \frac{1}{(N_V - 1)} \sum_{i=1}^{N_V} \max_j \log P\left(O_V/\lambda_V^j\right) - \log P\left(O_V/\lambda_V^i\right) \right\} \tag{6}$$

subject to: $1 \leq N_A, N_V \leq N$.

Then, the integration weight (γ) is calculated as in Eq. 3. Finally, the fitness function to be optimized is given as

$$\text{Recognition Accuracy} = \frac{\sum diag(R)}{\sum\sum(R)} \times 100 \tag{7}$$

where R is the confusion matrix. The proposed Algorithm 1 based on GA to solve Eqs. 5, 6 and 7 is explained step-by-step in the following procedure

Step 1 Initialization: Generate a random initial population of size $[N \times 2]$, for best acoustic and visual recognition hypotheses length to be considered to measure the correct reliability.

Step 2 Fitness Evaluation: Fitness of all the solutions $\{N_{A1}, N_{A2}....N_{AN}\}$ and $\{N_{V1}, N_{V2}....N_{VN}\}$ in

the population is evaluated. The steps for evaluating the fitness of a solution are given below:

Step 2a: Assume the confusion matrix R of size $[N_c \times N_c]$ with all zero values.

Step 2b: Class = 1: No of validation utterance class.

Step 2c: Datum = 1: No of validation utterance datum.

Step 2d: Get the acoustic log likelihood $\log P(O_A/\lambda_A^i)$; $(1 \le i \le N_c)$ for the Class and Datum. Each of its entries represents the log likelihood of the Datum O_A against all acoustic classes.

Step 2e: Find the maximum value in the acoustic log likelihoods and is represented as *amax*.

Step 2f: Compute the acoustic reliability S_A as:

$$S_A = \frac{1}{N_A - 1} \sum_{i=1}^{N_A} \left(amax - \log P\left(O_A/\lambda_A^i\right)\right) \tag{8}$$

where $N_A \in \{N_{A1}, N_{A2}....N_{AN}\}$ is the number of acoustic recognition hypotheses being considered to measure the correct acoustic reliability.

Step 2g: Similarly get the visual subsystem log likelihood $\log P(O_V/\lambda_V^i)$; $(1 \le i \le N_c)$ for the Class and Datum. Each of its entries represents the log likelihood of the Datum O_V against all visual classes.

Step 2h: Find the maximum value in the visual log likelihoods and is represented as *vmax*.

Step 2i: Compute the visual reliability S_V as:

$$S_V = \frac{1}{N_V - 1} \sum_{i=1}^{N_V} \left(vmax - \log P\left(O_V/\lambda_V^i\right)\right) \tag{9}$$

where $N_V \in \{N_{V1}, N_{V2} .... N_{VN}\}$ is the number of visual recognition hypotheses being considered to measure the correct visual reliability.

Step 2j: Estimate the integration weight γ as:

$$\gamma = \frac{S_A}{(S_A + S_V)} \tag{10}$$

Step 2k: Integrate the log likelihoods as follows:

$$C1 = \left\{\gamma \log P\left(O_A/\lambda_A^i\right) + (1-\gamma) \log P\left(O_V/\lambda_V^i\right)\right\} \tag{11}$$

using the estimated integration weight value γ in step 2j. Now $C1$ is a $[N_c \times 1]$ matrix which gives the audio-visual combined recognition hypotheses.

Step 2l: Find the maximum value of $C1$ and its corresponding position. The position represents the recognized utterance class.

Step 2m: Update the confusion matrix R as follows

$$R(class, position) = R(class, position) + 1 \tag{12}$$

Step 2n: Go to step 2*c* until all the Datum are over.

Step 2o: Go to step 2*b* until all the Classes are over.

Step 2p: The recognition accuracy or fitness value is calculated as

$$\text{Recognition accuracy} = \frac{\sum diag(R)}{\sum\sum(R)} \times 100 \tag{13}$$

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

The scattered crossover function creates a random binary vector and selects the genes from the 1st parent if the vector is 1, and the genes from the 2nd parent if the vector is 0, and combines the genes to form the next generation parents [16]. Similarly, the Gaussian mutation function adds a random number taken from a Gaussian distribution with zero mean and user defined variance to each entry of the current

parents to form the next generation parents [16]. The combination of scattered crossover function and the Gaussian mutation function converges quickly to the given fitness function.

