Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 86 (2006) 3549–3558
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Multimodal speaker/speech recognition using lip motion, lip texture and audio ☆

H.E. Çetingül*, E. Erzin, Y. Yemez, A.M. Tekalp

College of Engineering, Koç University, Sarıyer, Istanbul 34450, Turkey

Received 1 July 2005; received in revised form 5 December 2005; accepted 1 February 2006
Available online 2 June 2006

Abstract

We present a new multimodal speaker/speech recognition system that integrates audio, lip texture and lip motion modalities. Fusion of audio and face texture modalities has been investigated in the literature before. The emphasis of this work is to investigate the benefits of inclusion of lip motion modality for two distinct cases: speaker and speech recognition. The audio modality is represented by the well-known mel-frequency cepstral coefficients (MFCC) along with the first and second derivatives, whereas lip texture modality is represented by the 2D-DCT coefficients of the luminance component within a bounding box about the lip region. In this paper, we employ a new lip motion modality representation based on *discriminative analysis* of the dense motion vectors within the same bounding box for speaker/speech recognition. The fusion of audio, lip texture and lip motion modalities is performed by the so-called *reliability weighted summation* (RWS) decision rule. Experimental results show that inclusion of lip motion modality provides further performance gains over those which are obtained by fusion of audio and lip texture alone, in both speaker identification and isolated word recognition scenarios.

Keywords: Speaker identification; Isolated word recognition; Lip reading; Lip motion; Decision fusion

1. Introduction

Audio is probably the most natural modality to recognize speech content and a valuable source to identify a speaker [1]. Video also contains important biometric information, which includes face/lip texture and lip motion information that is correlated with the audio. Audio-only speaker/speech recognition systems are far from being perfect especially under noisy conditions. Furthermore, it is a known fact that the content of speech can be revealed partially through lip-reading. Performance problems are also observed in video-only speaker/speech recognition systems, where poor picture quality, changes in pose and lighting conditions, and varying facial expressions may have detrimental effects [2,3]. Hence, robust solutions for both speaker and speech recognition should employ multiple modalities, such as audio, lip texture and lip motion in a unified scheme.

The design of a multimodal recognition system requires addressing three basic issues: (i) Which

☆This work has been supported by TÜBİTAK under the project EEEAG-101E026 and by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).
*Corresponding author.
E-mail addresses: ertan@cis.jhu.edu (H.E. Çetingül), eerzin@ku.edu.tr (E. Erzin), yyemez@ku.edu.tr (Y. Yemez), mtekalp@ku.edu.tr (A.M. Tekalp).

0165-1684/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.sigpro.2006.02.045

modalities to fuse; (ii) How to represent each modality with a discriminative and low-dimensional set of features; and (iii) How to fuse existing modalities. Speech content and voice can be interpreted as two different, though correlated, information existing in audio signals. Likewise, video signal can be split into different modalities, such as face/lip texture and lip motion. The second issue, representative feature selection, also includes modeling of classifiers through which each class is represented with a statistical model or a representative feature set. Curse of dimensionality, computational efficiency, robustness, invariance and discrimination capability are the most important criteria in selection of the feature set and the recognition methodology for each modality. As for the final issue, that is, the fusion problem, different strategies are possible: in the so-called "early integration", modalities are fused at data or feature level, whereas in "late integration" decisions or scores resulting from each unimodal recognition are combined to give the final conclusion. Multimodal decision fusion can also be viewed from a broader perspective as a way of combining classifiers, which is a well-studied problem in pattern recognition. The main motivation for multimodal fusion is to compensate possible misclassification errors of a certain classifier with other available classifiers and to end up with a more reliable overall decision. Misclassification errors are in general inevitable due to numerous factors such as environmental noise, measurement and modeling errors or time-varying characteristics of signals. A comprehensive survey and discussion on classifier combination techniques can be found in [4].

State-of-art speech recognition systems have been jointly using lip information with audio [5–9]. For speech recognition, it is usually sufficient to extract the principal components of the lip information and to match the mouth openings–closings with the phonemes of speech. Speaker identification using audio and lip information, on the other hand, has been addressed in only few works such as [10–15]. The main challenge is that the principal components of the lip information are not usually sufficient to discriminate between speakers. Non-principal components are also valuable especially when the objective is to model the biometrics. In the speaker/speech recognition literature, audio is generally modeled by mel-frequency cepstral coefficients (MFCC) [16]. However for lip information, there are several approaches reported in the literature such as texture-based, motion-based, geometry-based and model-based [17]. In texture-based approaches, pure or DCT-domain lip image intensity is used as features [8,11,18]. Motion-based approaches compute motion vectors to represent the lip movement during speaking [10,19]. Geometry-based and model-based approaches, in fact, utilize similar processing methods such as active shape models [20,21], active contours [22,23] or parametric models [24] to segment the lip region. They differ in feature selection such that model-based approaches assign the fitted model parameters as features, while shape features such as lengths of horizontal and vertical lip openings, area, perimeter, pose angle are selected for lip representation in geometry-based approaches. In [10], the lip motion is represented by the full set of DCT coefficients of the dense optical flow vectors computed within the rectangular lip region, and then fused with the face texture and the acoustic features for multimodal speaker identification. However, no discrimination analysis and dimensionality reduction are performed in [10]. The speaker recognition schemes proposed in [10,12,13,25,26] are basically opinion fusion techniques that combine multiple expert decisions through adaptive or non-adaptive weighted summation of scores, whereas in [15,27], fusion is carried out at feature-level by concatenating individual feature vectors so as to exploit the temporal correlations that may exist between audio and video signals. In audio-visual speech recognition [18] concatenates audio and lip data, while in [28] unimodal decisions are combined to obtain the fused result. Furthermore, recent works show the success of multistream HMM structures in speech recognition [7–9,17].

