Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 294010, 14 pages
doi:10.1155/2011/294010

Research Article

On the Soft Fusion of Probability Mass Functions for Multimodal Speech Processing

D. Kumar, P. Vimal, and Rajesh M. Hegde

Department of Electrical Engineering, Indian Institute of Technology, Kanpur 208016, India

Correspondence should be addressed to Rajesh M. Hegde, rhegde@iitk.ac.in

Received 25 July 2010; Revised 8 February 2011; Accepted 2 March 2011

Academic Editor: Jar Ferr Yang

Multimodal speech processing has been a subject of investigation to increase robustness of unimodal speech processing systems. Hard fusion of acoustic and visual speech is generally used for improving the accuracy of such systems. In this paper, we discuss the significance of two soft belief functions developed for multimodal speech processing. These soft belief functions are formulated on the basis of a confusion matrix of probability mass functions obtained jointly from both acoustic and visual speech features. The first soft belief function (BHT-SB) is formulated for binary hypothesis testing like problems in speech processing. This approach is extended to multiple hypothesis testing (MHT) like problems to formulate the second belief function (MHT-SB). The two soft belief functions, namely, BHT-SB and MHT-SB are applied to the speaker diarization and audio-visual speech recognition tasks, respectively. Experiments on speaker diarization are conducted on meeting speech data collected in a lab environment and also on the AMI meeting database. Audiovisual speech recognition experiments are conducted on the GRID audiovisual corpus. Experimental results are obtained for both multimodal speech processing tasks using the BHT-SB and the MHT-SB functions. The results indicate reasonable improvements when compared to unimodal (acoustic speech or visual speech alone) speech processing.

1. Introduction

Multi-modal speech content is primarily composed of acoustic and visual speech [1]. Classifying and clustering multimodal speech data generally requires extraction and combination of information from these two modalities [2]. The streams constituting multi-modal speech content are naturally different in terms of scale, dynamics, and temporal patterns. These differences make combining the information sources using classic combination techniques difficult. Information fusion [3] can be broadly classified as sensor level fusion, feature level fusion, score-level fusion, rank-level fusion, and decision-level fusion. A hierarchical block diagram indicating the same is illustrated in Figure 1. Number of techniques are available for audio-visual information fusion, which can be broadly grouped into feature fusion and decision fusion. The former class of methods are the simplest, as they are based on training a traditional HMM classifier on the concatenated vector of the acoustic and visual speech features, or an appropriate transformation on it. Decision fusion methods combine the single-modality (audio-only and visual-only) HMM classifier outputs to recognize audio-visual speech [4, 5]. Specifically, class conditional log-likelihoods from the two classifiers are linearly combined using appropriate weights that capture the reliability of each classifier, or feature stream. This likelihood recombination can occur at various levels of integration, such as the state, phone, syllable, word, or utterance level. However, two of the most widely applied fusion schemes in multi-modal speech processing are concatenative feature fusion (early fusion) and coupled hidden Markov models (late fusion).

1.1. Feature Level Fusion. In the concatenative feature fusion scheme [6], feature vectors obtained from audio and video modalities are concatenated and the concatenated vector is used as a single feature vector. Let the time synchronous acoustic and visual speech features at instant t, be denoted by $O_S^{(t)} \in R^{D_s}$, where D_s is the dimensionality of the feature vector, and $s = A, V$, for audio and video modalities, respectively. The joint audio-visual

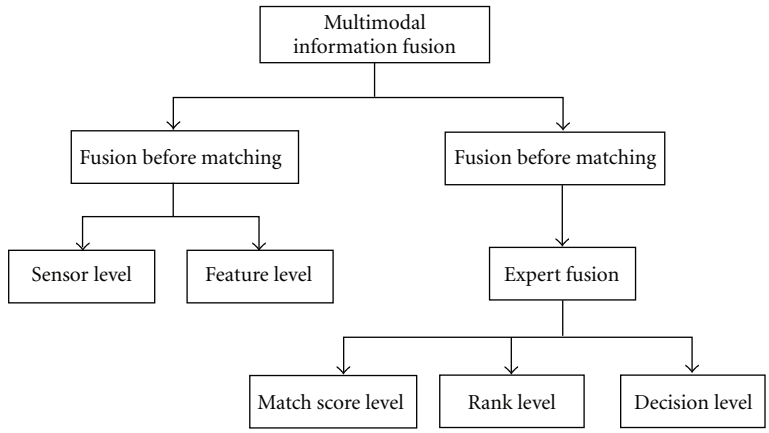

Figure 1: Levels of multi-modal information fusion.

feature vector is then simply the concatenation of the two, namely

$$O^{(t)} = \left[{O_A^{(t)}}^T, {O_V^{(t)}}^T \right]^T \in R^D, \tag{1}$$

where $D = D_A + D_V$. These feature vectors are then used to train HMMs as if generated from single modality and are used in the speech processing and recognition process. Hierarchical fusion using feature space transformations like hierarchical LDA/MLLT [6], are also widely used in this context. Another class of fusion schemes uses a decision fusion mechanism. Decision fusion with adaptive weighting scheme in HMM-based AVSR systems is performed by utilizing the outputs of the acoustic and the visual HMMs for a given audiovisual speech datum and then fuse them adaptively to obtain noise-robustness over various noise environments [7]. However, the most widely used among late fusion schemes is the coupled hidden Markov model (CHMM) [8].

1.2. Late Fusion Using Coupled Hidden Markov Models. A coupled HMM can be seen as a collection of HMMs, one for each data stream, where the discrete nodes at time t for each HMM are conditioned by the discrete nodes at time $t-1$ of all the related HMMs. Parameters of a CHMM are defined as follows

$$\begin{aligned} \pi_o^c(i) &= P(q_t^c = i), \\ b_t^c(i) &= P(O_t^c \mid q_t^c = i), \\ a_{i|j,k}^c &= P(q_t^c = i \mid q_{t-1}^0 = j,\ q_{t-1}^1 = k), \end{aligned} \tag{2}$$

where q_t^c is the state of the couple node in the cth stream at time t. In a continuous mixture with Gaussian components, the probabilities of the observed nodes are given by

$$b_t^c(i) = \sum_{m=1}^{M_i^c} w_{i,m}^c N\left(O_t^c, \mu_{i,m}^c, U_{i,m}^c\right), \tag{3}$$

where $\mu_{i,m}^c$ and $U_{i,m}^c$ are the mean and covariance matrix of the ith state of a coupled node, and mth component of the associated mixture node in the cth channel. M_i^c is the number of mixtures corresponding to the ith state of a coupled node in the cth stream and the weight $w_{i,m}^c$ represents the conditional probability $P(s_t^c = m \mid q_t^c = i)$ where s_t^c is the component of the mixture node in the cth stream at time t. A schematic illustration of a coupled HMM is shown in Figure 2. Multimodal information fusion can also be classified as hard and soft fusion. Hard fusion methods are based on probabilities obtained from Bayesian theory which generally place complete faith in a decision. However, soft fusion methods are based on principles of Dempster-Shafer theory or Fuzzy logic which involve combination of beliefs and ignorances. In this paper, we first describe a new approach to soft fusion by formulating two soft belief functions. This formulation uses confusion matrices of probability mass functions. Formulation of the two soft belief functions is discussed first. The first belief function is suitable for binary hypothesis testing (BHT) like problems in speech processing. One example for a BHT-like problem is speaker

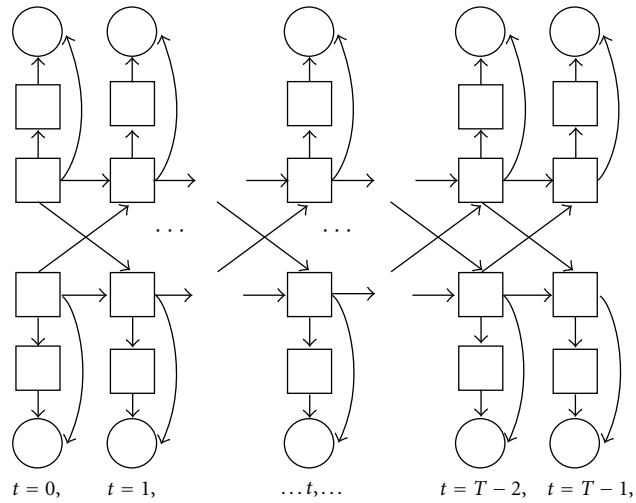

Figure 2: The audio-visual coupled HMM.

diarization. The second soft belief function is suitable for multiple hypothesis testing (MHT) like problems in speech processing, namely audio-visual speech recognition. These soft belief functions are then used for multi-modal speech processing tasks like speaker diarization and audio-visual speech recognition on the AMI meeting database and the GRID corpus. Reasonable improvements in performance are noted when compared to the performance using unimodal (acoustic speech or visual speech only) methods.

