Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Reliability

Engin Erzin, *Member, IEEE*, Yücel Yemez, *Member, IEEE*, and A. Murat Tekalp, *Fellow, IEEE*

***Abstract*—We present a multimodal open-set speaker identification system that integrates information coming from audio, face and lip motion modalities. For fusion of multiple modalities, we propose a new adaptive cascade rule that favors reliable modality combinations through a cascade of classifiers. The order of the classifiers in the cascade is adaptively determined based on the reliability of each modality combination. A novel reliability measure, that genuinely fits to the open-set speaker identification problem, is also proposed to assess accept or reject decisions of a classifier. A formal framework is developed based on probability of correct decision for analytical comparison of the proposed adaptive rule with other classifier combination rules. The proposed adaptive rule is more robust in the presence of unreliable modalities, and outperforms the hard-level max rule and soft-level weighted summation rule, provided that the employed reliability measure is effective in assessment of classifier decisions. Experimental results that support this assertion are provided.**

***Index Terms*—Classifier combining, modality reliability, multimodal speaker identification.**

I. Introduction

ALTHOUGH performances of different biometric technologies for speaker identification have been extensively studied individually, there is relatively little work reported in the literature on the fusion of various biometric technologies [1]. Audio is probably the most natural modality to identify a speaker. However, video also contains important biometric information, which includes still frames of face and temporal lip motion information that is correlated with the audio. Most speaker identification systems rely on audio-only data [2]. However, especially under noisy conditions, such systems are far from being perfect for high security applications. The same observation is also valid for systems using only visual data; where poor picture quality, changes in pose and lighting conditions or varying facial expressions may significantly degrade performance [3], [4]. Hence, a robust and precise solution should employ all available sources of information in a unified scheme.

The general speaker identification problem can be formulated as either an open-set or a closed-set identification problem. In the closed-set problem, a reject scenario is not defined and an unknown speaker is classified as one of the N registered people. In the open-set problem, the objective is, given the data of an unknown person, to find whether the person is registered in the database or not; the system identifies the person if there is a match and rejects otherwise. Hence, the problem can be thought of as an $N + 1$ 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. We can further classify the speaker identification problem 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 problem, each speaker is expected to utter a personalized secret phrase for the identification task. Particular attention needs to be paid in the latter case to handle impostor identity claims and the system has to be robust against unauthorized attempts to use the secret phrase of a registered speaker.

The design of a multimodal identification system requires addressing three basic issues. The first one is to decide which modalities to fuse. The word "modality" can be interpreted in various ways; in speaker identification it usually refers to a specific type of information that can be deduced from biometric signals. In this sense, speech, i.e. the content, and voice can be interpreted as two different, though correlated, modalities existing in audio signals. Likewise, video signal can be split into different modalities, face and motion being the major ones. The dominant modality in the motion of a speaking person is naturally the lip movement which is highly correlated with audio whereas gesture (or gait) could also be interpreted as a separate but less significant modality for speaker identification. The second issue is how to represent the raw biometric data for each modality with a discriminative and low-dimensional set of features and, in conjunction with this, to find the best matching metric in the resulting feature space for classification. This step also includes a training phase 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 classification methodology for each modality. The third issue is how to fuse different biometric signals. 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 classification are combined to give the final conclusion [5], [6]. This latter strategy is also referred to as decision or opinion fusion and is effective especially

Manuscript received March 3, 2004; revised August 18, 2004. This work was supported by TUBITAK under project EEEAG-101E038 and by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc). The associate editor coordinating the review of this paper and approving it for publication was Dr. John R. Smith.

The authors are with the Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, Sariyer, Istanbul 34450, Turkey (e-mail: eerzin@ku.edu.tr; yyemez@ku.edu.tr; mtekalp@ku.edu.tr).

Digital Object Identifier 10.1109/TMM.2005.854464

in case the contributing modalities are uncorrelated and thus the resulting partial decisions are statistically independent. Early integration techniques on the other hand, if adequately used, can be favored if a couple of modalities is highly correlated as in the fusion of audio and lip movement. 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 here 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 [7].

Multimodal speaker recognition systems existing in the literature are mostly bimodal, in the sense that they integrate multiple features from audio and face information as in [8]–[13] or from audio and lip information as in [14]–[16]. There are recent efforts to build multimodal databases, such as [17], which will provide valuable resources for the multimodal person recognition systems. The speaker identification and/or verification schemes proposed in [8], [11], [13]–[15], [18] are basically opinion fusion techniques that combine multiple expert decisions through adaptive or nonadaptive weighted summation of scores, whereas in [12] and [16], 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. The concurrent works [9], [10] use decision-level fusion for the verification problem, where scores resulting from each classifier are concatenated to form a feature vector, which is then fed into another classifier, e.g. a median radial basis function (MRBF) network as in [10] or support vector machines and Bayesian classifier as in [9]. The only work in the literature that addresses a multimodal speaker identification system using audio, face and lip motion at the same time is the one presented in [18]. In [18], the lip motion is represented by DCT coefficients of the optical flow vectors computed from lip frames extracted from the video signal. Face and lip features are stored as biometric templates and classified through a set of algorithms, so-called synergetic computer. The acoustic information on the other hand is represented by cepstral coefficients that are then classified by vector quantization using a minimum distance classifier.

In this paper, we present a new multimodality fusion strategy where some of the modalities might be corrupted by measurement noise and/or modeling errors. The basic idea is that a single highly reliable modality alone may sometimes yield a correct decision, whereas its linear fusion with some other less reliable modality may give incorrect results. On other occasions, results obtained by fusion of two modalities may outperform those obtained from each modality alone. Hence, our proposed scheme considers all possible linearly fused modality combinations (including single modalities) with their corresponding reliability measures, and aims at maximizing the benefit of multimodal fusion so that the upper bound for the system error rate becomes the expected occurrence rate of the cases where all classifier combinations fail. Thus, a critical feature of our system is to be able to adaptively assess each modality classifier with a reliability measure. There exist different approaches to measure reliability, such as taking into account statistical dispersion of scores [14], score rank correlation [11], time-varying stream reliability prediction by matching the test data to models, and predictability of the score stream [13] or noise level of the input signal [8]. The common way of incorporating these reliability values into decision fusion is to use them as weighting coefficients and to compute a weighted average of classifier output scores. When the output scores correspond to some probability or likelihood values in logarithmic domain, one can argue that the geometric average is the optimal fusion strategy in the Bayesian sense, given that the classifier decisions are statistically independent and free of modeling and measurement errors [19]. However, the optimality becomes questionable when the geometric average is weighted with reliability values to take errors into account. Weighting by reliability approach is usually formalized by using possibility and fuzzy set theory [20], trying to approximate probabilistic likelihoods from classifier opinions; there is in fact no formal justification that this strategy will probabilistically produce a minimum error classifier. In this work, we regard reliability as a means of giving priority to some single or combined modality in the fusion process, rather than using it as a numerical weight.

The main contributions of this paper to the multimodal speaker identification problem are the following.

i) We propose a new adaptive cascade rule for fusion of multiple modalities. The proposed rule uses an ordered cascade of classifiers each of which corresponds to a single modality or a linearly fused combination of modalities. The order of classifiers in the cascade is based on the estimated reliability of each modality, such that the goal is not to fail whenever at least one of the classifiers gives the correct accept or reject decision.
ii) We propose a new reliability measure to assess decisions of a classifier under both reject and accept scenarios. We also develop criteria that a good reliability measure has to meet.
iii) We develop a formal framework based on probability of correct decision for comparison of different classifier combination rules. We show analytically that the proposed rule outperforms the hard-level max rule and the soft-level weighted summation rule, provided that the employed reliability measure is effective enough in assessment of classifier decisions.
iv) The proposed multimodal speaker recognition system addresses the text-dependent open-set identification problem where the presence of a reject class necessitates specific considerations for the fusion process. Only a few multimodal speaker identification schemes proposed in the literature address the open-set identification problem, such as [11].

In Section II, we describe the probabilistic framework that we use for the open-set speaker identification problem, and discuss some critical problems with the fusion strategies that are commonly used in the literature. We present the proposed adaptive cascade rule for multimodal fusion in Section III together

with our reliability measure. We also introduce an error analysis scheme for comparison of the proposed rule with others. The details of the error analysis are provided in the Appendix. We describe the feature extraction and unimodal classification techniques that we use separately for audio, lip, fused audio-lip and face modalities in Section IV. Experimental results are presented and discussed in Section V, and finally concluding remarks are given in Section VI.