In this study, we use the lip motion features that are extracted by a novel discrimination analysis method [19]. Then we integrate lip texture, lip motion and audio features by the reliability-based decision fusion system reported in [11]. The main contribution of this paper is to investigate the fusion of audio modality with the best lip motion and texture representations for two distinct problems, speaker and speech recognition. In this investigation, the performance gain due to the fusion and the optimal modality selection for speaker and speech recognition problems are also discussed. The audio and lip features are presented in detail in Section 2. In Section 3, we describe the probabilistic framework that we use for the speaker/speech recognition problem, and present the reliability weighted

summation (RWS) rule for decision fusion of the multimodal system. Experimental results are presented and discussed in Section 4, and finally concluding remarks are given in Section 5.

2. Modalities and features

In this paper, audio, lip texture and lip motion are considered as different modalities. The MFCC are used as features for the audio modality. The features for the lip texture modality are 2D-DCT coefficients of the luminance component, and features for the lip motion modality are based on the dense motion vectors within a rectangular box about the lip region. The features for the different modalities are explained in more detail below.

2.1. Features for audio modality

Audio stream is represented with the MFCC, as they yield good discrimination of speech signal. The audio stream is processed over 10 ms frames centered on 25 ms Hamming window for 16 kHz sampled audio signal. Each analysis frame is first multiplied with a Hamming window and transformed to frequency domain using Fast Fourier Transform (FFT). Mel-scaled triangular filter-bank energies are calculated over the square magnitude of the spectrum and represented in logarithmic scale [16]. The resulting MFCC features, c_j, are derived using discrete cosine transform (DCT) over log-scaled filter-bank energies e_i:

$$c_j = \frac{1}{N_M}\sum_{i=1}^{N_M} e_i \cos\left((i-0.5)\frac{j\pi}{N_M}\right), \quad j = 1,2,\ldots,N, \tag{1}$$

where N_M is the number of mel-scaled filter banks and N is the number of MFCC features that are extracted. The MFCC feature vector is defined as $\boldsymbol{C} = [c_1 \;\; c_2 \;\; \cdots \;\; c_N]^{\mathrm{T}}$. The audio feature vector $\boldsymbol{f}_{\mathrm{A}}$ is formed as a collection of MFCC vector $\boldsymbol{C}$ along with the first and second delta MFCCs, $\boldsymbol{f}_{\mathrm{A}} = [\boldsymbol{C}\ \Delta\boldsymbol{C}\ \Delta\Delta\boldsymbol{C}]$. Audio feature extraction is briefly illustrated in Fig. 1a.

2.2. Features for lip texture modality

It has been a common practice to use intensity-based features for the representation of lip texture [8,11]. There are certain advantages and drawbacks of the intensity-based lip features, such as representing texture information as well as shape but being sensitive to illumination changes. Fig. 1b shows the intensity-based DCT feature extraction scheme that we employ in our system. The intensity-based lip features are extracted by applying the Bayesian discrimination technique [19] to the low-indexed 2D-DCT coefficients along the zig-zag scan.