2. Formulation of Soft Belief Functions Using Matrices of Probability Mass Functions

Soft information fusion refers to a more flexible system to combine information from audio and video modalities for making better decision. The Dempster Shafer (DS) theory is a mathematical theory of evidence [9]. It allows one to combine evidence from different sources and arrive at a degree of belief (represented by belief function) that takes into account all the available evidences. DS theory is a generalization of the Bayesian theory of subjective probability. While the Bayesian theory requires probabilities for each question of interest, belief functions allow us to have degrees of belief for one question on probabilities of a related question.

2.1. Belief Function in Dempster Shafer Theory. Dempster Shafer theory of evidence allows the representation and combination of different measures of evidence. It is essentially a theory that allows for soft fusion of evidence or scores. Let

$$\Theta = (\theta_1, \ldots, \theta_k) \tag{4}$$

be a finite set of mutually exclusive and exhaustive hypothesis referred as singletons and Θ is referred as a frame of discernment. A basic probability assignment is a function m such that

$$m : 2^{\Theta} \longrightarrow [0, 1] \tag{5}$$

where

$$\sum_{A \subset \Theta} m(A) = 1, \qquad m(\Phi) = 0. \tag{6}$$

If $\neg A$ is complementary set of A, then by DS Theory

$$m(A) + m(\neg A) < 1, \tag{7}$$

Which is in contrast to probability theory. This divergence from probability is called Ignorance. The function assigning sum of masses of all the subsets of the set of interest is called the belief function and is given by

$$\text{Bel}(A) = \sum_{B \subset A} m(B). \tag{8}$$

A belief function assigned to each subset of θ is a measure of total belief in the preposition represented by the subset. This definition of the belief function is used to formulate the soft belief functions proposed in the following sections.

Table 1: Reliability of the unimodal features.

Classification feature	Reliability
Acoustic speech: X_{audio}	x
Visual speech: X_{video}	y

3. A Soft Belief Function for Binary Hypothesis Testing-Like Problems in Speech Processing

This section describes the proposed methodology of using the confusion matrices of probability mass functions to combine decisions obtained from acoustic and visual speech feature streams. The degree of belief for a decision is determined from subjective probabilities obtained from the two modalities and then are combined using Dempster's rule, making a reasonable assumption that the modalities are independent.

3.1. Probability Mass Functions for Binary Hypothesis Testing-Like Problems. The probability mass function (PMF) in D-S theory defines a mass distribution based on the reliability of the individual modalities. Consider two unimodal (acoustic or visual speech feature) decision scenarios as follows

X_{audio}: the audio feature-based decision.

X_{video}: the video feature-based decision.

On the other hand let us consider a two hypothesis problem (H_1 or H_2) of two exclusive and exhaustive classes, which we are looking to classify with the help of above feature vectors. Both X_{audio} and X_{video} can hypothesize as H_1 or H_2. Thus the focal elements of both the features are H_1, H_2 and Ω, where Ω is the whole set of classes $\{H_1, H_2\}$. The unimodal source reliabilities provide us with a certain degree of trust that we should have on the decision of that modality. The reliabilities of acoustic and visual speech-based decisions is decided on the number of times the X_{audio} and X_{video} classifies the given data correctly. At a particular time interval, the acoustic speech features give a certain probability of classification. If $P(X_{\text{audio}} = H_1) = p_1$, then the mass distribution is $m_{\text{audio}}(H_1) = xp_1$. Similarly, the mass assigned to H_2 is $m_{\text{audio}}(H_2) = x(1 - p_1)$. The remaining mass, is allocated to the whole set of discernment, $m_{\text{audio}}(\Omega) = 1 - x$. Similarly we assign a mass function for the visual speech feature-based decision.

3.2. Generating Confusion Matrix of Probability Mass Functions. It is widely accepted that the acoustic and visual feature-based decisions are independent of each other. Dempster's rule of combination can therefore be used for arriving at a joint decision given any two modalities. However, there are three PMFs corresponding to the two hypothesis. The two mass functions with respect to hypothesis H_1 and H_2 and the mass function corresponding to the overall set of discernment make up the three PMFs. Since we have three mass functions corresponding to each modality, a confusion matrix of one versus the other can be formed.

The confusion-matrix of PMFs thus obtained for the audio-visual speech features combined is shown in Table 2.

3.3. Formulating the Soft Belief Function Using the Confusion Matrix of Mass Functions. The premise of coming up with such a confusion matrix is due to the fact that the two modalities under consideration carry complementary information. Hence if the decisions of the two modalities are inconsistent, their product of masses is assigned to a single measure of inconsistency, say k. From Table 2, total inconsistency k is defined as

$$k = xyp_1(1-p_2) + xyp_2(1-p_1). \tag{9}$$

Hence the combined belief in hypothesis H_1 and H_2, obtained from the multiple modalities (speech and video) can now be formulated as

$$\begin{aligned}\mathrm{Bel}(H_1) &= \frac{xyp_1p_2 + xp_1(1-y) + (1-x)yp_2}{(1-k)},\\ \mathrm{Bel}(H_2) &= \frac{xy(1-p_1)(1-p_2) + x(1-p_1)(1-y)}{(1-k)}\\ &\quad + \frac{(1-x)y(1-p_2)}{(1-k)}.\end{aligned} \tag{10}$$

Note that the mass functions have been normalized by the factor $(1-k)$. The soft belief function for BHT-like problems (BHT-SB), formulated in (10), gives a soft decision measure for choosing a better hypothesis from the two possible classifications.

3.4. Multimodal Speaker Diarization As a Binary Hypothesis Testing-Like Problem in Speech Processing. In the context of audio document indexing and retrieval, speaker diarization [10, 11], is the process which detects speakers turns and re groups those uttered by the same speaker. It is generally based on a first step of segmentation and often preceded by a speech detection phase. It also involves partitioning the regions of speech into sufficiently long segments of only one speaker. This is followed by a clustering step that consists of giving the same label to segments uttered by the same speaker. Ideally, each cluster corresponds to only one speaker and vice versa. Most of the systems operate without specific a priori knowledge of speakers or their number in the document. They generally need specific tuning and parameter training. Speaker diarization [10], can hence be considered as BHT-like problem since we only have two hypothesis to decide on. Hypothesis H_1 decides on a speaker change detected and hypothesis H_2 decides on speaker change not detected. Hence the aforementioned BHT-SB function is used on the multi-modal speaker diarization task [11], in the section on performance evaluation later in this paper.

4. A Soft Belief Function for Multiple Hypothesis Testing-Like Problems in Speech Processing

In this section we describe the formulation of a soft belief function for multiple hypothesis Testing-Like problems in speech processing, by taking an example of audio-visual speech recognition. Audio-visual speech recognition can be viewed as a multiple hypothesis testing problem, depending on the number of words in the dictionary. More specifically audio-visual speech recognition is an N hypothesis problem, where each utterance has N possible options to be classified into.

4.1. Probability Mass Functions for Multiple Hypothesis Testing-Like Problems. Consider the following multiple hypothesis testing scenario for word-based speech recognition

H_1: word 1

H_2: word 2

. . .

H_N: word N.

Recognition probabilities from individual modalities are given by (11)

$$P(X_{\text{audio}} = H_i) = A_i; \quad P(X_{\text{Video}} = H_i) = V_i; \quad 1 \le i \le N. \tag{11}$$

The problem is to find out most likely hypothesis by using X_{Audio} and X_{Video}, where

X_{audio}: the acoustic speech feature-based decision.

X_{video}: the visual speech feature-based decision.

The reliability of audio and video based decision is as given in Table 3.