II. Theoretical Framework

A. Unimodal Open-Set Identification

The speaker identification problem is often formalized within a probabilistic framework. The maximum *a posteriori* probability solution to the N-person open set problem requires computing $P(\lambda_n|\boldsymbol{f})$ for each class λ_n, $n = 1,\ldots,N+1$, given a feature vector $\boldsymbol{f}$ representing the sample data of an unknown individual. Alternatively, we can employ the maximum likelihood solution, which maximizes the class-conditional probability, $P(\boldsymbol{f}|\lambda_n)$, for $n = 1,\ldots,N+1$. Since it is difficult to accurately model the imposter class, λ_{N+1}, we employ the following solution which includes a reject strategy through the definition of the likelihood ratio

$$\rho(\lambda_n) = \log \frac{P(\boldsymbol{f}|\lambda_n)}{P(\boldsymbol{f}|\lambda_{N+1})} = \log P(\boldsymbol{f}|\lambda_n) - \log P(\boldsymbol{f}|\lambda_{N+1}). \tag{1}$$

The decision strategy can then be implemented in two steps. First, determine

$$\lambda_* = \underset{\lambda_1,\ldots,\lambda_N}{\arg\max}\, \rho(\lambda_n) \tag{2}$$

and then

$$\begin{array}{ll} \text{if } \rho(\lambda_*) \geq \tau, & \text{accept} \\ \text{otherwise,} & \text{reject} \end{array} \tag{3}$$

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 $n = 1,\ldots,N$ 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.

B. Multimodal Decision 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 P different classifiers, one for each of the P modalities $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_P$, are available. As described above, each classifier, say the pth classifier, produces a set of N-class log-likelihood ratios $\rho_p(\lambda_n)$, $n = 1,\ldots,N$. The problem then reduces to computing a single set of joint log-likelihood ratios $\rho(\lambda_1), \rho(\lambda_2), \ldots, \rho(\lambda_N)$ for these P modalities. In the Bayesian framework, assuming that $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_P$ are statistically independent, the joint log likelihood ratio is given by the sum of the individual ratios

$$\rho(\lambda_n) = \log \frac{P(\boldsymbol{f}_1|\lambda_n)\cdots P(\boldsymbol{f}_P|\lambda_n)}{P(\boldsymbol{f}_1|\lambda_{N+1})\cdots P(\boldsymbol{f}_P|\lambda_{N+1})} = \sum_p \rho_p(\lambda_n) \tag{4}$$

which is equivalent to the so-called product rule [7]. In practice, there are three main problems with the optimality of this rule. First, 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., $P(\boldsymbol{f}_p|\lambda_n)$, may not always comply with the actual distribution of test features. Third, the impostor model, i.e., $P(\boldsymbol{f}_p|\lambda_{N+1})$, is a mere approximation of the reality. As a result, the log likelihood ratios 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 they need to be normalized into the interval (0,1) before the fusion process, using methods such as sigmoid and variance normalization. Unfortunately there is no formally "correct" or optimal way of normalization, which is investigated in detail in [21]. In this paper, a sigmoid normalization is used as in [8], which maps likelihood ratios to the (0,1) interval by normalizing the likelihood ratio ρ using the function

$$g(\rho) = \left[1 + e^{-\left(\frac{\rho-\mu}{2\sigma}+1\right)}\right]^{-1} \tag{5}$$

where μ and σ are the mean and the standard deviation of the likelihood ratio ρ over the accept subjects, respectively.

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

$$\rho(\lambda_n) = \sum_{p=1}^{P} \omega_p \rho_p(\lambda_n), \quad \text{for } n = 1,2,\ldots,N \tag{6}$$

where ω_p denotes the weighting coefficient for modality p, such that $\sum_p \omega_p = 1$. Then, the fusion problem becomes finding the optimal weight coefficients. Note that when $\omega_p = (1/P)\forall p$, (6) is equivalent to the product rule. Most of the existing classifier fusion schemes [7], [22] actually vary in the way they interpret the weighting coefficients in (6). On one side, there are hard-level combination techniques such as max rule, min rule, and median rule [7] that use binary values for assignment of the weighting coefficients. These techniques combine decisions rather than likelihood scores and in this way try to filter out some of the erroneous likelihoods. The max rule and the min rule for example rely only on the classifier with the highest and the lowest best likelihood scores, respectively, and disregard the decisions of the other classifiers. In this sense, the max rule tends to have a high false accept rate, whereas the min rule is suited to high security applications. Both methods rely solely on likelihood scores and do not employ an additional reliability measure. Soft-level combination techniques, on the other hand,

regard each coefficient as a measure of the relative reliability R_p of each classifier so that each w_p becomes directly equal to R_p. We refer to this combination method as the reliability weighted summation (RWS) rule. Reliability values R_p can be set to some fixed values using some *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 [8], [11], [13], [14]. The problem of the reliability-based weighting approach is that the numerical estimation of reliability values itself, which is ideally feature and class dependent, is not in general very accurate; thus erroneous likelihood scores contribute to the joint score, corrupting correct partial decisions. Actually optimal assignment of the reliability weights still remains as an open problem [22]–[24].

III. Proposed Method

We start by introducing a modified version of the so-called max rule that will give us insight to develop our proposed scheme. The proposed adaptive cascade will eventually offer a compromise between the soft-level and the hard-level classifier combination strategies.

A. Confidence Measure and Modified Max Rule

We propose a confidence measure to define a modified max rule for the open-set problem. The conventional max rule sets the coefficients ω_p in (6) equal to

$$w_p = \begin{cases} 1, & \text{if } p = \arg\max\limits_i \rho_i\left(\lambda_*^i\right) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where λ_*^i denotes the decision class of the ith classifier, i.e. the decision with the corresponding highest best likelihood score. The max rule may filter out some of the erroneous contributions in the final decision; however, there are still three problems with this conventional hard-level classifier combination scheme.

1) The max rule is not well-suited to the open set identification problem, i.e., to detect impostors; it does not adequately take into account strong reject decisions and tends to yield a high false accept rate.
2) The fact that misclassification errors may occur even with high likelihood scores is not taken into account as the best likelihood itself is used to favor the decision of a classifier over the others.
3) Since only the best likelihood score is considered for each classifier, the correct decision cannot be made in cases where the correct decision does not show up as winner in any of the classifiers, although there may be strong evidence for it over the ensemble of all likelihood scores.

In the proposed modified max rule, the highest likelihood ratio in (7) is substituted with the highest confidence measure. Looking back to (3), once a threshold τ is set in the log-likelihood ratio test, one can claim that if the best likelihood score $\rho_p(\lambda_*^p)$ for modality p is much larger or much smaller than τ, the confidence of the accept or reject decision, respectively, is stronger. Hence, the absolute difference between the likelihood score $\rho_p(\lambda_*^p)$ and the threshold τ can be considered as a confidence measure

$$C_p = \left|\rho_p\left(\lambda_*^p\right) - \tau\right|, \quad p = 1, \ldots, P. \tag{8}$$

Then the modified max rule uses the following assignment for the weighting parameters:

$$w_p = \begin{cases} 1, & \text{if } p = \arg\max\limits_{1 \le i \le P} C_i \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

This addresses the first problem in the sense that, a strong reject decision can be favored even though the corresponding likelihood score is not the maximum of the best likelihoods resulting from the P modalities. However, the last two problems above still remain unaddressed.

B. Estimation of Modality Reliability

While the confidence measure C_p indicates our confidence in the accept/reject decision of the classifier for modality p, the proposed adaptive cascade rule will also need to assess the reliability of the source data for modality p. There are two main approaches in the speaker identification literature for adaptive estimation of the reliability of a modality. One approach is the analysis of the data itself from which the corresponding feature vector is extracted and fed to the classifier. Techniques based on this approach try to estimate how much the actual data deviates from the estimated distribution model as in [8], [13]. Such an analysis requires an *a priori* model for the corruption or the noise of the test data [25]. However, in practice the source of statistical deviation is various and difficult to model, such as acoustic/visual noise, time varying characteristics of signals, lighting and pose variations for visual data, etc.