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

The final solution of this Algorithm 1 gives the number of best acoustic and visual subsystems recognition hypotheses to be considered to measure the correct reliability of each modality. The performance of the proposed [1] method over the baseline reliability ratio-based and N best recognition hypotheses reliability ratio-based methods are shown in Table 1

3.2 Audio-visual Decision Fusion Based on GA Adaptive Reliability Measure and Optimum Integration Weight Method (Proposal 2)

The GA adaptive reliability measure proposed in Section 3.1 improves the recognition accuracy over the baseline reliability ratio-based and N best recognition hypotheses reliability ratio-based methods and its performance comparison is shown in Table 1. But still there is attenuating fusion at very low SNR conditions for noisy speech data. To solve this issue, we propose a scheme to further optimize the integration weight computed in Section 3.1 and thereby improves the recognition accuracy without attenuating fusion at any SNR conditions. The problem is formulated as follows:

Define the new integration weight $\overline{\gamma}$ as

$$\overline{\gamma} = \left[\frac{S_A}{(S_A + S_V)}\right] \times x \tag{14}$$

i.e. $\overline{\gamma} = \gamma \times x$. Then, for the given test datum O_A and O_V the recognized utterance C^* is obtained by solving

$$C^* = \arg\max_{i,x} \left\{\overline{\gamma} \log P\left(O_A/\lambda_A^i\right) + (1-\overline{\gamma}) \log P\left(O_V/\lambda_V^i\right)\right\} \tag{15}$$

$$\text{subject to} : 0 \leq \overline{\gamma} \leq 1$$

Finally, the objective function given in Eq. 7 based on this new integration weight is optimized using genetic algorithm. The procedure of the proposed [2] algorithm for optimizing the objective function using GA is explained in the following procedure

Step 1 Initialization: Generate a random initial population of size $[N \times 3]$, for best acoustic and visual recognition hypotheses length to be considered to measure the correct reliability, and the integration weight multiplier (x).

Step 2 Fitness Evaluation: Fitness of all the solutions $\{N_{A1}, N_{A2}....N_{AN}\}$, $\{N_{V1}, N_{V2}....N_{VN}\}$, and $\{x_1, x_2....x_N\}$ in the population is evaluated. The steps for evaluating the fitness of a solution are given below:

Step 2a–i: Follow the same steps as in Section 3.1.

Step 2j: Estimate the new integration weight $\overline{\gamma}$ as:

$$\overline{\gamma} = x_i \times \left(\frac{S_A}{(S_A + S_V)}\right) \tag{16}$$

based on the integration weight multiplier solution x_i. where $x_i \in \{x_1, x_2....x_N\}$.

Table 1 Recognition performance comparison of AV baseline-RR, AV N best-RR, and AV GA adaptive-RR bimodal systems.

SNR	AV baseline-RR (%)	AV N best-RR (%)	AV GA adaptive-RR (%)
Clean	85.24	86.38	88.47
20 dB	78.76	80.57	83.43
10 dB	56.67	63.43	66.28
5 dB	30.28	60.38	65.52
0 dB	26.19	42.76	59.86
−5 dB	21.52	27.52	47.14
Average(%)			
(−5 dB ∼ Clean)	49.77	60.17	68.45
(−5 dB ∼ 10 dB)	33.66	48.52	59.70

Step 2k: Integrate the log likelihoods as follows:

$$C2 = \left\{\overline{\gamma}\log P\left(O_A/\lambda_A^i\right) + (1-\overline{\gamma})\log P\left(O_V/\lambda_V^i\right)\right\} \tag{17}$$

using the estimated integration weight value $\overline{\gamma}$ in step 2j. Now $C2$ is a $[N_c \times 1]$ matrix which gives the audio-visual combined recognition hypotheses.

Step 2l: Find the maximum value of $C2$ and its corresponding position. The position represents the recognized utterance class.

Step 2m: Update the confusion matrix R as follows:

$$R(class, position) = R(class, position) + 1 \tag{18}$$

Step 2n–p: Follow the same steps as in Section 3.1.