4.2. Generating Confusion Matrix of Probability Mass Functions. The premise that acoustic and visual feature-based decisions are independent of each other can still be applied to a audio-visual speech recognition problem. Dempster's rule of combination can therefore be used for arriving at a joint decision given any two modalities even in this case. However, there are $(N+1)$ PMFs, as we are dealing with a N (multiple) hypothesis problem. The $N+1$ mass functions with respect to hypothesis H_1 through HN and the mass function corresponding to the overall set of discernment make up the $N+1$ PMFs. Since we have three mass functions corresponding to each modality, a confusion matrix of one versus the other can be formed. The confusion-matrix of probability mass functions (PMFs), for this "N" hypothesis problem is shown in Table 4.

4.3. Formulating the Soft Belief Function Using the Confusion Matrix of Mass Functions. From Table 4, the total inconsistency k is given by

$$k = \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N} xyA_iV_j. \tag{12}$$

Table 2: The confusion-matrix of probability mass functions (PMFs) for multi-modal features.

	$m_v(H_1) = yp_2$	$m_v(H_2) = y(1-p_2)$	$m_v(\Omega) = 1-y$
$m_a(H_1) = xp_1$	$m_{a,v}(H_1) = xyp_1p_2$	$k = xyp_1(1-p_2)$	$m_{a,v}(H_1) = x(1-y)p_1$
$m_a(H_2) = x(1-p_1)$	$k = xyp_2(1-p_1)$	$m_{a,v}(H_2) = xy(1-p_1)(1-p_2)$	$m_{a,v}(H_2) = x(1-y)(1-p_1)$
$m_a(\Omega) = 1-x$	$m_{a,v}(H_1) = (1-x)yp_2$	$m_{a,v}(H_2) = (1-x)y(1-p_2)$	$m_{a,v}(\Omega) = (1-x)(1-y)$

Table 3: Reliability of the unimodal features.

Classification feature	Reliability
Acoustic speech: X_{audio}	x
Visual speech: X_{video}	y

Hence, the combined belief in hypothesis H_k, $1 \le k \le N$, obtained from the multiple modalities (speech and video) can now be formulated as

$$\text{Bel}(H_k) = \frac{xyA_kV_k + x(1-y)A_k + (1-x)yV_k}{(1-k)}. \quad (13)$$

The soft belief function for MHT-like problems (MHT-SB), formulated in (13), gives a soft decision measure for choosing a better hypothesis from the N possible options.

4.4. Audio-Visual Speech Recognition As a Multiple Hypothesis Testing Problem. Audio-visual speech recognition (AVSR) is a technique that uses image processing capabilities like lip reading to aid audio-based speech recognition in recognizing indeterministic phones or giving preponderance among very close probability decisions. In general, lip reading and audio-based speech recognition works separately and then the information gathered from them is fused together to make a better decision. The aim of AVSR is to exploit the human perceptual principle of sensory integration (joint use of audio and visual information) to improve the recognition of human activity (e.g., speech recognition, speech activity, speaker change, etc.), intent (e.g., speech intent) and identity (e.g., speaker recognition), particularly in the presence of acoustic degradation due to noise and channel, and the analysis and mining of multimedia content. AVSR can be viewed as a multiple hypothesis Testing-Like problem in speech processing since there are multiple words to be recognized in a typical word-based audio-visual speech recognition system. The application of the aforementioned MHT-SB function to such a problem is discussed in the ensuing section on performance evaluation.

5. Performance Evaluation

5.1. Databases Used in Experiments on Speaker Diarization. In order to evaluate and compare the performance of the soft belief function for BHT-like problems, the BHT-SB is applied to a speaker diarization task on two databases. The first database is composed of multi-modal speech data recorded on the lab test bed and the second database is the standard AMI meeting corpus [12].

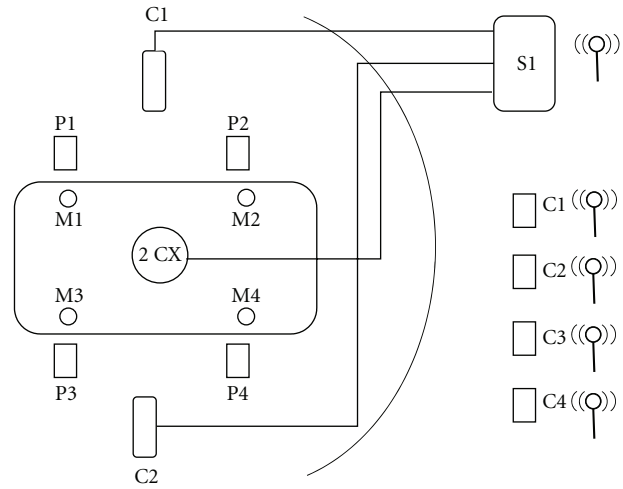

Figure 3: Layout of the lab test bed used to collect multi-modal speech data.

5.1.1. Multimodal Data Acquisition Test Bed. The experimental lab test bed is a typical meeting room setup which can accommodate four participants around a table. It is equipped with an eight-channel linear microphone array and a four channel video array, capable of recording each modality synchronously. Figure 3 represents layout of the test bed used in data collection for this particular set of experiments. C1, and C2 are two cameras; P1, P2, P3, P4 are four participants of the meeting; M1, M2, M3, M4 represents four microphones and S is the screen. It is also equipped with a two-channel microphone array (2CX), a server and computing devices. A manual timing pulse is generated to achieve start to end multi-modal synchronization. For the purpose of speaker diarization we use only one channel of audio data and two-channel of video data with each camera focusing on the participants face. The multi-modal data used in our experiments is eighteen minutes long, consisting of 3 speakers taking turns as in a dialogue, and the discussion was centered around various topics like soccer, research, and mathematics. Figure 4 shows the snapshot of the lab test bed used for acquiring the multi-modal data.

5.1.2. AMI Database. The AMI (augmented multi-party interaction) project [12] is concerned with the development of technology to support human interaction in meetings, and to provide better structure to the way meetings are run and documented. The AMI meeting corpus contains 100 hours of meetings captured using many synchronized recording devices, and is designed to support work in speech and video processing, language engineering, corpus linguistics,

Table 4: The confusion-matrix of probability mass functions for multi-modal features.

	$m_v(H_1) = yV_1$	$m_v(H_2) = yV_2$	$\cdots$	$m_v(H_N) = yV_N$	$m_v(\Omega) = 1 - y$
$m_a(H_1) = xA_1$	$m_{a,v}(H_1) = xyA_1V_1$	$k = xyA_1V_2$	$\cdots$	$k = xyA_1V_N$	$m_{a,v}(H_1) = x(1-y)A_1$
$m_a(H_2) = xA_2$	$k = xyA_2V_1$	$m_{a,v}(H_2) = xyA_2V_2$	$\cdots$	$k = xyA_2V_N$	$m_{a,v}(H_2) = x(1-y)A_2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$m_a(H_N) = xA_N$	$k = xyA_NV_1$	$k = xyA_NV_2$	$\cdots$	$m_{a,v}(H_N) = xyA_NV_N$	$m_{a,v}(H_N) = x(1-y)A_N$
$m_a(\Omega) = 1 - x$	$m_{a,v}(H_1) = (1-x)yV_1$	$m_{a,v}(H_2) = (1-x)yV_2$	$\cdots$	$m_{a,v}(H_N) = (1-x)yV_N$	$m_{a,v}(\Omega) = (1-x)(1-y)$

Figure 4: Snapshot of the actual test bed used to acquire multi-modal speech data.

Figure 5: AMI's instrumented meeting room (source: AMI website).

and organizational psychology. It has been transcribed orthographically, with annotated subsets for everything from named entities, dialogue acts, and summaries to simple gaze and head movement. Two-thirds of the corpus consists of recordings in which groups of four people played different roles in a fictional design team that was specifying a new kind of remote control. The remaining third of the corpus contains recordings of other types of meetings. For each meeting, audio (captured from multiple microphones, including microphone arrays), video (coming from multiple cameras), slides (captured from the data projector), and textual information (coming from associated papers, captured handwritten notes and the white board) are recorded and time-synchronized. The multi-modal data from the augmented multi-party interaction (AMI) corpus is used here to perform the experiments. It contains the annotated data of four participants. The duration of the meeting was around 30 minutes. The subjects in the meeting are carrying out various activities such as presenting slides, white board explanations and discussions round the table.