An alternative method is to analyze directly the statistics and rank correlation of the resulting likelihood scores [11], [14], [22]. Reliability estimates based on this approach might be less accurate compared to the first approach in some controlled environments; but the latter is more general, addressing all kinds of possible corruption. It is a known fact that a correct speaker model would create a likelihood ratio that would be significantly higher than the likelihood ratios of the other speaker models. Therefore, the difference between the best two likelihood ratios is commonly used as a reliability measure for the accept scenario [14], [26]. Let $\rho_p(\lambda_*)$ and $\rho_p(\lambda_{**})$ denote the best and the second best likelihood ratios, respectively, resulting from the pth classifier. Then the associated likelihood ratio difference Δ_p is defined as

$$\Delta_p = \rho_p(\lambda_*) - \rho_p(\lambda_{**}). \tag{10}$$

However, in the presence of a reject class, Δ_p does not convey a reliability measure for true reject decisions. In the $N+1$ class open-set identification problem that includes a reject class, we should consider a reliability measure that would also favor true reject decisions as well as true accept decisions. We would expect that a high likelihood ratio $\rho_p(\lambda_*)$ and a high likelihood ratio difference Δ_p are evidences of a true accept decision, and alternatively a low likelihood ratio and a low Δ_p are evidences

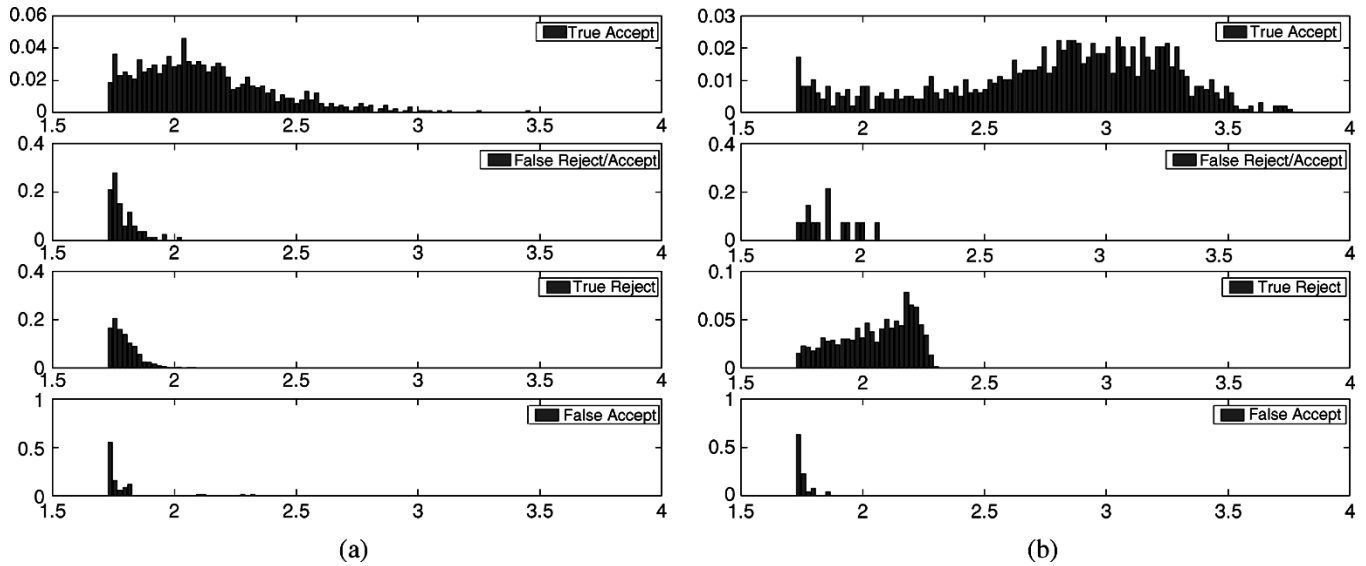

Fig. 1. Histograms of the reliability measure R_p for different classifiers for accept (top two rows) and reject (bottom two rows) scenarios. (a) Audio only classifier at 15-dB noise with EER 6.1%. (b) RWS rule for audio, face, and audio-lip modalities at 10-dB noise with EER 0.7%.

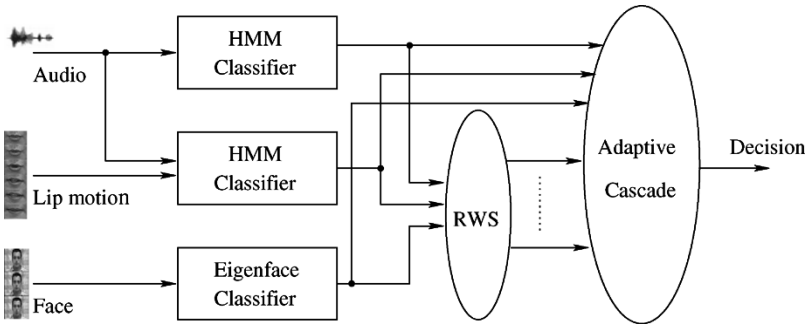

Fig. 2. Block diagram of the proposed modality fusion system for open-set speaker identification.

of a true reject decision. Thus, we propose a new reliability measure R_p given by

$$R_p = \frac{1}{\sum_i \gamma_i}\gamma_p \tag{11}$$

where

$$\gamma_p = \left(e^{(\rho_p(\lambda_*)+\Delta_p)} - 1\right) + \left(e^{(\kappa-\rho_p(\lambda_*)-\Delta_p)} - 1\right). \tag{12}$$

The first and second terms in γ_p are associated with the true accept and true reject, respectively, and κ is a factor that sets the relative weight of the true reject case. Hence, the reliability measure R_p increases when there is an evidence of either true accept or true reject, otherwise stays low. Fig. 1 provides an illustration that the reliability measure R_p attains low values for false accept and false reject decisions as compared to true accept and true reject cases. It should also be noted that when the equal error rate is smaller, the separation of R_p values for true and false decisions is better, which is an expected indication that better classifiers will produce better reliability measures.

C. Adaptive Cascade Rule

Our objective now is to define a new modality fusion rule that addresses all problems associated with the max and RWS rules, inheriting the merits of each of them. This strategy should be able to switch between two modes (hard-level or soft-level) on each decision instant depending on both the reliability of the modalities and the confidence measures. When there is evidence for a reliable strong accept or reject decision in at least one of the classifiers, the strongest decision that is most likely to be true should be favored disregarding the other modalities. When there is ambiguity in the decisions coming from all modalities, the decision is rather not be made based only on a single modality. It may even be the case that all classifiers are wrong and the true decision can be deduced by taking into account the whole ensemble of likelihood scores that they produce. Hence, the need to incorporate reliability weighted modality combinations into the decision scheme. To this effect, assume that there are P different classifiers, each associated with a single modality. Theoretically, it is possible to create a total of $P' = 2^P - 1$ classifier combinations, including P unimodal classifiers and $2^P - 1 - P$ for new combined modalities. Each of these multimodal combinations corresponds to a classifier that produces another set of likelihood scores by some linear combination, i.e., the RWS rule, of the corresponding likelihoods. A reliability measure can also be estimated for each of the combined classifiers and then be incorporated into the decision scheme. The block diagram of the proposed overall modality fusion system is shown in Fig. 2, where all classifier combinations, including unimodal and multimodal combinations, are input to a decision fusion rule which is called the adaptive cascade rule.