Step 3 Updating Population: As similar to step 3 of the algorithm in Section 3.1, 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 scattered crossover function and Gaussian mutation function.

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

The final solution of Algorithm 2 gives the number of best acoustic and visual recognition hypotheses to be considered to measure the correct reliability of each modality and the optimum integration weight multiplier (x).

4 Experimental Database, Audio, and Visual Speech Features

This paper focuses on a slightly different type of AVSR system which use visual features extracted from side-face mouth region images rather than frontal face images. 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 [27]. With reference to the above, as well as to make the system more practical, around 70 commonly used mobile functions (isolated words) were recorded 30 times each by a microphone and web camera located near the speaker's right cheek mouth region. Samples of the recorded side-face videos are shown in Fig. 1. 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 [10]. Most of the audio-visual 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 speakers mouth, view angle of the camera etc. as constant. But in this work, the recording was done in the office environment on different days with different values for the above factors to make the database suitable for real life applications. Also, the database included natural environment noises such as fan noise, birds sounds, sometimes other people speaking and shouting sounds.

Figure 1 Example video frames of multi-speaker side-face audio-visual speech database recorded in a typical office environment.

4.1 Acoustic Feature Extraction

The acoustic speech was recorded at the rate of 8 kHz with 16-bit resolution. The popular Mel-frequency cepstral coefficients (MFCC) are extracted from the acoustic speech signal [13]. The frequency analysis of the signal is performed for each frame segmented by the Hamming window having the length of 32 ms and an overlap of 12.5 ms. For each frame, we perform the Fourier analysis, computation of the logarithm of the Mel-scale filter bank energy, and the discrete cosine transformation. The cepstral mean subtraction (CMS) method is applied to remove channel distortions existing in the speech data [15]. As a result we obtain 39 acoustic parameters: 12 MFCCs, 12 ΔMFCCs, 12 ΔΔMFCCs, log energy, Δ log energy, and ΔΔ log energy.

4.2 Visual Feature Extraction

Visual features proposed in the literature of AVSR can be categorized into shape-based, pixel-based and

motion-based features [28]. 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 [10]. This work focuses on the relative benefits of both static and dynamic features for improved AVSR recognition.

4.2.1 DCT Based Static Feature Extraction

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/s and [240 × 320] pixel resolutions. Prior to the image transform the recorded video frames are converted to equivalent RGB image. This RGB image is converted to 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 obtain [16 × 16] and then passed as the input to the DCT. The DCT returns a 2D matrix 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 × 6] triangle region DCT coefficients without the DC component are considered as 20 static features of a frame.

4.2.2 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. The MHI is defined as

$$MHI = Max \bigcup_{t=1}^{N-1} DOF_t(m,n) \times t \tag{19}$$

where N represents the number of frames used to capture the side-face mouth region motion and DOF is the binarized difference image over a threshold. The threshold is optimized through experimentation. In Eq. 19, to show the recent movements with brighter value, the binarized version of the DOF is multiplied with a ramp of time and integrated temporally. Next, DCT was applied to MHI and the transformed coefficients are obtained. Similar to static feature extraction, only [6 × 6] triangle region DCT coefficients without the DC component are considered as the dynamic features. Finally, the static and dynamic features are concatenated to represent visual speech.

5 HMM Training and Recognition Results

The HMM is a commonly used classifier in speech recognition, since it has the desirable property that it can readily be model the time-varying speech signal [15]. This work 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 [15]. 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.

5.1 Recognition Results

The bimodal decision fusion speech recognition system using side-face mouth region image is shown in Fig. 2. The dataset was recorded in an office environment with a background noise. Each word was recorded 30 times for each speaker, hence we have a total of 90 samples/word. Out of these 90 recorded samples, 60 samples were taken randomly for training the HMMs and 15 samples have been used as a validation data to estimate the best acoustic and visual recognition hypotheses length N_A and N_V to measure the correct reliabilities using the proposed Algorithm 1. The same set of samples have been used to estimate the optimum integration weight using the proposed Algorithm 2. The remaining 15 samples were artificially degraded with additive white Gaussian noise at SNRs of 20, 10, 5, 0, and −5 dB. Theses noisy samples have been used as a test data to compute the recognition accuracy. The experiment was conducted three times for each SNR, in