5.2. Database Used in Experiments on Audio-Visual Speech Recognition: The GRID Corpus. GRID [13] corpus is a large multitalker audio-visual sentence corpus to support joint computational behavioral studies in speech perception. In brief, the corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). Sentences are of the form "put red at g nine now".

5.2.1. Sentence Design. Each sentence consisted of a six word sequence of the form indicated in Table 5. Of the six components, three-color, letter, and digit were designated as keywords. In the letter position, "*w*" was excluded since it is the only multisyllabic English alphabetic letter. "Zero" was used rather than "oh" or "naught" to avoid multiple pronunciation alternatives for orthographic 0. Each talker produced all combinations of the three keywords, leading to a total of 1000 sentences per talker. The remaining components command, preposition, and adverb were fillers.

5.2.2. Speaker Population. Sixteen female and eighteen male talkers contributed to the corpus. Participants were staff and students in the Departments of Computer Science and Human Communication Science at the University of Sheffield. Ages ranged from 18 to 49 years with mean age being 27.4 years.

5.2.3. Collection. Speech material collection was done under computer control. Sentences were presented on a computer screen located outside the booth, and talkers had 3 seconds to produce each sentence. Talkers were instructed to speak in a natural style. To avoid overly careful and drawn-out utterances, they were asked to speak sufficiently quickly to fit into the 3 seconds time window.

5.3. Experiments on Speaker Diarization. In the ensuing sections we describe the experimental conditions for uni-modal speech diarization [14], and the proposed multi-modal speaker diarization using the BHT-SB function.

Table 5: Sentence structure for the Grid corpus. Keywords are identified with asterisks.

Command	Color*	Preposition	Letter*	Digit*	Adverb
bin	blue	at	A–Z	1–9, 0	again
lay	green	by	excluding W		now
place	red	in			please
set	white	with			soon

5.3.1. Speech-Based Unimodal Speaker Diarization. The BIC (bayesian information criterion) for segmentation and clustering based on MOG (mixture of gaussian) is used for the purpose of speech-based unimodal speaker diarization. The likelihood distance is calculated between two segments to determine whether they belong to the same speaker or not. The distances used for acoustic change detection can also be applied to speaker clustering in order to infer whether two clusters belong to the same speaker. For a given acoustic segment $\mathcal{X}_i$, the BIC value of a particular model $\mathcal{M}_i$, indicates how well the model fits the data, and is determined by (16). In order to detect the audio scene change between two segments with the help of BIC, one can define two hypothesis. Hypothesis 0 is defined as

$$H_0 : \mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N \sim \mathcal{N}(\mu, \Sigma), \tag{14}$$

which considers the whole sequence to consist no speaker change. Hypothesis 1 is defined as

$$\begin{aligned} H_1 : \mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_L &\sim \mathcal{N}(\mu_1, \Sigma_1), \\ \mathrm{x}_{L+1}, \mathrm{x}_{L+2}, \ldots, \mathrm{x}_N &\sim N(\mu_2, \Sigma_2) \end{aligned} \tag{15}$$

is the hypothesis that a speaker change occurs at time L. A check of whether the hypothesis H_0 better models the data as compared to the hypothesis H_1, for a mixture of Gaussian case can be done by computing a function similar to the generalized likelihood ratio as

$$\begin{aligned} \Delta\mathrm{BIC}(\mathcal{M}_i) = {} & \log(\mathcal{L}(\mathcal{X}, \mathcal{M})) \\ & - \left(\log(\mathcal{L}(\mathcal{X}_i, \mathcal{M}_i)) + \log\left(\mathcal{L}\left(\mathcal{X}_j, \mathcal{M}_j\right)\right)\right) \\ & - \lambda\Delta\#(i, j)\log(N), \end{aligned} \tag{16}$$

where $\Delta\#(i, j)$ is the difference in the number of free parameters between the combined and the individual models.

When the BIC value based on mixture of Gaussian model exceeds a certain threshold, an audio scene change is declared. Figure 6, illustrates a sample speaker change detection plot with speech information only using BIC. The illustration corresponds to the data from the AMI multimodal corpus. Speaker changes have been detected at 24, 36, 53.8 and 59.2 seconds. It is important to note here that the standard mel frequency cepstral coefficients (MFCC) were used as acoustic features in the experiments.

5.3.2. Video Based Unimodal Speaker Diarization Using HMMs. Unimodal speaker diarization based on video features uses frame-based video features for speaker diarization.

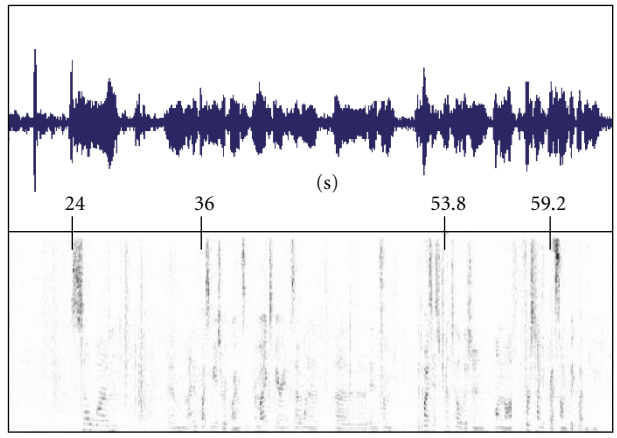

Figure 6: Speech-based unimodal speaker change detection.

Figure 7: Video frame of silent speaker.

Figure 8: Video frame of talking speaker.

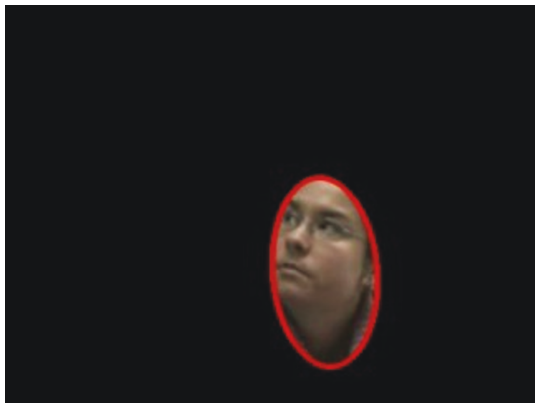

Figure 9: Extracted face of silent speaker.

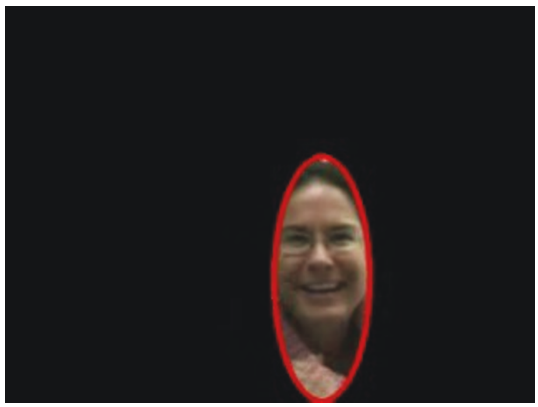

Figure 10: Extracted face of talking speaker.

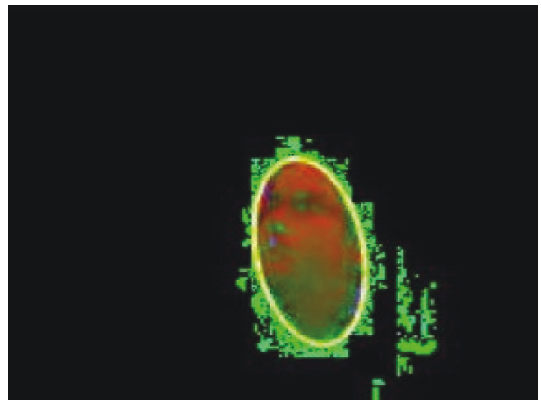

Figure 11: Hue plane of silent speaker.