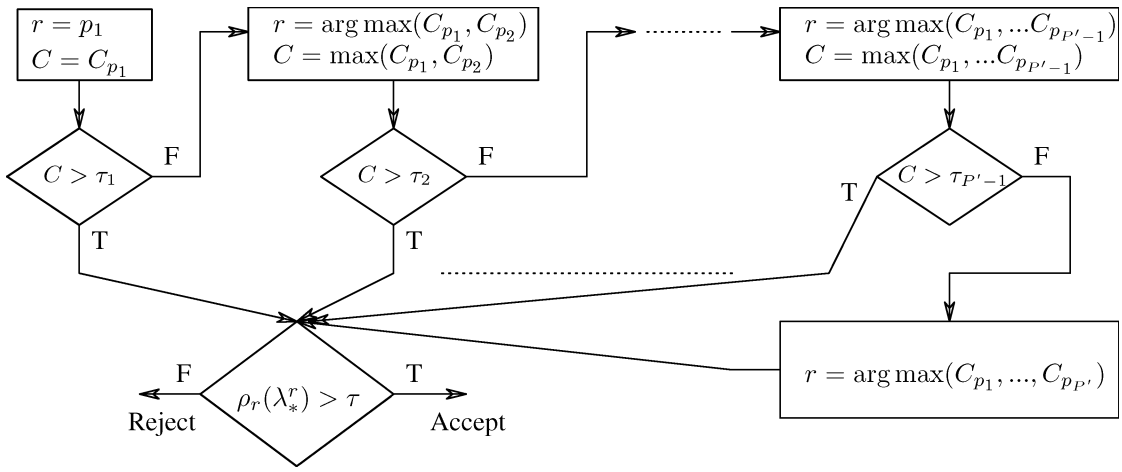

Fig. 3. Flowchart of the adaptive cascade rule. The modality classifiers in the cascade are ordered with respect to their reliability values. C_{p_1} stands for the confidence of the most reliable modality p_1 and $\rho_r(\lambda_*^r)$ is the log-likelihood ratio of the decision coming from the rth modality.

The proposed adaptive cascade rule employs a cascade of P' classifiers, which are adaptively ordered based on their reliability estimates, given P' reliability measures $R_1, R_2, \ldots, R_{P'}$ and P' confidence measures $C_1, C_2, \ldots, C_{P'}$. The order in the classifier cascade $\{p_i\}$ is then arranged such that $R_{p_1} \geq R_{p_2} \geq \ldots \geq R_{p_{P'}}$. This order implicitly defines a priority on each modality or modality combination. Then starting with the most reliable classifier p_1, the cascade rule successively searches for a decision with a sufficiently high confidence measure. As soon as a classifier p_i with sufficiently high confidence measure is encountered, the decision cascade is concluded with accept or reject decision. The exact structure of the adaptive cascade rule is depicted in Fig. 3. Note that the adaptive cascade rule uses P' confidence measure thresholds τ and $(\tau_1, \ldots, \tau_{P'-1})$, each of which has to be determined experimentally. Determining the minimum equal error rate or the receiver operating surface for the above algorithm requires $O(N^{P'})$ computation, where N is the number of classes. For large P', this much computation is usually infeasible. Two ways of improving the algorithm complexity is possible with no significant performance loss. The first improvement is achieved initially by selecting a reduced set of modality combinations that are statistically meaningful and discarding the rest. A further reduction is obtained by adaptive selection of the most reliable $\tilde{P}$ classifiers from the reduced set depending on the reliability order that varies from one decision instant to another. Setting $\tilde{P} = 3$ is usually sufficient resulting in $O(N^3)$ complexity for determination of the three corresponding thresholds τ, τ_1, and τ_2.

D. Error Analysis

The comparison of conventional classifier combination rules such as the max, min, sum, product, or majority vote rules, is usually based on the analysis of the error sensitivity in their estimates of joint likelihood scores or *a posteriori* probabilities [7], [27]. This framework, however, is not appropriate when reliability estimates are incorporated into fusion scheme as in the RWS and the adaptive cascade rules. This is mainly because these reliability-based rules are in fact heuristic methods to approximate the joint likelihoods in the presence of error and can not formally be justified in the Bayesian framework. Moreover, although different techniques exist in the literature to estimate reliability, there is no formal clear definition of what reliability is; nor are there any means to assess how good or effective a given reliability measure is. In the Appendix, we present a detailed error analysis of the proposed adaptive rule in comparison with the RWS and the max rules. The analysis is based on a probabilistic framework that allows the comparison of the probabilities of correct decision for the three rules under question. The probability of correct decision in each case mainly depends on how good the reliability and likelihood estimates are.

To assess the effectiveness of a given reliability measure, we propose to use the probability of that reliability measure to differentiate between true and false decisions coming from different classifiers. Consider two classifiers, one of which gives a true decision (accept or reject) for a given feature sample and the other gives a false decision. Let R_{T} and C_{T} denote respectively the reliability estimate and the confidence measure for the true classifier, that is the classifier with true decision. Similarly, the reliability and the confidence measure for the false classifier are denoted by R_{F} and C_{F}. Then the probability that the true classifier has higher reliability, i.e. $P(R_{\mathrm{T}} > R_{\mathrm{F}})$, can be used to measure the effectiveness of the employed reliability measure whereas the probability $P(C_{\mathrm{T}} > C_{\mathrm{F}})$ similarly measures the effectiveness of the confidence measure. For a given rule, the true accept and true reject probabilities can then be expressed in terms of these two probabilistic measures as given in Appendix. In fact, one can at once intuitively see that a reasonable reliability measure has to at least meet the following condition: $P(R_{\mathrm{T}} > R_{\mathrm{F}}) > P(C_{\mathrm{T}} > C_{\mathrm{F}})$ since otherwise a reliability-based technique would hardly be of any use. The main conclusion of the error analysis in Appendix is that the adaptive cascade rule is expected to outperform the max and the RWS rules, provided that the above condition is satisfied and the employed reliability measure is effective enough in assessment of classifier decisions.

IV. Feature Representation and Classification

In this section we describe the feature representation and classification technique that we use for each modality, i.e. for audio, face and lip motion information. We consider a

text-dependent open-set speaker identification scenario, where the database consists of audio and video signals belonging to individuals of a certain population.

A. Face Modality

The eigenface technique [4], or more generally the principal component analysis, has proved itself as an effective and powerful tool for recognition of still faces. The core idea is to reduce the dimensionality of the problem by obtaining a smaller set of features than the original dataset of intensities. In eigenface technique, every face image is expressed as a linear combination of some basis vectors, i.e. eigenfaces, that best describe the variation of intensities from their mean. Obtaining principal components of a face image can be thought of as an eigenvalue problem. Suppose that the training set consists of M mean-removed face image vectors $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M-1}$. Then the eigenfaces $\boldsymbol{v}_m$, $m = 0, 1, \ldots, M-1$, can be computed as the eigenvectors of the following covariance matrix $\boldsymbol{X}$:

$$\boldsymbol{X} = \frac{1}{M} \sum_{m=0}^{M-1} \boldsymbol{x}_m \boldsymbol{x}_m^T. \tag{13}$$

Each eigenface $\boldsymbol{v}_m$ is associated to an eigenvalue, and principal components are given by the first R eigenfaces associated to the first R eigenvalues when ordered with respect to their magnitudes. Usually, the reduced dimension R is much smaller than M, and the rth eigenimage coefficient ω_r is obtained by the projection $\omega_r = \boldsymbol{v}_r^T \boldsymbol{x}$ for a given image vector $\boldsymbol{x}$.

When a face image sequence, rather than a single image, is available for each speaker as in our case, the images in the sequence can all be used to enforce the classification performance. The eigenface coefficients, when computed for every frame i of a given test sequence, constitute a set of face feature vectors $\boldsymbol{f}_{\mathrm{F}}^i = [\omega_1, \omega_2, \cdots, \omega_r]$, $i = 1, 2, \ldots, K$, that needs to be matched with features of the registered speaker classes. Suppose that the training set contains L face sequences from each speaker class λ_n and let $\boldsymbol{f}_{\mathrm{F}n}^j$, $j = 1, \ldots, K \cdot L$, denote the feature vectors of these images belonging to the class λ_n. Then the minimum distance d_n between these two sets of feature vectors can be used as a similarity metric between the speaker class λ_n and the unknown person

$$d_n = \min_{i,j} \left\| \boldsymbol{f}_{\mathrm{F}}^i - \boldsymbol{f}_{\mathrm{F}n}^j \right\|. \tag{14}$$