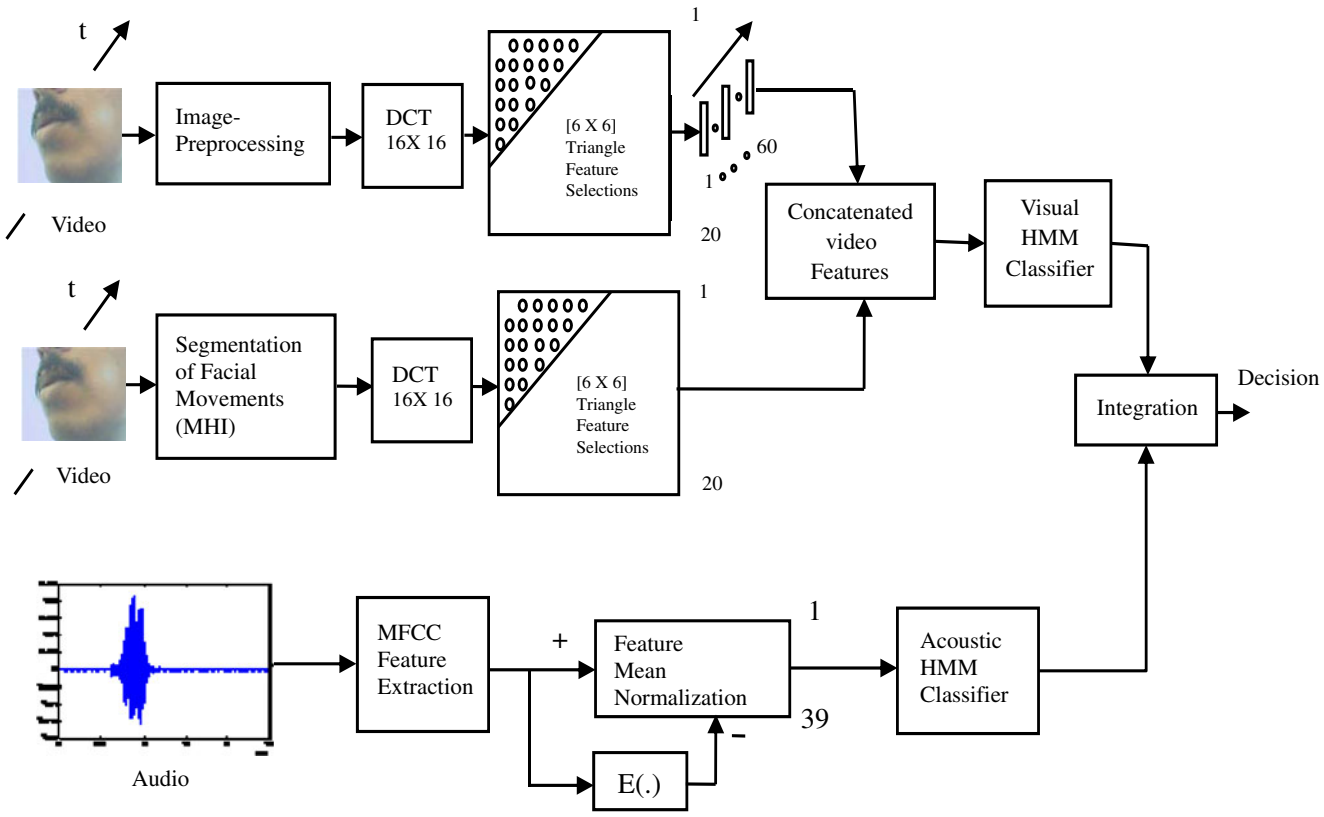

Figure 2 Block diagram of the proposed audio-visual decision fusion speech recognition system using mouth region side-face images.

each trial 60 samples were taken randomly for training and 15 samples for testing. Finally, the average of all three trials are taken as the recognition accuracy.