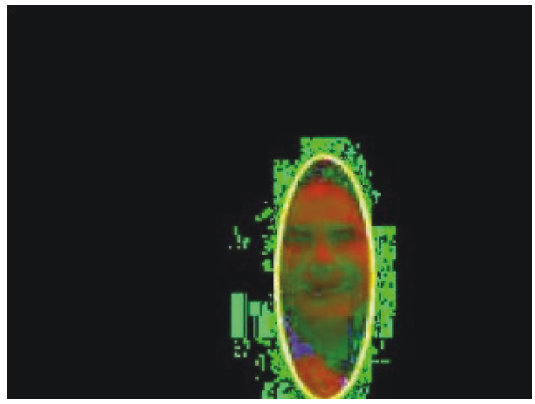

Figure 12: Hue plane of talking speaker.

The feature used is the histogram of the hue plane of the face pixels. The face of the speaker is first extracted from the video. The hue plane of the face region of each frame is then determined. The histogram of this hue plane in thirty-two bins is used as video feature vector. Hue plane features of the whole face are used and not just of the lips. This is primarily because the face contains a considerable amount of information from the perspective of changes in the hue plane. It was also noted from initial experiments that the changes in the hue plane of the face pixels when a person is speaking compared to when silent are significant. This histogram is then used as feature vector for training hidden Markov models. Figures 7, 8, 9, shows a frame of the video of a silent speaker from the AMI database, whose skin colored pixels are tracked and then the hue plane of the frame extracted. In Figures 10, 11, 12, a similar set of results are illustrated for the same speaker and from the same video clip, when she is speaking. Using the features extracted from the histogram of the hue plane, speaker diarization is now performed over a video segment of a certain duration by calculating the likelihood of the segment belonging to a model. The segment is classified as belonging to that speaker, for which the model likelihood is maximum. HMMs for each speaker are trained a priori using the video features. A speaker change is detected if the consecutive segments are classified as belonging to different models. The probability of speaker change is computed as the probability of two consecutive video segments belonging to two different models.

5.4. Experimental Results on Multimodal Speaker Diarization Using the BHT-Soft Belief Function. To facilitate for the synchronization of multi-modal data, that is, the video frame rate of 25 fps, and the speech sampling rate of 44100 Hz, we consider frame-based segment intervals for evaluating speaker change detection and subsequent speaker diarization. An external manual timing pulse is used for synchronization. The results obtained are compared with the annotated data of the AMI corpus. The multi-modal data recorded from the test bed has video frame rate of 30 fps and is manually annotated. Speaker diarization performance is usually evaluated in terms of diarization error rate (DER), which is essentially a sum of three terms namely, missed speech (speech in the reference but not in the hypothesis), false alarm speech (speech in the hypothesis but not in the reference), and speaker match error (reference and hypothesized speakers differ). Hence the DER is computed as

$$\mathrm{DER} = \frac{\mathrm{FA} + \mathrm{MS} + \mathrm{SMR}}{\mathrm{SPK}}\%, \tag{17}$$

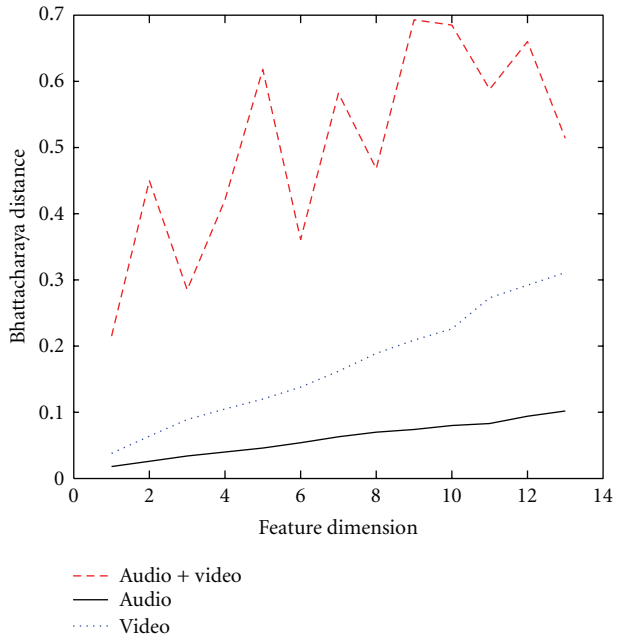

Figure 13: Separability analysis results as the BD versus the feature dimension for unimodal and multi-modal features.

where missed speaker time (MS) is the total time when less speakers are detected than what is correct, false alarm speaker time (FA) is the total time when more speakers are detected than what is correct, speaker match error time (SMR) is the total time when some other speaker is speaking rather than the speaker detected and scored speaker time (SPK) is the sum of every speakers utterance time as indicated in the reference.

5.4.1. Separability Analysis for Multimodal Features. In order to analyze the complementary nature of the acoustic and visual speech features, separability analysis is performed using the Bhattacharya distance as a metric. The Bhattacharya distance (BD), which is a special case of the Chernoff distance is a probabilistic error measure and relates more closely to the likelihood maximization classifiers that we have used for performance evaluation. Figure 13 illustrate the separability analysis results as the BD versus the feature dimension for both unimodal (speech only & video only) and multi-modal (speech + video) features in Figure 13. The complementarity of the multi-modal features when compared to unimodal speech features can be noted from Figure 13.

5.4.2. Experimental Results. The reliability of each feature is determined by its speaker change detection performance on a small development set created from unimodal speech or video data. The reliability values of the audio and video features computed from the development data set are given in Table 6, for the two corpora used in our experiments. The speaker diarization error rates (DER) for both the multi-modal corpora used is also shown in Figure 14. Reasonable reduction in DER is noted on using the BHT-SB function

Table 6: Reliability of the unimodal information as computed from their feature vectors on the two multi-modal data sets.

Unimodal Feature	Reliability on AMI corpus	Reliability on test bed data
Audio: X_{audio}	90.47	87.50
Video: X_{video}	87.50	78.04

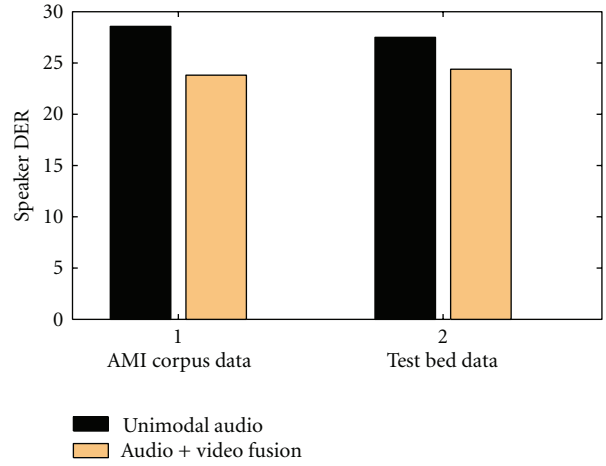

Figure 14: Speaker DER using unimodal audio and multi-modal information fusion on the two data sets.

as a soft fusion method when compared to the experimental results obtained from unimodal speech features.

5.4.3. Discussion on Speaker Diarization System Performance. Performance of speaker diarization system increases considerably when video information is fused with audio information as compared to audio only based system. A measure of the system performance, Diarization Error Rate (DER), is considerably low for system based on the proposed method of fusion of audio and visual information, as compared to audio only system. This result is shown in Figure 14, for the AMI database and also for multi-modal data from the lab testbed. Table 6, indicates that audio has been more reliable than video, which is quiet evident as there are certain sounds which can be produced without involving mouth movement (e.g., nasals). This fact is also reflected in Figure 13.

5.5. Experiments on Audio-Visual Speech Recognition. The potential for improved speech recognition rates using visual features is well established in the literature on the basis of psychophysical experiments. Canonical mouth shapes that accompany speech utterances have been categorized, and are known as visual phonemes or "visemes". Visemes [15], provide information that complements the phonetic stream from the point of view of confusability. A viseme is a representational unit used to classify speech sounds in the visual domain. This term was introduced based on the interpretation of the phoneme as a basic unit of speech in the acoustic domain. A viseme describes particular facial and

Table 7: Visemes as phoneme classes.