The similarity metric defined in (14) can also be expressed as a probabilistic likelihood by making use of the Gibbs distribution: Given the face texture feature vectors $\boldsymbol{f}_{\mathrm{F}}^1, \boldsymbol{f}_{\mathrm{F}}^2, \ldots, \boldsymbol{f}_{\mathrm{F}}^K$, the class conditional probability of the test feature set can be written as

$$P\left(\boldsymbol{f}_{\mathrm{F}}^1, \boldsymbol{f}_{\mathrm{F}}^2, \ldots, \boldsymbol{f}_{\mathrm{F}}^K | \lambda_n\right) = \frac{1}{\nu} e^{-\frac{d_n}{\sigma}} \tag{15}$$

where $\nu = \sum_d e^{-d/\sigma}$ and σ is the decay coefficient of the Gibbs distribution function. The log-likelihood ratio is then defined as

$$\begin{aligned} \rho\left(\boldsymbol{f}_{\mathrm{F}}^1, \boldsymbol{f}_{\mathrm{F}}^2, \ldots, \boldsymbol{f}_{\mathrm{F}}^K | \lambda_n\right) &= \log P\left(\boldsymbol{f}_{\mathrm{F}}^1, \boldsymbol{f}_{\mathrm{F}}^2, \ldots, \boldsymbol{f}_{\mathrm{F}}^K | \lambda_n\right) \\ &\quad - \log P\left(\boldsymbol{f}_{\mathrm{F}}^1, \boldsymbol{f}_{\mathrm{F}}^2, \ldots, \boldsymbol{f}_{\mathrm{F}}^K | \lambda_{N+1}\right) \\ &= \frac{\tilde{d} - d_n}{\sigma} \end{aligned} \tag{16}$$

where σ can be set to 1 or can be used for variance normalization of the likelihood scores. The log-likelihood ratio as defined in (16) requires the definition of a universal background class λ_{N+1}. For this, we adapt the faceness measure defined by the authors in [4] and use the face eigenspace to represent the face universal background class. Hence, $\tilde{d}$ is defined as the distance (or the error) between the original mean-adjusted image vector and its projection to the eigenspace. The log-likelihood ratio in (16) is computed for each class λ_n, and can then be fused with decision scores coming from other available modalities.

B. Audio and Lip Modalities

The two synchronized modality streams, audio and lip, are considered as valuable information sources for the speaker identification problem, especially when they are used jointly under adverse environmental conditions. Audio and lip features are extracted separately from these two synchronized streams at different rates. In our identification system, we employ separate unimodal classifiers for audio and lip modalities as well as bimodal classifiers to exploit the correlation between these two modalities. In the bimodal scenario audio and lip features are concatenated together so that a rate adjustment of audio and lip features becomes necessary.

Audio stream is represented with the mel frequency cepstral coefficients (MFCC), as they yield good discrimination of speech signal. The audio stream is processed over 10-ms frames centered on a 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 [28]. The resulting MFCC features, c_j, are derived using discrete cosine transform (DCT) over log-scaled filter-bank energies e_i

$$c_j = \frac{1}{N_M} \sum_{i=1}^{N_M} e_i \cos\left((i - 0.5)\frac{j\pi}{N_M}\right), \quad j = 1, 2, \ldots, N \tag{17}$$

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 for the kth frame is defined as, $\boldsymbol{C}_k = [c_1 c_2 \cdots c_N]^T$. The audio feature vector $\boldsymbol{f}_{\mathrm{A}}^k$ for the kth frame is formed as a collection of MFCC vector $\boldsymbol{C}_k$ along with the first and second delta MFCC's, $\boldsymbol{f}_{\mathrm{A}}^k = [\boldsymbol{C}_k \Delta\boldsymbol{C}_k \Delta\Delta\boldsymbol{C}_k]$.

The video stream is processed to label lip center locations, where each lip stream is extracted by cropping 64×40 lip frames that are centered to these locations. The gray scale lip stream is transformed into two-dimensional DCT domain and then each lip frame is represented by the first M DCT coefficients of the zig-zag scan excluding the dc-term [29]. The lip feature vector for the ith lip frame is denoted by $\boldsymbol{f}_{\mathrm{L}}^i$.

The unimodal and bimodal temporal characterizations of the audio and the lip modalities are performed using hidden Markov models (HMM), which are reliable structures to model human hearing system and thus widely used for speech recognition and speaker identification problems [2]. In this paper, a word-level continuous-density HMM structure is built for the speaker identification task. Each speaker in the database population is mod-

eled using a separate HMM and is represented with the feature sequence that is extracted over the audio/lip stream while uttering the secret phrase. First a world HMM model is trained over the whole training data of the population. Then each HMM associated to a speaker is trained over some repetitions of the audio-video utterance of the corresponding speaker. In the identification process, given a test feature set, each HMM structure, associated with speakers and the world class, produces a likelihood. The log-ratio of the speaker likelihoods to the world class likelihood results in a stream of log-likelihood ratios to be used in the multimodal decision fusion.

The feature level audio-lip fusion is carried out by concatenating these two features. As the audio features are extracted at a rate of 100 fps and the lip features are extracted at a rate of 15 fps, rate synchronization should be performed prior to any data fusion. The rate of the lip features is increased to match the rate of the audio features by linear interpolation, and it is formulated as follows:

$$\tilde{\boldsymbol{f}}_{\mathrm{L}}^{k} = (1-\alpha_k)\boldsymbol{f}_{\mathrm{L}}^{i^*} + \alpha_k \boldsymbol{f}_{\mathrm{L}}^{i^*-1} \tag{18}$$

where $\tilde{\boldsymbol{f}}_{L}^{k}$ is the interpolated lip feature, which is synchronized by the kth audio feature $\boldsymbol{f}_{A}^{k}$ using the i^*th and (i^*-1)th lip features $\boldsymbol{f}_{\mathrm{L}}^{i^*}$ and $\boldsymbol{f}_{\mathrm{L}}^{i^*-1}$; further i^* and α_k are defined as $i^* = \lfloor 3k/20 \rfloor$, $\alpha_k = (3k/20) - i^*$. After synchronization of the features, the audio and the lip modalities can be fused to each other using data concatenation [29]. The data concatenation is based on the early integration model, where the integration is performed in the feature space to form a composite feature vector of audio and lip modalities. Hence, the joint audio-lip feature $\boldsymbol{f}_{\mathrm{AL}}^{k}$ is formed by combining the audio feature $\boldsymbol{f}_{\mathrm{A}}^{k}$ and the interpolated lip features $\tilde{\boldsymbol{f}}_{\mathrm{L}}^{k}$ for the kth audio-visual frame as

$$\boldsymbol{f}_{\mathrm{AL}}^{k} = \begin{bmatrix} \boldsymbol{f}_{\mathrm{A}}^{k} & \tilde{\boldsymbol{f}}_{\mathrm{L}}^{k} \end{bmatrix}. \tag{19}$$

The temporal characterization of the fused audio-lip feature stream $\boldsymbol{f}_{\mathrm{AL}}^{k}$ can be performed using a single-stream or a multistream HMM structure [30]. In the multistream HMM, each observation vector at kth frame is represented as a collection of audio and lip features, that is $\boldsymbol{f}_{\mathrm{AL}}^{k}$. The observation probability for the multistream HMM at state j, $b_j(\boldsymbol{f}_{\mathrm{AL}}^{k})$, is computed as a synchronous function of audio and lip features

$$b_j\left(\boldsymbol{f}_{\mathrm{AL}}^{k}\right) = \left[\sum_{m=1}^{M_{\mathrm{A}}} c_{\mathrm{A}jm}\mathcal{N}\left(\boldsymbol{f}_{\mathrm{A}}^{k}; \boldsymbol{\mu}_{\mathrm{A}jm}, \boldsymbol{\Sigma}_{\mathrm{A}jm}\right)\right]^{\gamma_{\mathrm{A}}} \times \left[\sum_{m=1}^{M_{\mathrm{L}}} c_{\mathrm{L}jm}\mathcal{N}\left(\tilde{\boldsymbol{f}}_{\mathrm{L}}^{k}; \boldsymbol{\mu}_{\mathrm{L}jm}, \boldsymbol{\Sigma}_{\mathrm{L}jm}\right)\right]^{\gamma_{\mathrm{L}}} \tag{20}$$

where M_{A} and M_{L} are respectively the number of mixture components in the audio and the lip streams, $c_{\mathrm{A}jm}$ and $c_{\mathrm{L}jm}$ are the weights of the mth component and $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The exponents γ_{A} and γ_{L} are the stream weights. Note that, for single-stream HMM the observation probability is simply given as

$$b_j\left(\boldsymbol{f}_{\mathrm{AL}}^{k}\right) = \sum_{m=1}^{M} c_{\mathrm{AL}jm}\mathcal{N}\left(\boldsymbol{f}_{\mathrm{AL}}^{k}; \boldsymbol{\mu}_{\mathrm{AL}jm}, \boldsymbol{\Sigma}_{\mathrm{AL}jm}\right). \tag{21}$$

For statistical modeling of the audio feature $\boldsymbol{f}_{\mathrm{A}}$ and the lip feature $\boldsymbol{f}_{\mathrm{L}}$, single-stream HMM structures are used whereas for the fused audio-lip feature $\boldsymbol{f}_{\mathrm{AL}}$, both single-stream and multistream HMM structures are employed. As presented in the experiments, the most encouraging results are obtained with audio only and multistream audio-lip modalities. Hence, these two modalities along with the face constitute the three unimodal components of our multimodal speaker identification system.