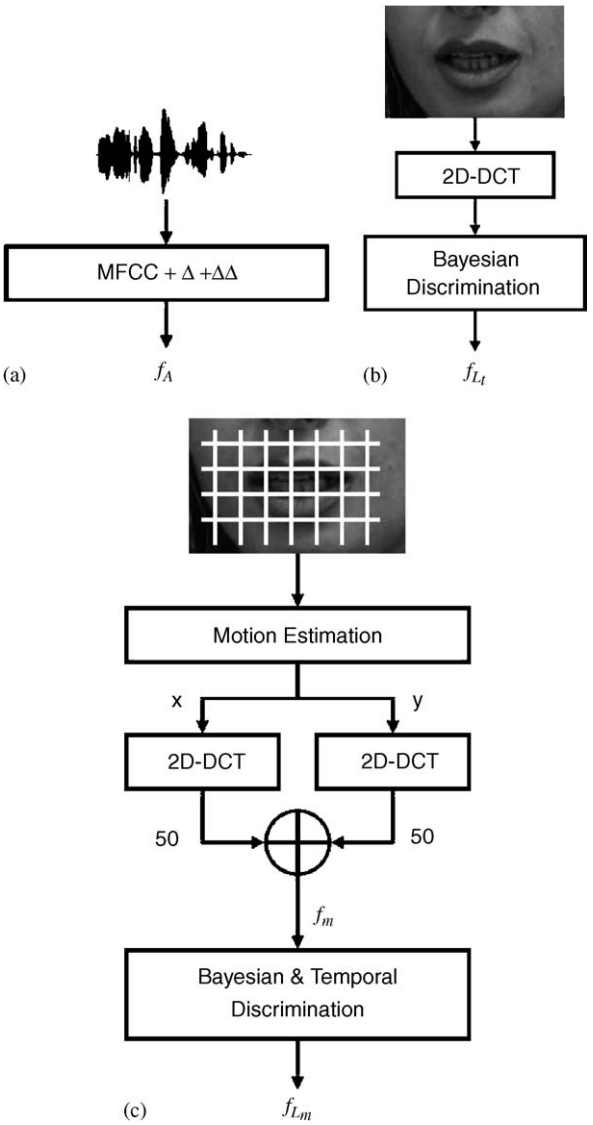

Fig. 1. Block diagrams of the feature extraction for modalities: (a) audio; (b) lip texture; and (c) lip motion.

In the Bayesian discrimination, the DCT coefficients are ordered based on the discrimination measure presented in [19]. Let f_k and $P(f_k|\lambda_i)$ denote the kth component of a feature vector $\boldsymbol{f}$ and its class conditional probability for the class λ_i, respectively. Then the ratio of intra-class and

inter-class probabilities,

$$I(\lambda_i|f_k) = \frac{P(f_k|\lambda_i)P(\lambda_i)}{\sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)}, \quad (2)$$

which appears in the maximum mutual information estimator (MMIE) [29], is employed to define the discrimination content of each feature component. The discrimination content of the kth component of feature $\boldsymbol{f}$, $d(f_k)$, is computed as the average of discriminative ratio over all training repetitions,

$$d(f_k) = \sum_i \frac{1}{M} \sum_{m=0}^{M-1} I(\lambda_i|f_k^i(m)), \quad (3)$$

where $f_k^i(m)$ is the kth feature component of the mth feature vector observation for the class λ_i and M the total number of training repetitions for each class. After finding the discrimination content $d(f_k)$ for each feature component, the most discriminative 50 DCT coefficients are concatenated to form the lip-texture feature vector $\boldsymbol{f}_{L_t}$.

A preprocessing step is required to locate the lip region and eliminate the global motion of the head between the frames so that the extracted motion features within the lip region provides us with the pure movement of the speaking act. To this effect, each face frame is aligned with the first frame of the sequence using a 2D parametric motion estimator. For every two consecutive face images, global head motion parameters are calculated using hierarchical Gaussian image pyramids and 12-parameter quadratic motion model [30]. The face images are successively warped according to these calculated parameters [19]. In the resulting aligned image sequence, the location of the lip region remains almost unchanged except for local movements. Thus, by only hand-labeling the mid-point of the lip region on the first frame, we automatically extract a region of interest around this point so as to obtain a sequence of lip frames of size 128×80.

2.3. Features for lip motion modality

Although lip movement is considered as the primary source for visual speech applications, it is rarely represented by its pure motion features. There are few studies incorporating the pure lip motion as the visual feature [10]. In [10], the lip motion is represented by the full set of 2D-DCT coefficients of the vectors. In this study the best lip motion representation that is found in [19,31] is employed. A brief summary of this representation is presented in the following.

After performing global head motion compensation and lip region extraction as defined in Section 2.2, the use of a dense uniform grid of size 64×40 on the intensity lip image is considered. This grid definition allows us to analyze the whole motion information contained within the rectangular mouth region and it has proven its identification performance [31]. We use hierarchical block matching to estimate the lip motion in quarter-pixel accuracy by interpolating the original lip image with appropriate 6-tap Wiener and bilinear filters as used in H.264/MPEG-4 AVC [32]. The motion estimation procedure yields two 64×40 2D matrices $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$, each of which stores the motion vector components at grid points of the mouth region. The x and y components of the motion vector computed at the grid point (i,j) is given by the (i,j)th entries of $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$, respectively. The motion matrices, $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$, are separately transformed via 2D-DCT. The first 50 DCT coefficients of the zig-zag scan both on x and y directions are combined to form a feature vector of dimension 100.

In [19], we proposed a two-stage discriminative feature selection approach to determine the best lip motion features. It takes into account the temporal discrimination information as well as the intra-class and inter-class distribution of individual single-frame lip feature vectors. At the first stage, we achieve discrimination in the Bayesian sense using a probabilistic measure that maximizes the ratio of intra-class and inter-class probabilities. The most discriminative features among the whole set of 100 features are selected. At the second stage, the successively concatenated lip feature vectors are created as a new sequence of higher dimensional feature vectors, each centered at the current frame instant. Then, they are projected to a lower dimensional feature space using linear discriminant analysis (LDA). The resulting lower dimensional feature vector representing the dense grid motion will be denoted by $\boldsymbol{f}_{L_m}$. Fig. 1c presents a block diagram for the lip motion feature extraction.