Viseme	Phoneme class
0	silence
1	f v w
2	s z
3	S Z
4	p b m
5	g k x n N r j
6	t d
7	l
8	I e:
9	E E:
10	A
11	@
12	i
13	O Y y u 2: o: 9 9: O:
14	a:

oral positions and movements that occur alongside the voicing of phonemes. Phonemes and visemes do not always share a one-to-one correspondence. Often, several phonemes share the same viseme. Thirty two visemes are required in order to produce all possible phoneme with the human face. If the phoneme is distorted or muffled, the viseme accompanying it can help to clarify what the sound actually was. Thus, visual and auditory components work together while communicating orally. Earlier experimental work on audio-visual speech recognition for recognizing digits can be found in [16]. Experimental work on recognizing words can be referred to in [17, 18], while the recognition of continuous speech is dealt with in [19–21].

5.5.1. Feature Vectors for Acoustic and Visual Speech. Acoustic features used in the experiments are the conventional mel frequency cepstral coefficients (MFCC), appended with delta and acceleration coefficients. Visual speech features are computed from the histogram of the lip region. To compute the visual speech feature, lip region is assumed to be in the lower half of the face part. We have used 70×110 pixel sized region, in the lower part of the face as lip region. To find out video feature vector, first we subtract RGB values of consecutive frames, so as to get motion vector video from the original video. Lip region is then extracted from this video and is converted to gray scale image by adding up the RGB values. A non-linear scale histogram of the pixel values of each frame, in 16 bins is found out and is used as feature vector. The sixteen bins are on a nonlinear scale. HMM models for these video features of each word utterance are trained for video only speech recognition. The visual evidence for the complementary information present in acoustic and visual features are illustrated in Figures 15, 16, and 17, 18. Illustrations for two situations where clean speech and noisy videos are available and vice versa are given in Figures 15, 16 and 17, 18, respectively.

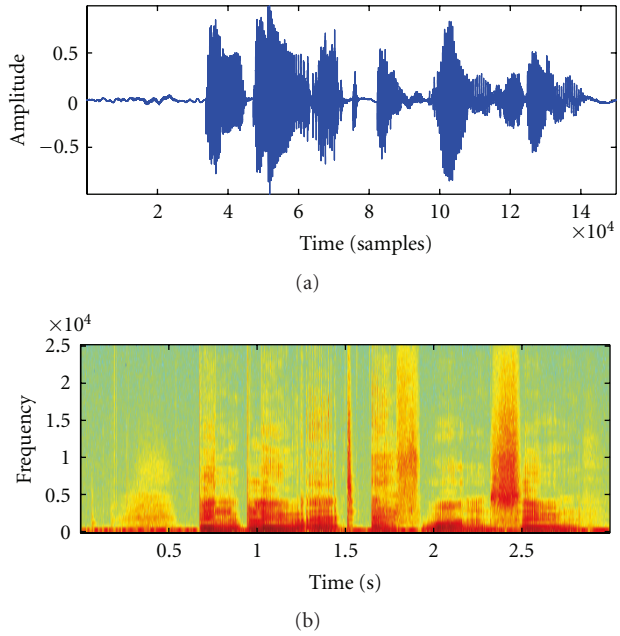

Figure 15: Clean speech signal and its spectrogram.

Figure 16: Noisy video signal.

5.5.2. Experimental Results on Audio-Visual Speech Recognition on the GRID Corpus. As described earlier, the GRID corpus sentence consists of 6 words. The organization of these words as sentences is as follows

Word 1: bin — lay — place — set;

Word 2: blue — green — red — white;

Word 3: at — by — in — with;

Word 4: a — b — c — d — e — f — g — h — i — j — k — l — m — n — o — p — q — r — s — t — u — v — x — y — z;

Word 5: zero — one — two — three — four — five — six — seven — eight — nine;

Word 6: again — now — please — soon.

In order to use the proposed MHT-SB function for a soft combination of the decisions made from audio and video

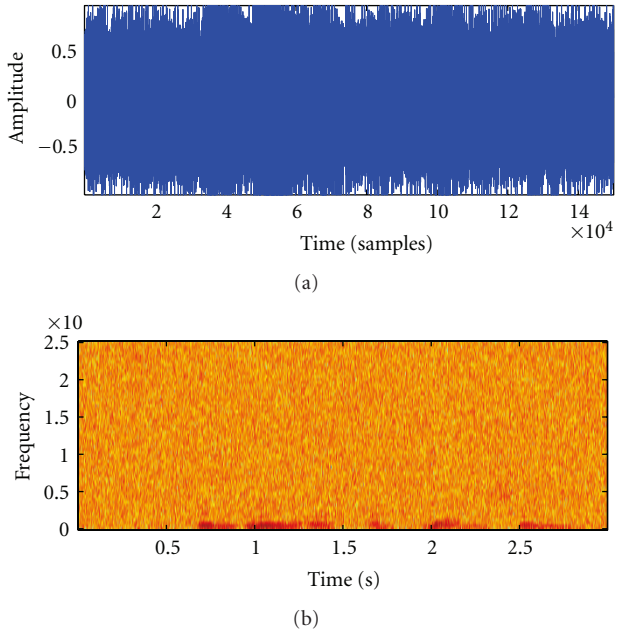

Figure 17: Noisy speech signal and its spectrogram.

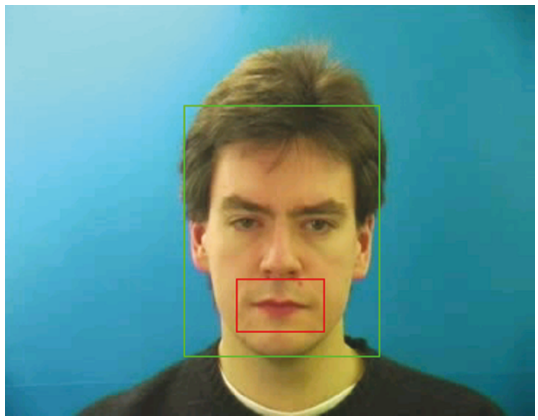

Figure 18: Clean video signal.

Table 8: Percentage word recognition for clean speech.

	Word 1	Word 2	Word 3
Reliability of video	52.33%	43.54%	44.04%
Reliability of audio	98.12%	95.74%	77.08%
Unimodal video features	50.33%	39.05%	40.55%
Unimodal audio features	96.99%	96.49%	79.89%
A-V feature fusion	82.98%	74.44%	57.62%
Coupled HMM	92.98%	94.44%	77.62%
Fusion using MHT-SBF	99.00%	96.63%	80.12%
	Word 4	Word 5	Word 6
Reliability of video	11.31%	21.89%	39.93%
Reliability of audio	72.10%	95.94%	99.94%
Unimodal video features	10.92%	25.24%	43.50%
Unimodal audio features	72.83%	96.65%	99.89%
A-V Feature fusion	48.13%	71.19%	83.48%
Coupled HMM	66.98%	90.44%	92.62%
Fusion using MHT-SBF	74.34%	97.52%	99.89%

Table 9: Percentage word recognition for speech at SNR of 30 dB.

	Word 1	Word 2	Word 3
Reliability of video 52.33%	43.54%	44.04%	
Reliability of audio	81.43%	89.50%	63.45%
Unimodal video features	50.33%	39.05%	40.55%
Unimodal audio features	81.94%	91.47%	64.16%
A-V feature fusion	80.42%	72.30%	54.41%
Coupled HMM	73.98%	85.44%	60.62%
Fusion using MHT-SBF	83.95%	91.97%	64.96%
	Word 4	Word 5	Word 6
Reliability of video	11.31%	21.89%	39.93%
Reliability of audio	59.61%	85.60%	88.64%
Unimodal video features	10.92%	25.24%	43.50%
Unimodal audio features	64.91%	86.36%	90.84%
A-V feature fusion	35.93%	58.49%	56.88%
Coupled HMM	61.98%	79.44%	87.62%
Fusion using MHT-SBF	65.97%	87.28%	92.67%

modalities, the reliability of acoustic and visual speech features is found by carrying out recognition experiments on the development data. This gives the reliability of acoustic and visual speech data. The weighted likelihoods corresponding to acoustic and visual speech features are found using

$$SL_\gamma = \text{antilog}\left(\frac{L}{1000\gamma}\right). \tag{18}$$

In (18), SL_γ is the weighted log likelihood, γ is the weighting factor, and L the original likelihood obtained from the unimodal visual speech feature. The variable γ, represents the weight being given to likelihood obtained from the video modality, while making the combined decision. The values of the log likelihood obtained from the recognizer is small. For audio it is of the order of -300 to -200, whereas for video is of the order of -3000 to -2000. Because of exponential function, for large values of weight, difference in the probabilities of different words is larger for video than audio. So large value of the weighting factor γ, represents more weight being given to video than audio. Out of the total data available 80% of the data is used as training data and remaining 20% as test data.