V. Experimental Results

The proposed multimodal speaker identification system has been tested on the audio-visual database MVGL-AVD [29]. The database includes 50 subjects, where each subject utters ten repetitions of her/his name as the secret phrase. A set of impostor data is also available for each subject in the population uttering five different names from the population. The performance of the open-set speaker identification system will be presented in terms of the equal error rate (EER) figure. The EER is calculated as the operating point, where the false accept rate (FAR) equals the false reject rate (FRR). FARs and FRRs are defined as

$$\begin{aligned} \mathrm{FAR} &= 100 \times \frac{\text{number of false accepts}}{N_a + N_r} \\ \mathrm{FRR} &= 100 \times \frac{\text{number of false rejects}}{N_a} \end{aligned} \tag{22}$$

where N_a and N_r are the total number of trials for the true and imposter clients in the testing, respectively. False accepts include the cases where a registered speaker is accepted with a wrong identification or an unregistered speaker is classified as one of the speakers in the database. The database is partitioned into two equal sets in two different ways, so that four different and independent training and testing sessions are deployed. Let $\boldsymbol{D}_T$ represents the whole database for the true clients. In the experimental simulations, the true clients database $\boldsymbol{D}_T$ is partitioned in two ways as $\{\boldsymbol{D}_{T_A}, \boldsymbol{D}_{\bar{T}_A}\}$ and $\{\boldsymbol{D}_{T_B}, \boldsymbol{D}_{\bar{T}_B}\}$, where $\boldsymbol{D}_{T_A}$ and $\boldsymbol{D}_{T_B}$ are disjoint sets, each having five repetitions from each subject in the database. Training and testing are performed over four independent sessions, where $\{\boldsymbol{D}_{T_A}, \boldsymbol{D}_{\bar{T}_A}\}$, $\{\boldsymbol{D}_{\bar{T}_A}, \boldsymbol{D}_{T_A}\}$, $\{\boldsymbol{D}_{T_B}, \boldsymbol{D}_{\bar{T}_B}\}$, and $\{\boldsymbol{D}_{\bar{T}_B}, \boldsymbol{D}_{T_B}\}$ pairs are, respectively, used for training and testing. As there are 50 subjects and five repetitions for each true and imposter client tests, the resulting total number of trials are given by $N_a = 1000$ and $N_r = 1000$.

The temporal characterization of the audio, lip and multistream audio-lip has been achieved using a 6-state left-to-right HMM structure for each speaker. The acquired video data is first split into segments of secret phrase utterances. The visual and audio streams are then separated into two parallel streams, where the visual stream has gray-level video frames of size 720 $\times$ 576 pixels containing the frontal view of a speaker's head at a rate of 15 fps and the audio stream has 16 bits/sample at 16-kHz sampling rate. 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.

In the analysis of audio stream, the MFCC feature vector, $\boldsymbol{C}_k$, is composed of 13 cepstral coefficients using 26 mel frequency bins. The first mel-band that corresponds to the first energy term

TABLE I
ABBREVIATIONS AND DESCRIPTIONS FOR MODALITIES AND FUSION TECHNIQUES

A	Audio only modality
L	Lip only modality
F	Face only modality
AL	Audio-Lip data fusion with single-stream HMM
ALms	Audio-Lip data fusion with multi-stream HMM
$+$	Product rule
$\oplus$	RWS rule
$\star$	Modified max rule
$\circ$	Non-adaptive cascade rule (fixed a priori reliability ordering, where leftmost being the most reliable modality)
$\bullet$	Adaptive cascade rule (adaptive reliability ordering)

TABLE II
SPEAKER IDENTIFICATION RESULTS: EQUAL ERROR RATES AT VARYING NOISE LEVELS FOR DIFFERENT MODALITIES

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	25	20	15	10	7	5
A	2.4	2.5	3.7	6.1	11.0	18.9	26.5
F	8.4						
L	18.0						
AL	15.6	15.8	15.8	16.2	16.4	16.7	16.7
ALms	13.5	13.8	13.8	13.8	14.9	15.3	15.4

e_1 [see (17)] is picked to start at 250 Hz and the dc-term c_0 is excluded, as the low frequency components and the dc-term do not carry valuable information for the speaker identification process [29]. The resulting audio feature vector, $\boldsymbol{f}_{\mathrm{A}}^{k}$ of size 39, includes the MFCC vector together with the first and second delta MFCC vectors.

Each video stream for a single secret phrase uttering is around 1 second in duration and during this time it is assumed that the subject does not considerably move her/his head. The MVGL-AVD database includes the hand-labeled lip center locations for the first frames of each video stream. Hence, each lip stream is extracted by cropping 64×40 lip frames to form the lip sequence of each secret phrase utterance. The lip feature vectors $\boldsymbol{f}_{\mathrm{L}}^{k}$, which are used in both training and testing of the HMM-based classifier, are obtained as described in Section IV-B with $M = 60$. The stream weights γ_A and γ_L are picked, respectively, as 0.7 and 0.3 for the multistream HMM structure. Similarly, face image streams are extracted and an eigenspace of dimension $R = 20$ is computed using a collection of face images that includes two face images from each utterance in the training part of the MVGL-AVD database.

A summary of the employed modalities and the decision fusion techniques is given in Table I that also describes the notation used in presenting the experimental results. The unimodal identification results are shown in Table II, where we observe the equal error rates at varying levels of acoustic noise. In the audio-only scenario, the identification performance degrades rapidly with decreasing SNR. For the face-only case, images in the training and test sets have varying backgrounds and lighting; this is why the face-only identification performance may seem to be worse than expected. The lip modality alone yields 18% equal error rate performance. When lip features are fused with audio using single-stream or multistream HMM structures, the identification performance improves with respect to lip-only performance only at low SNR levels and even degrades under acoustically clean conditions. However, they still carry important information on the temporal correlations of audio-lip modalities that can be exploited during the multimodal fusion process to improve the overall performance. We also note that multistream audio-lip fusion visibly outperforms single-stream fusion at all SNR levels.

A. Performance of the RWS

In the targeted audio-visual speaker identification problem, audio stream is correlated with the audio-lip stream, and under different environmental conditions such as additive noise or varying lighting conditions, some of the modalities do not produce unbiased features. Under such varying environmental conditions, if a classifier is not capable of producing reliable decisions, its contribution should better be reduced in the classifier combining process. The RWS rule addresses this classifier combining process with properly selected weights. The proposed reliability measure in (11) assumes a κ factor that needs to be determined so that it sets an optimal compromise between accept and reject scenarios. The audio-only, face-only and multistream audio-lip modalities are combined using RWS rule at varying acoustical noise levels with a set of possible κ factors. The average equal-error-rate curve over different SNR levels is plotted in Fig. 4(a) for varying values of κ factor. The optimal value is observed to be $\kappa = 0.65$ that minimizes the average EER curve in Fig. 4(a). The value of the factor $\kappa = 0.65$ sets some bias to weight true accept decisions slightly more than true reject decisions, which is expected from the nature of product rule that normally favors true accepts.