Table 2 shows recognition accuracies obtained by the audio-only, visual-only, audio-visual baseline reliability ratio (AV baseline-RR), audio-visual N best recognition hypotheses reliability ratio (AV N-best RR), and the proposed GA adaptive reliability ratio (AV-GA adaptive-RR) and GA adaptive reliability and optimized (AV-GA adaptive RR & GA optimized) bimodal systems at various SNR conditions. Similarly Fig. 3 compares the unimodal and bimodal system's recognition performance. In Table 2, "Clean" means the recorded speech samples without any additional white Gaussian noise. From the results (Table 2), the following observations were made,

1. The acoustic-only recognition shows nearly 77% for the recorded speech but, as the speech contains more artificially added white Gaussian noise, its performance is degraded sharply; the recognition is even less than 2% at −5 dB SNR conditions. Since the maximum recognition accuracy for the recorded speech is 77%, it shows that the recorded speech itself is highly noisy.
2. The average recognition accuracy for visual-only system is 62.57%, which appears constant regardless of acoustic noise conditions.
3. The baseline reliability ratio-based and N best recognition hypotheses reliability ratio-based bi-

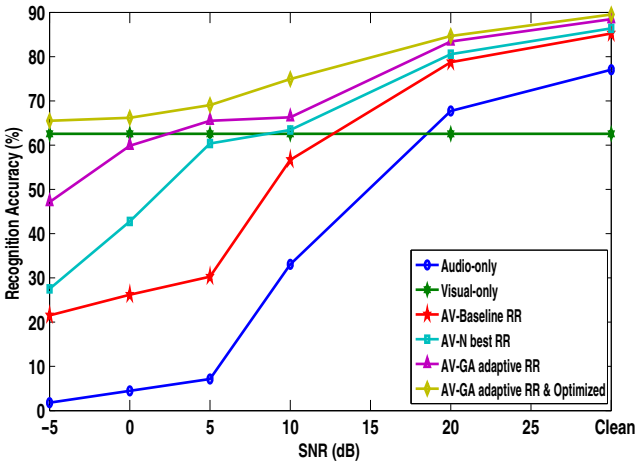

Figure 3 Performance of the existing unimodal and bimodal systems and the proposed AV-GA adaptive-RR and AV-GA adaptive RR & GA optimized bimodal systems in recognition accuracy (%).

Table 2 AVSR performance in recognition accuracy (%) by the audio-only, visual-only, unimodal systems and AV baseline-RR, AV *N*-best RR, and the proposed [1] & [2] bimodal systems.

SNR	Audio only (%)	Visual only (%)	AV baseline-RR (%)	AV *N*-best RR (%)	AV-GA adaptive-RR (%)	AV-GA adaptive RR & GA optimized (%)
Clean	77.05	62.57	85.24	86.38	88.47	89.52
20 dB	67.71	62.57	78.76	80.57	83.43	84.67
10 dB	33.05	62.57	56.67	63.43	66.28	74.95
5 dB	07.14	62.57	30.28	60.38	65.52	69.05
0 dB	04.87	62.57	26.19	42.76	59.86	66.19
−5 dB	01.81	62.57	21.52	27.52	47.14	65.52
Average(%)						
(−5 dB ∼ Clean)	31.87	62.57	49.77	60.17	68.45	74.98
(−5 dB ∼ 10 dB)	11.62	62.57	33.66	48.52	59.70	68.93

modal systems show performance improvement only under Clean and 20 dB SNR conditions over the acoustic-only and visual-only systems but, under the remaining SNR conditions (i.e., −5 dB ∼ 10 dB) their performances are inferior to that of visual-only system i.e. they show attenuating fusion at these SNR conditions.

4. The proposed [1] GA adaptive reliability ratio-based system improves the recognition accuracy over baseline reliability ratio and *N* best recognition hypotheses reliability ratio-based systems under most of the SNR conditions.
5. The proposed [2] GA adaptive reliability ratio and optimized bimodal system further improves the recognition accuracy and shows performance improvement as compared to all other systems at all SNR conditions. Especially, the performance improvement is larger when the SNR is small i.e. −5 dB ∼ 10 dB. This demonstrates that the noise-robustness of recognition is achieved by the proposed [2] system for the recorded noisy audio-visual speech recognition task.