3. Multimodal fusion

When more than one information source is available, the fusion of information from different sources can reduce overall uncertainty and increase the robustness of a classification system. Suppose that a different classifier, which employs the

maximum likelihood solution using the class-conditional probabilities $P(\boldsymbol{f}_n|\lambda_r)$, is available for each of the N modalities $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N$. Equivalently, each classifier, say the nth classifier, produces a set of R log-likelihood values $\rho_n(\lambda_r) = \log P(\boldsymbol{f}_n|\lambda_r)$ for each of the R classes $\lambda_1, \lambda_2, \ldots, \lambda_R$. The problem then reduces to compute a single set of joint log-likelihood values $\rho(\lambda_1), \rho(\lambda_2), \ldots, \rho(\lambda_R)$ for these N modalities. In the Bayesian framework, assuming that $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N$ are statistically independent, the joint log-likelihood is given by the sum of the individual log-likelihoods:

$$\rho(\lambda_r) = \log P(\boldsymbol{f}_1|\lambda_r) \cdots P(\boldsymbol{f}_N|\lambda_r) = \sum_n \rho_n(\lambda_r), \qquad (4)$$

which is equivalent to the so-called product rule [4]. In practice, there are a couple of problems with the optimality of this rule. First, the partial decisions coming from different classifiers may be correlated. Second, due to modeling errors and/or measurement noise, the estimated distribution model of training features, i.e., $P(\boldsymbol{f}_n|\lambda_r)$, may not always comply with the actual distribution of the test features. As a result, the log-likelihood values coming from separate classifiers should each be considered as an opinion or a likelihood score rather than a probabilistic value. The statistics and the numerical range of these likelihood scores mostly vary from one classifier to another, and thus using sigmoid or variance normalization, they are often normalized into $(0, 1)$ interval before the fusion process. In this work, we employ the sigmoid normalization described in [11].

In order to cope with the above problems, various approximation approaches have been proposed in the literature as alternatives to the product rule (i.e., the sum rule in log domain) such as max rule, min rule and reliability-based weighted summation. In fact, the most generic way of computing joint ratios (or scores) can be expressed as a weighted summation

$$\rho(\lambda_r) = \sum_{n=1}^{N} \omega_n \rho_n(\lambda_r) \quad \text{for } r = 1, 2, \ldots, R, \qquad (5)$$

where ω_n denotes the weighting coefficient for modality n, such that $\sum_n \omega_n = 1$. Then, the fusion problem becomes finding the optimal weight coefficients. Note that when $\omega_n = 1/N \ \forall n$, (5) is equivalent to the product rule. Since the ω_n values can be regarded as the reliability values of the classifiers, we referred to this combination method as the reliability weighted summation (RWS) rule in [11]. Reliability values ω_n can be set to some fixed values using a priori knowledge about the performance of each modality classifier or can be estimated adaptively for each decision instant via various methods such as those in [11,12,25].

In this work, we employed the RWS rule for the fusion of audio, lip texture and lip motion modalities using the reliability value estimation which is described in Section 3.3.

3.1. Speaker recognition

The recognition task can be formulated as either a verification or an identification problem. The latter can further be classified as open-set or closed-set identification. In the closed-set identification problem, a reject scenario is not defined and an unknown observation is classified as belonging to one of the R registered pattern classes. In the open-set problem, the objective is, given the observation from an unknown pattern, to find whether it belongs to a pattern class registered in the database or not; the system identifies the pattern if there is a match and rejects otherwise. Hence, the problem can be thought of as an $R+1$ class identification problem, including also a reject class. Open-set identification has a variety of applications such as the authorized access control for computer and communication systems, where a registered user can log onto the system with her/his personalized profile and access rights. In this paper, we formulate the speaker recognition problem in an open-set identification framework, which is a more challenging and realistic way of addressing the problem as compared to closed-set speaker identification and verification. Note that verification is a special case of the general open-set identification problem.

In the open-set identification problem, an imposter class λ_{R+1} is introduced as the $(R+1)$th class. Since it is difficult to accurately model the imposter class, λ_{R+1}, we employ the following solution which includes a reject strategy through the definition of the likelihood ratio:

$$\bar{\rho}(\lambda_r) = \log \frac{P(\boldsymbol{f}|\lambda_r)}{P(\boldsymbol{f}|\lambda_{R+1})} = \log P(\boldsymbol{f}|\lambda_r) - \log P(\boldsymbol{f}|\lambda_{R+1}). \qquad (6)$$

Then, the decision strategy of the open-set identification can be implemented in two steps. First, determine

$$\lambda_* = \arg\max_{\lambda_1,\ldots,\lambda_R} \bar{\rho}(\lambda_r) \qquad (7)$$

and then

$$\begin{array}{ll} \text{if } \bar{\rho}(\lambda_*) \geqslant \tau & \text{accept,} \\ \text{otherwise} & \text{reject,} \end{array} \tag{8}$$

where τ is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate.