The RWS rule based performances are presented in Table III together with the performances of the so-called product rule, that is the summation of log-likelihoods with uniform weighting. One can observe that at all SNR levels, any combination of modalities obtained by both the product rule and the RWS rule performs at least better than the worst unimodal performance. When audio and face are employed in the fusion (see also Table I), bimodal performances are all better than the best unimodal performance. The RWS rule yields a significant improvement over product rule for all test conditions, and the best equal-error-rate scores are obtained with the fusion of audio, face image, and multistream audio-lip modalities. Note that, when the multistream audio-lip modality is fused with the $(\mathbf{A} \oplus \mathbf{F})$ combination by RWS rule, that is $\mathbf{A} \oplus \mathbf{F} \oplus \mathbf{ALms}$, the performance significantly improves over all SNR range except for the clean condition. This performance improvement is expected as the multistream audio-lip modality sustains fairly

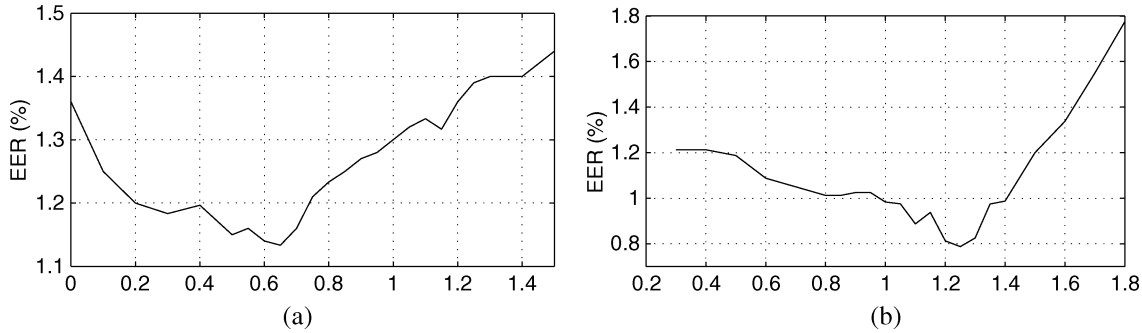

Fig. 4. Average EER performances over different SNR levels for varying values of κ factor: (a) with the RWS rule and (b) with the adaptive cascade rule.

TABLE III
SPEAKER IDENTIFICATION RESULTS: EQUAL ERROR RATES FOR PRODUCT AND RWS RULES AT VARYING NOISE LEVELS

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	25	20	15	10	7	5
A + F	1.9	2.0	2.6	3.8	6.6	8.5	11.9
A ⊕ F	0.5	1.0	1.1	1.2	3.2	5.2	8.2
A + L	8.0	8.2	8.7	9.3	11.6	16.8	22.5
A ⊕ L	5.6	6.0	7.2	9.2	11.4	16.7	22.1
A + AL	7.3	7.5	7.4	7.8	10.3	14.9	16.4
A ⊕ AL	5.0	5.2	6.0	6.9	8.9	13.1	15.9
A + ALms	6.1	6.4	6.4	7.1	9.9	15.5	18.2
A ⊕ ALms	4.6	4.7	5.4	6.3	8.7	13.6	17.1
F + ALms	5.8	5.9	5.9	5.8	6.5	6.7	6.9
F ⊕ ALms	2.9	3.0	3.2	3.4	3.6	3.8	3.9
A + F + ALms	1.1	1.1	1.1	1.3	1.7	2.8	4.3
A ⊕ F ⊕ ALms	0.7	0.7	0.7	1.0	1.4	2.2	3.5

TABLE IV
SPEAKER IDENTIFICATION RESULTS: EQUAL ERROR RATES FOR DIFFERENT DECISION FUSION TECHNIQUES AT VARYING NOISE LEVELS

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	25	20	15	10	7	5
A ⋆ F ⋆ ALms	1.8	1.8	2.6	4.9	6.4	7.0	9.0
A ∘ F ∘ ALms	2.3	2.2	2.5	3.3	5.3	7.4	12.6
A • F • ALms	1.4	1.5	1.7	1.9	2.5	3.9	6.3
A • F	0.8	0.8	0.8	1.5	2.6	4.5	7.9
M_0 = (**A ⊕ F ⊕ ALms**)	0.7	0.7	0.7	1.0	1.4	2.2	3.5
M_1 = (**A ⊕ F**)	0.5	1.0	1.1	1.2	3.2	5.2	8.2
M_2 = (**F ⊕ ALms**)	2.9	3.0	3.2	3.4	3.6	3.8	3.9
$M_0 \oplus M_1 \oplus M_2 \oplus$ **A ⊕ F**	0.4	0.6	0.6	1.0	1.6	1.9	2.6
M_0 • **A • F**	0.3	0.3	0.4	0.5	1.2	2.1	4.5
M_0 • M_1 • **A • F**	0.2	0.2	0.3	0.4	1.1	2.1	4.3
M_0 • M_1 • M_2 • **A • F**	0.2	0.2	0.3	0.6	1.0	1.6	2.3

robust performance, which is partially uncorrelated from the audio-only performance under noisy conditions.

B. Performance of the Adaptive Classifier Cascade

Classifier cascade sets a reliability order for each modality. The reliability ordering of the modalities could be set *a priori* or modality reliability could adaptively be estimated using the proposed reliability measure in (11). When the proposed reliability measure is used, an optimal κ factor should be extracted that best fits the adaptive cascade rule. In order to set such a κ factor, some selected modalities are combined with the adaptive cascade rule at varying acoustical noise levels with a set of possible κ factors. The average equal-error-rate curve over different SNR levels and over classifier cascades of various selected modality combinations is plotted in Fig. 4(b) for varying values of κ factor. The optimal factor is measured as $\kappa = 1.25$ from Fig. 4(b). This κ value weights the reliability measure almost equally for accept and reject decisions.

Decision fusion results are presented in Table IV, where the product, the max and the adaptive cascade rules are evaluated. The first three rows compare respectively the max, the nonadaptive cascade and the adaptive cascade rules by combining audio, face and audio-lip streams. The performance of the adaptive cascade clearly outperforms both the max and the nonadaptive (fixed reliability ordered) cascade rules, and achieves 1.4% and 6.3% EER rates at clean and 5-dB SNR conditions, respectively.

Our reliability measure is a function of the difference between the best and the second best likelihood scores as well as the best likelihood itself. If a modality stream is well separated for true and imposter claims, it yields a better EER performance and also a better estimation for the proposed reliability measure. Hence, when the single stream modalities are adaptively cascaded, the individual streams that have the best EER performances are expected to yield the best cascade performance. This tendency is expected since better estimation of the modality reliability yields better cascade performance as can also be observed from the decision fusion results that are presented in Table IV. The adaptive cascade of audio, face, and multistream audio-lip ($\mathbf{A} \bullet \mathbf{F} \bullet \mathbf{ALms}$) versus audio and face cascade ($\mathbf{A} \bullet \mathbf{F}$) is investigated and the corresponding EER results are presented in the third and fourth rows, respectively. The audio/face cascade is found to perform significantly better than the audio/face/audio-lip cascade, especially under high SNR conditions. This result is also as expected in view of the individual EER performances of audio, face and multistream audio-lip modalities displayed in Table II, where the multistream audio-lip modality has significantly poorer EER performance among the three unimodal streams over 10-dB SNR level, that also yields a poor estimation of the reliability of the $\mathbf{ALms}$ stream.

The analysis of the adaptive cascade performance of audio, face and audio-lip modalities reveals an important finding, that one should not include a stream with a poor EER performance to the cascade rule since it also yields a poor reliability estimation. This finding leads us to examine the RWS modality combinations, as defined in Section III-C, that have better EER performances than the unimodal streams; the reliability estimation of these combinations are expected to be better than the

unimodal reliability estimates. Three such combined modalities are considered by fusing audio, face and multistream audio-lip modalities in different combinations using RWS rule, specifically $(\mathbf{A} \oplus \mathbf{F} \oplus \mathbf{ALms})$, $(\mathbf{A} \oplus \mathbf{F})$ and $(\mathbf{F} \oplus \mathbf{ALms})$. The EER performances of these three combined modalities are given in Table IV. Once these combined modalities are adaptively cascaded with relatively reliable unimodal streams, i.e., audio and face, a further performance gain is achieved. This performance gain is an indicator of robust reliability estimates for each single or combined modality included in the adaptive cascade.

The best EER performances, i.e. 0.2% at clean conditions and 2.3% at 5-dB SNR, are obtained with the adaptive classifier cascade including five different modality combinations; at each decision instant, only the three most reliable of these combinations are actually used in the decision mechanism as described in Section III-C. One can also consider the same set of modalities (single or combined) in a possible RWS-only rule fusion scheme, as presented in the eighth row of Table IV. The eighth row performances indicate that in this case RWS fusion does slightly better than the individual RWS combinations, but still the overall performance gain is not as good as the adaptive cascade fusion. Hence, one can suggest to use the RWS rule for combining unimodal classifiers to create stronger modality combinations that can further be fused using adaptive cascade rule to boost the overall performance.