6 Discussion and Conclusion

In this paper, influence of the reliability measure on integration weight estimation is demonstrated via 70 commonly used mobile functions isolated words audio-visual speech database of three speakers. The proposed systems use 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. Generally, the dynamic visual speech features are obtained by derivative 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 systems, the recognition accuracy is compared with other two related methods namely reliability ratio based method and *N*-best recognition hypothesis reliability ratio based method. The results in Table 2 clearly show overall performance improvement by the proposed systems over the existing unimodal and bimodal systems in a noisy side face audio-visual speech database. Especially at low SNRs, all the reported methods show very poor recognition accuracy. But the proposed methods solve this issue and improve the recognition accuracy considerably.

In the proposed works, the fusion happens at the end of each utterance based on a single reliability measure for the whole utterance. This will not be able to effectively account for time-varying noise conditions where the reliability will also vary during the duration of the utterance. This problem can be solved to a certain extent, when we measure the reliability of both acoustic and visual signal in a frame-by-frame basis and fuse the decision to find the correct utterance. If we do so, the complexity of the algorithms become very high even for the isolated word recognition task. This is one of the drawbacks of the proposed algorithms, in view of handling time-varying noises in the acoustic signal. Also, the amount of training data used to model the HMMs are low, hence we obtained very poor recognition accuracy at low SNRs. In future, we will record more samples from other speakers and model the HMMs in more reliable form to obtain reasonable accuracy at low SNRs.

The proposed works consider effect of noise only in audio signal. Since we recorded the video signal so close to the speaker's mouth region the possibility of video distortion is less. This assumption may not be

correct always. So, in future we will include the effect of noise in the recorded video signal and also apply these proposed algorithms to see their performance.

And more over, the proposed algorithms in their current form are suitable for isolated word recognition because, we can easily calculate the combined likelihood based on a single reliability measure. But, if we extend these algorithms to the continuous speech recognition task it become very challenging because we have to unmanageably consider many possible word or phoneme sequence hypotheses to calculate the combined likelihood. With these considerations, further investigations on applying the proposed algorithms to complex tasks such as continuous speech recognition are in progress.