Computation of class-conditional probabilities needs a prior modeling step, through which a probability density function of feature vectors is estimated for each class $r = 1, 2, \ldots, R$ by using available training data. A common and effective approach to model the impostor class is to use a universal background model, which is estimated by using all available training data regardless of which class they belong to.

3.2. Speech recognition

The speech recognition problem can be formulated so as to identify a specific utterance, such as in the isolated word recognition task. Therefore, the closed-set identification framework can be used to address the speech recognition problem with an isolated word dictionary.

We address the closed-set identification problem within the maximum likelihood framework that maximizes the class-conditional probability, $P(\boldsymbol{f}|\lambda_r)$, for $r = 1, \ldots, R$. Hence a decision in the closed-set identification is taken as

$$\lambda_* = \arg\max_{\lambda_1,\ldots,\lambda_R} \log P(\boldsymbol{f}|\lambda_r) = \arg\max_{\lambda_1,\ldots,\lambda_R} \rho(\lambda_r). \tag{9}$$

3.3. The reliability estimation for the RWS

Among various reliability estimation techniques existing in the literature, we favor the one proposed in [11], since it is better suited to the open-set speaker identification problem by assessing both accept and reject decisions of a classifier, and it can also be easily used for the closed-set identification problem.

The RWS rule combines the likelihood ratio values of the N modalities weighted by their reliability values ω_n as in (5). The reliability value ω_n is estimated based on the difference of likelihood ratios of the best two candidate classes λ_* and λ_{**}, that is, $\Delta_n = \rho_n(\lambda_*) - \rho_n(\lambda_{**})$, for modality n. In the absence of reject class, that is for closed-set identification, the likelihood difference of the best two candidates, Δ_n, can be used as the reliability value. However, in the presence of a reject class, one would expect a high-likelihood ratio $\rho_n(\lambda_*)$ and a high Δ_n value for true accept decisions, and a low-likelihood ratio $\rho_n(\lambda_*)$ and a low Δ_n value for true reject decisions. Hence, a normalized reliability measure ω_n can be estimated by

$$\omega_n = \frac{1}{\sum_i \gamma_i}\,\gamma_n, \tag{10}$$

where

$$\gamma_n = \begin{cases} \Delta_n & \text{closed-set,} \\ (e^{(\rho_n(\lambda_*)+\Delta_n)} - 1) & \\ \quad +(e^{(\kappa-\rho_n(\lambda_*)-\Delta_n)} - 1) & \text{open-set.} \end{cases} \tag{11}$$

The first and second terms for open-set identification in γ_n are associated with the true accept and true reject, respectively. The symbol κ stands for an experimentally determined factor to reach the best compromise between accept and reject scenarios. The κ value is set to 0.65 as it is found to be optimal for the open-set speaker identification task in [11].

4. Experimental results

Hidden Markov models (HMM) are known to be as effective structures to model the temporal behavior of the speech signal, and thus they are widely used both in audio-based speaker identification and speech recognition applications [1]. The speaker identification problem can be classified as text-dependent and text-independent depending on the audio content. In the text-independent problem, identification is performed over a content free utterance of the speakers, whereas in the text-dependent case, each speaker is expected to utter a personalized secret phrase for the identification task. State-of-the-art systems use HMMs for text-dependent and Gaussian mixture models (GMM) for text-independent speaker identification [33]. HMM-based techniques are preferred in text-dependent scenarios since HMM structures can successfully exploit the temporal correlations of a speech signal. Since lip motion is strongly coupled with audio utterance, HMMs can also be employed for temporal characterization of lip features. Hence, class-conditional probabilities of both audio and lip features are modeled and estimated using HMM architectures in our experiments.

In this work, we consider a text-dependent scenario for the speaker recognition problem and address it in the open-set identification framework,

whereas for the speech recognition problem, the closed-set identification framework is employed. We use word-level continuous-density HMM structures for both speaker identification and speech recognition tasks. Each speaker or utterance in the database is modeled using a separate HMM that is trained over some repetitions of the lip-motion streams of the corresponding speaker or utterance. In the recognition process, given a test feature set, each HMM structure associated with a speaker or an utterance produces a likelihood. In the speaker identification case, a world HMM model is also trained over the whole training data of the population. The log-ratio of the speaker likelihoods to the world class likelihood results in a stream of log-likelihood ratios that are used in the speaker identification process. The system identifies the person if there is a match and rejects otherwise. In speech recognition, the impostor or world class is not defined; thus the best match is given by the utterance class that maximizes the produced likelihood as described in Section 3.2.

The performance of speaker verification systems is often measured using the equal error rate (EER) figure. The EER is calculated as the operating point where false accept rate (FAR) equals false reject rate (FRR). In the open-set identification case, FAR and FRR can be defined as

$$\mathrm{FAR} = 100 \times \frac{F_{\mathrm{a}}}{N_{\mathrm{a}} + N_{\mathrm{r}}} \quad \text{and} \quad \mathrm{FRR} = 100 \times \frac{F_{\mathrm{r}}}{N_{\mathrm{a}}}, \tag{12}$$

where F_{a} and F_{r} are the number of false accepts and rejects, and N_{a} and N_{r} are the total number of trials for the true and imposter clients in the testing, respectively. The performance of speech recognition systems, on the other hand, is usually measured with the recognition rate, that is, the ratio of the true matches to the total number of trials.