Recognition is performed on every word of the sentence separately as well as on the whole sentence as continuous speech recognition. Experiments were carried out for four sets of noise conditions at an SNR 40 dB (clean), 30 dB, 20 dB, and 10 dB. Isolated word recognition results for all the noise conditions are given in Tables 8, 9, 10, and 11. Figure 19, illustrates the bar chart of percentage word (letter set) recognition rates for various signal to noise ratios using unimodal video features, unimodal audio features, audio-visual feature fusion, coupled HMM, and the MHT-SB fusion methods. Similar plots are illustrated in Figure 20, for

Table 10: Percentage word recognition for speech at SNR of 20 dB.

	Word 1	Word 2	Word 3
Reliability of video	52.33%	43.54%	44.04%
Reliability of audio	54.72%	78.34%	58.17%
Unimodal video features	50.33%	39.05%	40.55%
Unimodal audio features	53.35%	78.92%	57.00%
A-V feature fusion	56.67%	64.77%	48.97%
Coupled HMM	60.98%	72.44%	52.62%
Fusion using MHT-SBF	64.47%	79.80%	58.35%
	Word 4	Word 5	Word 6
Reliability of video	11.31%	21.89%	39.93%
Reliability of audio	35.91%	60.66%	68.92%
Unimodal video features	10.92%	25.24%	43.50%
Unimodal audio features	37.25%	67.90%	72.25%
A-V feature fusion	26.7%	45.38%	50.06%
Coupled HMM	39.98%	64.44%	70.62%
Fusion using MHT-SBF	42.67%	69.87%	75.91%

Table 11: Percentage word recognition for speech at SNR of 10 dB.

	Word 1	Word 2	Word 3
Reliability of video	52.33%	43.54%	44.04%
Reliability of audio	48.78%	64.12%	48.72%
Unimodal video features	50.33%	39.05%	40.55%
Unimodal audio features	48.95%	57.06%	51.51%
A-V Feature fusion	62.91%	59.61%	44.45%
Coupled HMM	60.98%	56.44%	49.62%
Fusion using MHT-SBF	64.63%	60.38%	52.87%
	Word 4	Word 5	Word 6
Reliability of video	11.31%	21.89%	39.93%
Reliability of audio	18.92%	43.17%	49.70%
Unimodal video features	10.92%	25.24%	43.50%
Unimodal audio Features	17.60%	47.84%	55.67%
A-V feature fusion	17.98%	36.39%	48.01%
Coupled HMM	19.98%	47.44%	55.62%
Fusion using MHT-SBF	21.86%	50.56%	59.45%

digit recognition under various SNR for all the methods used in this work.

5.5.3. Experimental Results As a Function of the Weighting Factor. As described in the previous section, the variable γ, represents the weight being given to likelihood obtained from the video modality, while making the combined decision. In order to analyze the importance of the weight applied to the video modality experiments are performed again for various values of γ. Figures 21, 22, 23, and 24, show graphs of percent word recognition against the weighting factor γ, as described in (18), for different noise conditions.

5.5.4. Discussion on Audio-Visual Speech Recognition System Performance. Speech recognition problem is more challenging than speaker diarization problem because it is a multiple hypothesis problem. Moreover video information for speech

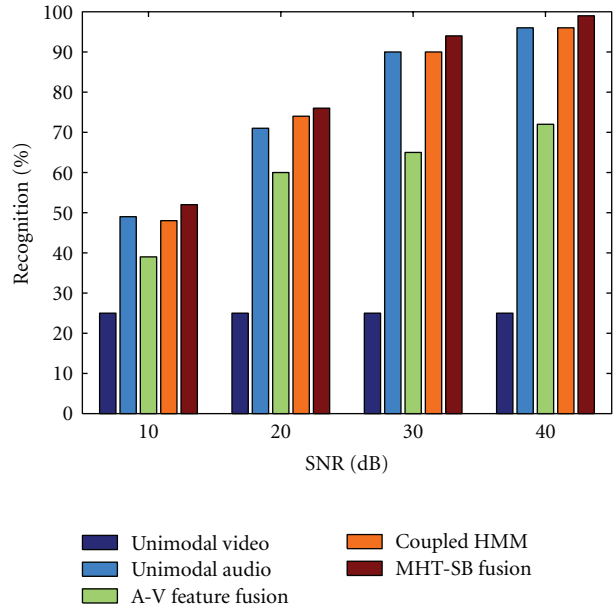

Figure 19: Recognition results for the letter set "A–Z, except W".

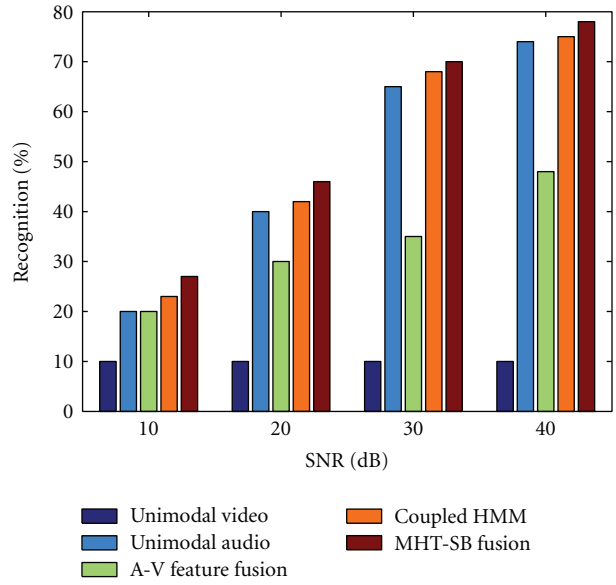

Figure 20: Recognition results for the digit set "zero-nine".

recognition case is even less reliable as many type of sounds can be generated from one type of lip shape and different speakers have different speaking styles. For clean speech, audio provides excellent recognition results but with increase in noise content, audio performance falls drastically. In high noise conditions even this less reliable video information can be quiet helpful in improving recognition results as listed in Tables 8, 9, 10, and 11. This is also illustrated in Figures 19 and 20. The results indicate that the proposed soft fusion method performs reasonably better than the audio only recognition results and recognition results of concatenated audio-visual features. The weight given to video information

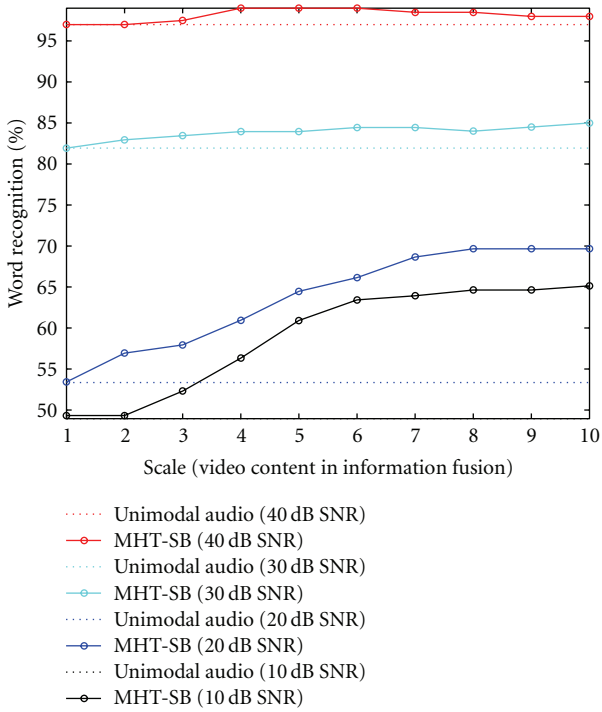

Figure 21: Recognition results for word 1= "bin — lay — place — set" as a function of the weight γ.