References

1. Iwano, K., Yoshinaga, T., Tamura, S., & Furui, S. (2007). Audio-visual speech recognition using lip information extracted from side-face images. *EURASIP Journal on Audio, Speech, and Music Processing, 2007*, 64 506+.
2. Lee, J. S., & Park, C. H. (2008). Adaptive decision fusion for audio-visual speech recognition. In F. Mihelic & J. Zibert (Eds.), *Speech recognition, technologies and applications* (pp. 550). Vienna, Austria.
3. Lee, J. S., & Park, C. H. (2008). Robust audio-visual speech recognition based on late integration. *IEEE Transaction on Multimedia, 10*, 767–779.
4. Meyer, G. F., Mulligan, J. B., & Wuerger, S. M. (2004). Continuous audiovisual digit recognition using *N*-best decision fusion. *Information Fusion, 5*, 91–101.
5. Rogozan, A., & Deléglise, P. (1998). Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication, 26*, 149–161.
6. Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audio-visual speech. In *Proceddings of IEEE* (Vol. 91, no. 9).
7. Dupont, S., & Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia, 2*, 141–151.
8. Seymour, R., Stewart, D., & Ming, J. (2007). Audio-visual integration for robust speech recognition using maximum weighted stream posteriors. In *Proc. of INTERSPEECH* (pp. 654–657).
9. Chibelushi, C. C., Deravi, F., & Mason, J. S. D. (2002). A review of speech based bimodal recognition. *IEEE Transactions on Multimedia, 4*, 23–37.
10. Rajavel, R., & Sathidevi, P. S. (2009). Static and dynamic features for improved HMM based visual speech recognition. In *Proc. of 1st international conference on intelligent human computer interaction* (pp. 184–194). Allahabad, India.
11. Yau, W. C., Kumar, W. C., & Arjunan, S. P. (2006). Voiceless speech recognition using dynamic visual speech features. In *Proc. of HCSNet workshop on the use of vision in HCI*. Canberra, Australia.
12. Yau, W. C., Kumar, D. K., & Weghorn, H. (2007). Visual speech recognition using motion features and Hidden Markov models. In M. Kampel & Hanbury, A. (Eds.), *LNCS* (pp. 832–839). Heidelberg: Springer.
13. Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). 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.
14. Seymour, R., Stewart, D., & Ming, J. (2008). Comparison of image transform-based features for visual speech recognition in clean and corrupted videos. *EURASIP Journal on Image and Video Processing, 2008*. doi:10.1155/2008/810362.
15. Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs: Prentice Hall.
16. Haupt, R. L., & Haupt, S. E. (2004). *Practical genetic algorithm*. Hoboken, NJ: Wiley.
17. Chen, T. (2001). Audiovisual speech processing. Lip reading and lip synchronization. *IEEE Signal Processing Magazine, 18*, 9–21.
18. Petajan, E. D. (1984). Automatic lipreading to enhance speech recognition. In *Proc. global telecommunications conf.* (265–272). Atlanta.
19. Wang, X., Hao, Y., Fu, D., & Yuan, C. (2008). Audio-visual automatic speech recognition for connected digits. In *2nd international symposium on intelligent information technology application* (pp. 328–332). China.
20. Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*, 339–355.
21. Summerfield, A. Q. (1987). 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* (pp. 3–51). London: Lawrence Erlbarum.
22. Benoit, C., Mohamadi, T., & Kandel, S. D. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research, 37*, 1195–1203.
23. Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., et al. (2000). *Audio visual speech recognition, final workshop 2000 report, CLSP*. The Johns Hopkins University, Baltimore.
24. Teissier, P., Robert-Ribes, J., & Schwartz, J. L. (1999). Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing, 7*, 629–642.
25. Chibelushi, C. C., Deravi, F., & Mason, J. S. D. (2002). A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia, 4*(1), 23–37.
26. Silsbee, P. L. (1994). Sensory integration in audiovisual automatic speech recognition. In *28th annual asilomar conference on signals, systems, and computers* (Vol. 1, pp. 561–565).
27. Potamianos, G., Neti, C., Huang, J., Connell, J. H., Chu, S., Libal, V., et al. (2004). Towards practical development of audio-visual speech recognition. In *IEEE international conf. on acoustic, speech, and signal processing*, (pp. iii777–780). Canada.
28. Foo, S. W., & Dong, L. (2002). 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* (pp. 607–614). Berlin Heidelberg: Springer-Verlag.
29. Verma, A., Faruquie, T., Neti, C., & Basu, S. (1999). Late integration in audiovisual continuous speech recognition. In *Proc. workshop on automatic speech recognition and understanding* (pp. 71–74). Keystone.
30. Tamura, S., Iwano, K., & Furui, S. (2005). A stream-weight optimization method for multi-stream HMMs based on like-

lihood value normalization. In *Proc. of ICASSP* (Vol. 1, pp. 469–472). Philadelphia.
31. Adjoudani, A., & Benoit, C. (1996). On the integration of auditory and visual parameters in an HMM-based ASR. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and speech recognition, technologies and applications* (pp. 461–472). Berlin, Germany: Springer.
32. Lewis, T. W., & Powers, D. M. W. (2004). Sensor fusion weighting measures in audio-visual speech recognition. In *Proceedings of 27th conf. Australasian computer science* (pp. 305–314). Dunedin, New Zealand.

R. Rajavel obtained his Bachelor's Degree in Electronics and Communication Engineering from the University of Madras in 1999. He received the Master of Engineering Degree in Instrumentation and Process Control Engineering from the Annamalai University in 2002. His general research interests include signal processing, speech recognition and medical image processing. He was a research scholar in National Institute of Technology Calicut, India from June 2006 to April 2010. Currently, he is on the research staff of Network Systems and Technologies, at Technopark, Trivandrum, India.

P. S. Sathidevi is currently serving as Professor in the Department of Electronics and Communication Engineering at National Institute of Technology Calicut, India. She received B.Tech. Degree in Electronics Engineering from Regional Engineering College, Calicut, India in 1985, M.Tech. in Electronics from Cochin University of Science and Technology, Cochin, India in 1987 and Ph.D from Regional Engineering College, Calicut, India in 2003, in the field of Speech and Audio Processing. Her current research interests include speech processing, perceptual audio coding, image processing, Computational Auditory Scene Analysis (CASA) and cryptography. She has over 50 publications to her credit in various international journals and conferences.