The speaker and speech recognition experiments have been conducted using the MVGL-AVD audio-visual database [34]. The database includes 50 subjects and considers two distinct text-dependent speaker identification scenarios, which are the name ($\mathscr{D}_n$) and the digit ($\mathscr{D}_d$) scenarios. In the name scenario, each subject utters 10 repetitions of her/his name as the secret phrase. A set of impostor data is also collected with each subject in the population uttering five different names from the population. In the digit scenario, each subject utters 10 repetitions of a fixed digit password 348 572. Although we have a limited variation in the name scenario, each name is considered as an isolated word, and a subset of the name scenario, $\mathscr{D}_s \subset \mathscr{D}_n$, which includes each name utterance with more than 12 repetitions, is considered as the testbed of the speech recognition experiments.

The audio recordings are perturbed with varying levels of additive noise during the testing sessions to simulate adverse environmental conditions. The additive acoustic noise is picked to be a mixture of office and babble noise. Abbreviations and descriptions for the modalities and fusion techniques are given in Table 1.

4.1. Speaker recognition: name scenario

The $\mathscr{D}_n$ database is partitioned into two sets namely $\{\mathscr{D}_{n_A}$ and $\mathscr{D}_{\bar{n}_A}\}$, where $\mathscr{D}_{n_A}$ and $\mathscr{D}_{\bar{n}_A}$ are mutually exclusive sets each having five repetitions from each subject in the database. The subsets $\mathscr{D}_{n_A}$ and $\mathscr{D}_{\bar{n}_A}$ are used for training and testing, respectively. Since there are 50 subjects and five repetitions for each true and imposter client tests, the resulting total number of trials for the true accepts and true rejects become, respectively, $N_{\mathrm{a}} = 250$ and $N_{\mathrm{r}} = 250$.

Table 2 presents the EER performances of the unimodal and multimodal open-set speaker identification systems with audio, lip texture and lip motion modalities. The EER performances of the lip texture and lip motion modalities are 5.6% and 5.2%, which are close to each other and better than the audio modality at 15 dB SNR and below. When the product rule and the RWS rule are applied to fuse a pair of modalities or all the three modalities, the EER performance increases significantly. The RWS rule is observed to perform better than the product rule, especially under noisy conditions. The best EER performance is achieved with the fusion of all three modalities at 15 dB SNR and below. Above 15 dB SNR, the best performance is achieved with the fusion of lip texture and audio modalities.

Table 1
Abbreviations and descriptions for modalities and fusion techniques

A	Audio modality
L_{t}	Lip texture modality
L_{m}	Lip motion modality
$+$	Product rule
$\oplus$	RWS rule

Table 2
Speaker identification results for name scenario: equal error rates at varying noise levels for different modalities and multimodal fusion structures

Source modality	Noise level (dB SNR)						
	Clean	25	20	15	10	7	5
EER (%)							
A	1.0	1.6	2.4	5.3	14.8	25.4	31.5
L_t	5.6						
L_m	5.2						
$L_m + A$	2.6	3.2	3.6	4.4	7.2	17.5	22.8
$L_m \oplus A$	0.8	1.2	1.8	3.2	5.6	13.6	19.2
$L_t + A$	0.4	0.4	0.8	2.0	4.4	11.2	15.9
$L_t \oplus A$	1.0	0.8	1.0	1.8	3.0	6.8	9.6
$L_m + L_t + A$	1.6	1.4	1.4	1.4	1.74	3.6	4.4
$L_m \oplus L_t \oplus A$	1.2	1.2	1.2	1.2	1.4	3.2	3.2

Table 3
Speaker identification results for digit scenario: equal error rates at varying noise levels for different modalities and multimodal fusion structures

Source modality	Noise level (dB SNR)						
	Clean	25	20	15	10	7	5
EER (%)							
A	2.4	3.4	6.9	12.2	24.9	33.1	37.1
L_t	1.7						
L_m	5.2						
$L_m + A$	2.4	2.4	2.4	4.0	10.4	18.0	23.2
$L_m \oplus A$	2.4	2.4	2.4	4.0	10.0	16.8	22.0
$L_t + A$	0.4	0.4	0.4	1.4	6.8	14.0	18.4
$L_t \oplus A$	0.4	0.4	0.4	0.8	4.0	10.0	13.8
$L_m + L_t + A$	0.8	0.8	1.2	1.2	2.6	4.2	5.2
$L_m \oplus L_t \oplus A$	0.4	0.4	0.6	0.8	2.4	3.8	5.2