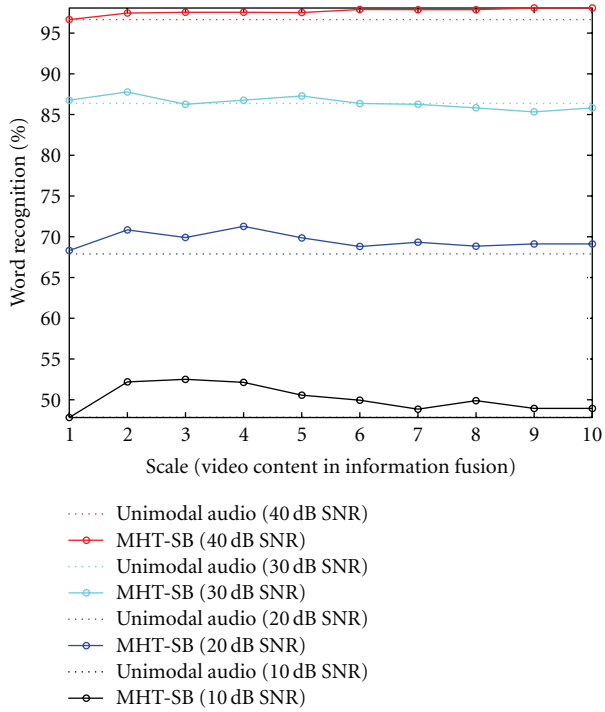

Figure 23: Recognition results for word 5 = "zero-nine" as a function of the weight γ.

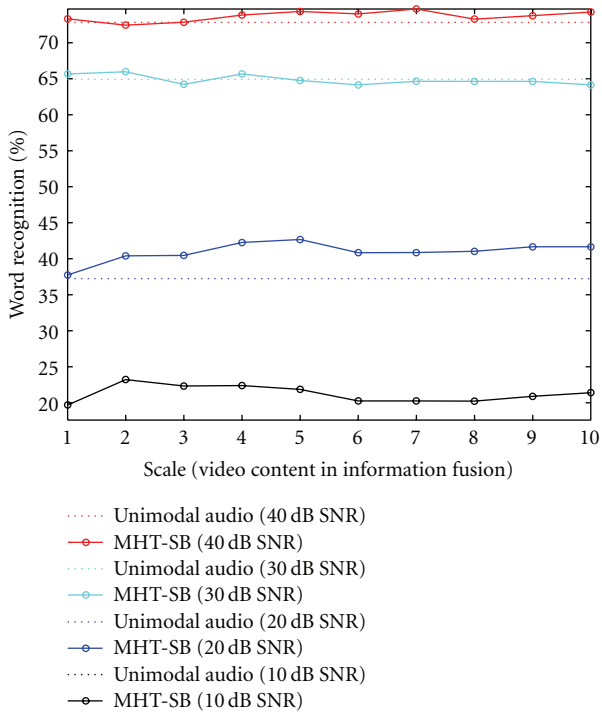

Figure 22: Recognition results for word 4 = "A–Z, except W" as a function of the weight γ.

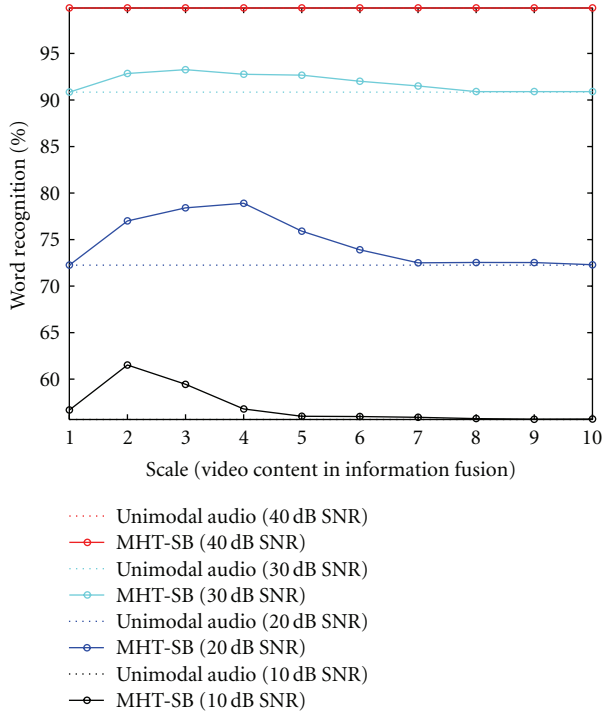

Figure 24: Recognition results for word 6 = "again — now — please — soon" as a function of the weight γ.

alters the recognition results, according to its reliability. For the example of, word 1 = "bin — lay — place — set", the reliability of video information is good (52.33%). In this case, when video information is given a higher weight recognition results improve, especially for noisy conditions, as shown in Figure 21. Similar results illustrating the effect of the weighting factor γ, on word recognition can also be seen in Figures 22, 23, and 24. In general, the weight of the individual modalities can be set based on the reliability obtained for each modality.

6. Conclusions

A methodology to fuse information from multiple modalities using soft belief functions has been proposed for multi-modal speech processing. This method uses a confusion matrix of probability mass functions and combines both the belief and ignorance computed from acoustic and visual speech. As the experimental results show there is a significant improvement in the system performance due to the use of multiple modalities and subsequent soft fusion. This method also provides a framework for soft fusion when compared to the conventional probabilistic fusion framework used in multi-modal speech applications. The results listed in this paper are for a small vocabulary. Hence future work will focus on potential application areas based on small vocabulary recognition, such as assistive driving and assistive living.

Acknowledgment

The work described in this paper was supported by BITCOE and IIT Kanpur under project nos. 20080252, 20080253 and 20080161.

References

[1] M. Gentilucci and L. Cattaneo, "Automatic audiovisual integration in speech perception," *Experimental Brain Research*, vol. 167, no. 1, pp. 66–75, 2005.

[2] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley, New York, NY, USA, 2004.

[3] J.-P. Thiran, F. Marques, and H. Bourlard, *Multi Modal Signal Processing: Theory and Applications for Human-Computer Interaction*, Academic Press, New York, NY, USA, 2010.

[4] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an hmm-based ASR," in *Proceedings of NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, D. Stork and M. Hennecke, Eds., pp. 461–472, 2001.

[5] C. Bregler and Y. Konig, "'eigenlips' for robust speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '94)*, pp. 669–672, 1994.

[6] C. Neti, G. Potamianos, J. Luettin et al., "Audio visual speech recognition, final workshop 2000 report," Tech. Rep., Center for Language and Speech Processing, 2000.

[7] J.-S. Lee and C. H. Park, "Adaptive decision fusion for audio-visual speech recognition," in *Speech Recognition, Technologies and Applications*, pp. 275–296, I-Tech, Vienna, Austria, 2008.

[8] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '02)*, pp. 2013–2016, May 2002.

[9] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton. NJ, USA, 1976.

[10] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 4069–4072, April 2009.

[11] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, 2007.

[12] I. McCowan, J. Carletta, W. Kraaij et al., "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, September 2005.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[14] J. Ajmera, G. Lathoud, and I. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 605–608, May 2004.

[15] M. Visser, M. Poel, and A. Nijholt, "Classifying visemes for automatic lipreading," in *Proceedings of the 2nd International Workshop on Text, Speech and Dialogue*, V. Matousek et al., Ed., vol. 1692 of *Lecture Notes in Computer Science*, p. 843, Plzen, Czech Republic, 1999.

[16] X. Wang, Y. Hao, D. Fu, and C. Yuan, "Audio-visual automatic speech recognition for connected digits," in *Proceedings of the 2nd International Symposium on Intelligent Information Technology Application (IITA '08)*, pp. 328–332, December 2008.

[17] P. Wiggers, J. C. Wojdel, and L. J. M. Rothkrantz, "Medium vocabulary continuous audio-visual speech recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '02)*, 2002.

[18] T. J. Hazen, K. Saenko, C. H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: data collection, development, and initial experiments," in *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04)*, pp. 235–242, October 2004.

[19] L. Liang, X. Liu, Y. Zhao, X. Pi, and A. V. Nefian, "Speaker independent audio-visual continuous speech recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 25–28, 2002.

[20] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1274–1288, 2002.

[21] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1325, 2003.