4.2. Speaker recognition: digit scenario

The $\mathscr{D}_d$ database is partitioned into two sets $\{\mathscr{D}_{d_A}$ and $\mathscr{D}_{\bar{d}_A}\}$, where $\mathscr{D}_{d_A}$ and $\mathscr{D}_{\bar{d}_A}$ are mutually exclusive sets each having five repetitions of the same 6-digit number from each subject in the database. The subsets $\mathscr{D}_{d_A}$ and $\mathscr{D}_{\bar{d}_A}$ are used for training and testing, respectively. Note that, in the digit scenario no imposter recordings are performed since every subject utters the same 6-digit number. Hence, the imposter clients are generated by the *leave-one-out* scheme, where each subject, let us denote her/him by S, becomes the imposter of the remaining $R-1$ subjects in the population. Since, the class S is out of the population during the imposter tests, every test utterance that belongs to S becomes an imposter test. Having $R = 50$ subjects and five testing repetitions the resulting total number of trials for the true accepts and true rejects (imposters) become, respectively, $N_a = 250$ and $N_r = 250$.

Table 3 presents the EER performances of the unimodal and multimodal open-set speaker identification systems with audio, lip texture and lip motion modalities. The EER performances of the lip texture and lip motion modalities are 1.7% and 5.2%. Since every subject utters the same six digit password in the digit scenario, the discrimination of true and imposter clients is poor for audio modality. However, this discrimination is better for the lip texture modality, since true and imposter clients carry different lip textures. When the product rule and the RWS rule are applied to fuse a pair of modalities or all the three modalities, the EER performance increases significantly. The RWS rule is observed to perform better than the product rule at all SNR conditions. The best EER performance is achieved with the fusion of all three modalities at all SNR levels.

4.3. Speech recognition

In this scenario, the database $\mathscr{D}_s$ includes 35 different phrases (isolated words) where each phrase is actually names of the subjects in the database and repeated at least 12 times. The $\mathscr{D}_s$ database is partitioned into two sets $\mathscr{D}_{s_A}$ and $\mathscr{D}_{\bar{s}_A}$, where they are mutually exclusive sets each having equal number of utterance repetitions. The subsets $\mathscr{D}_{s_A}$ and $\mathscr{D}_{\bar{s}_A}$ are used for training and testing, respectively.

Table 4 presents the recognition performances of the unimodal and multimodal speech recognition systems with audio, lip texture and lip motion modalities. The recognition performances of the lip texture and lip motion modalities are 62.86% and 72.86%. The recognition rate of the lip texture modality is poorer than the lip motion modality. This is as expected since motion information is more important than texture in lip reading. When the product rule and the RWS rule are applied to fuse a pair of modalities or all the three modalities, the recognition performance increases when the lip texture modality is not included in the fusion. The best recognition performance is achieved with the RWS fusion of audio and lip motion modalities at all SNR levels.

Table 4
Speech recognition results: recognition rates at varying noise levels for different modalities and multimodal fusion structures

Source modality	Noise level (dB SNR)						
	Clean	25	20	15	10	7	5
Recognition (%)							
A	90.00	88.57	87.62	86.67	80	62.86	39.05
L_t	62.86						
L_m	72.86						
$L_m + A$	86.19	84.28	84.28	84.28	80.95	72.38	63.33
$L_m \oplus A$	91.43	90.95	88.57	88.10	84.76	75.71	69.05
$L_t + A$	76.67	77.14	78.57	76.67	76.19	69.04	54.76
$L_t \oplus A$	76.67	77.14	75.24	74.76	73.33	68.57	61.69
$L_m + L_t + A$	80.95	81.42	80.95	81.90	79.04	75.71	69.52
$L_m \oplus L_t \oplus A$	78.57	78.57	76.19	77.14	74.28	72.38	68.10

5. Conclusions

A multimodal speaker/speech recognition system that integrates audio, lip texture and lip motion modalities has been investigated, where the lip motion modality is represented by the dense-motion-based features within a rectangular grid. We emphasize that the lip motion modality carries additional useful information over that is present in the lip texture modality for both speaker and speech recognition applications. Hence, the fusion of lip motion with audio and lip texture modalities is observed to provide additional performance gains. Furthermore, the lip motion is found to be more valuable than the lip texture modality for speech recognition. The fusion of audio, lip texture and lip motion modalities is performed by the so-called *reliability weighted summation* (RWS) decision rule, which is observed to perform better than the product rule.

References

[1] J. Campbell, Speaker recognition: a tutorial, Proc. IEEE 85 (9) (1997) 1437–1462.
[2] Y.Y.J. Zhang, M. Lades, Face recognition: eigenface, elastic matching, and neural nets, Proc. IEEE 85 (9) (1997) 1423–1435.
[3] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 586–591.
[4] J. Kittler, M. Hatef, R. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Machine Intell. 20 (3) (1998) 226–239.
[5] I. Matthews, T. Cootes, J. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Machine Intell. 24 (2) (2002) 198–213.
[6] C. Chibelushi, F. Deravi, J. Mason, A review of speech-based bimodal recognition, IEEE Trans. Multimedia 4 (1) (2002) 23–37.
[7] X. Zhang, C. Broun, R. Mersereau, M. Clements, Automatic speechreading with applications to human–computer interfaces, EURASIP J. Appl. Signal Process. (2002) 1228–1247.
[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, A. Senior, Recent advances in the automatic recognition of audio-visual speech, Proc. IEEE 91(9).
[9] J. Perez, A. Frangi, E. Solano, K. Lukas, Lip reading for robust speech recognition on embedded devices, Proceedings of the International Conference on Acoustics, Speech and Signal Processing 2005 (ICASSP '05), vol. I, 2005, pp. 473–476.
[10] R. Frischholz, U. Dieckmann, BioID: a multimodal biometric identification system, J. IEEE Comput. 33 (2) (2000) 64–68.
[11] E. Erzin, Y. Yemez, A. Tekalp, Multimodal speaker identification using an adaptive classifier cascade based on modality reliability, IEEE Trans. Multimedia 7 (5) (2005) 840–852.
[12] T. Wark, S. Sridharan, Adaptive fusion of speech and lip information for robust speaker identification, Digital Signal Process. 11 (3) (2001) 169–186.
[13] P. Jourlin, J. Luettin, D. Genoud, H. Wassner, Acoustic-labial speaker verification, Pattern Recognition Lett. 18 (9) (1997) 853–858.
[14] L. Mok, W. Lau, S. Leung, S. Wang, H. Yan, Lip features selection with application to person authentication, Proceedings of the International Conference on Acoustics, Speech and Signal Processing 2004 (ICASSP '04), vol. III, 2004, pp. 397–400.
[15] M. Civanlar, T. Chen, Password-free network security through joint use of audio and video, Proc. SPIE Photonic (1996) 120–125.
[16] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[17] S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition, IEEE Trans. Multimedia 2 (3) (2000) 141–151.
[18] C. Bregler, Y. Konig, Eigenlips for robust speech recognition, Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing, 1994, pp. 669–672.
[19] H. Cetingul, Y. Yemez, E. Erzin, A. Tekalp, Discriminative lip-motion features for biometric speaker identification, Proceedings of the International Conference on Image Processing 2004 (ICIP 2004), 2004, pp. 2023–2026.
[20] S. Lucey, S. Sridharan, V. Chandran, Initialised eigenlip estimator for fast lip tracking using linear regression, Proceedings of the 15th International Conference on Pattern Recognition 2000, vol. 3, 2000, pp. 178–181.
[21] S. Wang, W. Lau, S. Leung, H. Yan, A real-time automatic lipreading system, Proceedings of the 2004 International Symposium on Circuits and Systems (ISCAS 2004), vol. 2, 2004, pp. 101–104.
[22] T. Wakasugi, M. Nishiura, K. Fukui, Robust lip contour extraction using separability of multi-dimensional distributions, Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04), 2004, pp. 415–420.

[23] P. Aleksic, J. Williams, Z. Wu, A. Katsaggelos, Audio-visual speech recognition using MPEG-4 compliant visual features, EURASIP J. Appl. Signal Process. (2002) 1213–1227.
[24] N. Eveno, A. Caplier, P.-Y. Coulon, Accurate and quasi-automatic lip tracking, IEEE Trans. Circuits Systems Video Technol. 14 (5) (2004) 706–715.
[25] C. Sanderson, K. Paliwal, Noise compensation in a person verification system using face and multiple speech features, Pattern Recognition 36 (2) (2003) 293–302.
[26] R. Brunelli, D. Falavigna, Person identification using multiple clues, IEEE Trans. Pattern Anal. Machine Intell. 17 (1995) 955–966.
[27] U. Chaudhari, G. Ramaswamy, G. Potamianos, C. Neti, Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction, Proceedings of the International Conference on Multimedia & Expo 2003 (ICME2003), vol. 3, 2003, pp. 9–12.
[28] D. Zhang, Automated Biometrics, Kluwer Academic Publishers, Dordrecht, 2000.
[29] X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice-Hall, Englewood Cliffs, NJ, 2001.
[30] J.-M. Odobez, P. Bouthemy, Robust multiresolution estimation of parametric motion models, J. Visual Comm. Image Representation 6 (4) (1995) 348–365.
[31] H. Cetingul, Y. Yemez, E. Erzin, A. Tekalp, Robust lip-motion features for speaker identification, Proceedings of the International Conference on Acoustics, Speech and Signal Processing 2005 (ICASSP '05), vol. I, 2005, pp. 509–512.
[32] A. Puri, X. Chen, A. Luthra, Video coding using the H.264/MPEG-4 AVC compression standard, Signal Processing: Image Communications 19 (2004) 793–849.
[33] D. Reynolds, Speaker identification and verification using gaussian mixture speaker models, Speech Comm. 17 (1995) 91–108.
[34] E. Erzin, Y. Yemez, A. Tekalp, Joint audio-video processing for robust biometric speaker identification in car, in: DSP in Mobile and Vehicular Systems, Springer, Berlin, 2005.