VI. Conclusions

We have presented a multimodal (audio-lip-face) open-set speaker identification system that aims at robust performance under adverse environmental conditions. The proposed adaptive cascade rule outperforms traditional fusion schemes such as the product rule and the max rule. Since the performance of the adaptive cascade rule depends on the effectiveness of the employed reliability measure, a novel modality reliability estimation scheme that performs successfully under both accept and reject scenarios has also been proposed. This measure, that is based on the likelihood ratio scores, differentiates the best likelihood ratio score from the rest of the scores, creating a relative assessment on the reliability of each modality. We have also analytically shown that the proposed adaptive cascade rule outperforms the hard-level max rule and soft-level weighted summation rule, provided that the employed reliability measure is effective enough in assessment of classifier decisions.

An important feature of this work is the use of combined modalities in the decision fusion scheme. Some modality combinations, obtained via for instance RWS rule, may achieve much better EER performances than the single modalities; such combined modalities can be considered as additional reliable sources for the adaptive cascade rule. The experimental findings support that the adaptive cascade of the strong modality combinations together with the reliable unimodal streams can further boost the overall performance. The speaker identification results that have been presented are encouraging for robust speaker identification systems. The adaptive cascade rule, as a high performance classifier combining scheme, can also be used in many other multimodal identification applications.

Appendix
Error Analysis for the Adaptive Cascade Rule

We provide a probabilistic error analysis to compare the performance of the proposed adaptive cascade rule with those of the conventional max and RWS rules. Suppose that there are P unimodal classifiers, that can be combined. Three possible scenarios exist: The first case is that all P unimodal classifiers individually give the true decision (accept or reject); in this case all three classifier combining rules, i.e. the max, adaptive cascade and RWS rules, will each yield the true decision. The second case is that all unimodal classifiers give individually false decisions. In this case the max rule has no chance to yield the correct decision. Even though the RWS rule and the adaptive cascade will most likely fail as well in this case, they still have some chance to extract the correct decision through multiple modality combinations. In the third case, at least one of the classifiers gives the correct decision and one gives a false decision.

In the following, for the sake of simplicity, we will consider an open-set identification problem with $N = 2$ and $P = 2$. The error analysis involves the estimation of false (or equivalently true) accept and reject probabilities, and can be generalized to an $(N + 1)$ class and P modality problem. Let us first assume that the adaptive cascade rule does not include any RWS classifier combinations and works only on two unimodal classifiers. Let $\rho_{1\mathrm{T}}$ and $\rho_{2\mathrm{T}}$ denote the best and the second best likelihood scores of the classifier that gives the true decision. Similarly, $\rho_{1\mathrm{F}}$ and $\rho_{2\mathrm{F}}$ will stand for the likelihood scores of the other classifier that gives a false decision. Thus, for a given fixed likelihood test threshold τ, if the true decision is an accept decision, then $\rho_{1\mathrm{T}} > \tau$ and it is associated with the correct class, whereas $\rho_{1\mathrm{F}} < \tau$ or it is associated with an incorrect class. Let also R_{T}, R_{F} and C_{T}, C_{F} denote the respective reliability estimates and confidence measures of the classifiers.

Consider the max rule given by (9) in Section III-A. The probability of true accept decision $P_{\mathrm{MAX}}(\mathrm{TA})$ can be written as

$$P_{\mathrm{MAX}}(\mathrm{TA}) = P(C_{\mathrm{T}} > C_{\mathrm{F}}) \tag{23}$$

and for the RWS rule, we have

$$\begin{aligned} P_{\mathrm{RWS}}(\mathrm{TA}) &= P(R_{\mathrm{T}}\rho_{1\mathrm{T}} + R_{\mathrm{F}}\rho_{2\mathrm{F}} > \tau) \\ &\quad \times P(R_{\mathrm{T}}\rho_{1\mathrm{T}} + R_{\mathrm{F}}\rho_{2\mathrm{F}} > R_{\mathrm{T}}\rho_{2\mathrm{T}} + R_{\mathrm{F}}\rho_{1\mathrm{F}}) \end{aligned} \tag{24}$$

whereas for the cascade rule

$$\begin{aligned} &P_{\mathrm{CASC}}(\mathrm{TA}) \\ &\quad = P(R_{\mathrm{T}} > R_{\mathrm{F}})\,[P(C_{\mathrm{T}} > \tau_1) + P(C_{\mathrm{T}} < \tau_1)P(C_{\mathrm{T}} > C_{\mathrm{F}})] \\ &\qquad + P(R_{\mathrm{T}} < R_{\mathrm{F}})P(C_{\mathrm{F}} < \tau_1)P(C_{\mathrm{T}} > C_{\mathrm{F}}). \end{aligned} \tag{25}$$

We need to compare the probabilities given by (23), (24), and (25). For comparison, we fix the threshold τ for all three rules as the value that gives the equal error rate for the RWS rule.

Since likelihood estimates are in general very sensitive to modeling/measurement errors, one of the criteria that a reasonable reliability measure has to meet can be stated as follows:

$$P(R_{\mathrm{T}} > R_{\mathrm{F}}) > P(C_{\mathrm{T}} > C_{\mathrm{F}}) \tag{26}$$

since otherwise incorporating reliability values into a decision scheme would be of no use. The expression in (25) can now be viewed as a function of τ_1 for which two extreme cases can be observed

$$P_{\mathrm{CASC}}(\mathrm{TA}) = \begin{cases} P(R_{\mathrm{T}} > R_{\mathrm{F}}), & \text{if } \tau_1 = 0 \\ P(C_{\mathrm{T}} > C_{\mathrm{F}}), & \text{if } \tau_1 = 1. \end{cases} \tag{27}$$

An observation that can be deduced from (25) and (27) is that the true accept probability of the max rule, $P(C_{\mathrm{T}} > C_{\mathrm{F}})$, is a lower bound for $P_{\mathrm{CASC}}(\mathrm{TA})$. It can also be shown that $P_{\mathrm{CASC}}(\mathrm{TA})$, i.e., the expression in (25), takes a local maximum value between the two extreme values of τ_1. Thus, with a proper choice of τ_1, it is possible to have

$$P_{\mathrm{CASC}}(\mathrm{TA}) > P(R_{\mathrm{T}} > R_{\mathrm{F}}). \tag{28}$$

It is now easy to see that $P_{\mathrm{CASC}}(\mathrm{TA}) > P_{\mathrm{MAX}}(\mathrm{TA})$, and the magnitude of the difference between the two true accept probabilities depends on how good the estimate or the measure of the classifier reliability is as compared to likelihood estimates.

To compare $P_{\mathrm{CASC}}(\mathrm{TA})$ and $P_{\mathrm{RWS}}(\mathrm{TA})$, first note that

$$\begin{aligned} &P(R_{\mathrm{T}}\rho_{1\mathrm{T}} + R_{\mathrm{F}}\rho_{2\mathrm{F}} > R_{\mathrm{T}}\rho_{2\mathrm{T}} + R_{\mathrm{F}}\rho_{1\mathrm{F}}) \\ &\quad = P\left(R_{\mathrm{T}}(\rho_{1\mathrm{T}} - \rho_{2\mathrm{T}}) > R_{\mathrm{F}}(\rho_{1\mathrm{F}} - \rho_{2\mathrm{F}})\right) \\ &\quad = P(R_{\mathrm{T}}\Delta_{\mathrm{T}} > R_{\mathrm{F}}\Delta_{\mathrm{F}}) \end{aligned} \tag{29}$$

where $\Delta_{\mathrm{T}} = \rho_{1\mathrm{T}} - \rho_{2\mathrm{T}}$ and $\Delta_{\mathrm{F}} = \rho_{1\mathrm{F}} - \rho_{2\mathrm{F}}$. When $N = 2$, Δ_{T} and Δ_{F} become identical to the likelihood score differences as defined by (10) in Section III-B. Recalling the discussion in Section III-B, if we have a reliability measure that is better than Δ_{T} and Δ_{F}, we can write

$$P(R_{\mathrm{T}} > R_{\mathrm{F}}) > P(\Delta_{\mathrm{T}} > \Delta_{\mathrm{F}}). \tag{30}$$

Hence, $P(R_{\mathrm{T}} > R_{\mathrm{F}}) > P(R_{\mathrm{T}}\Delta_{\mathrm{T}} > R_{\mathrm{F}}\Delta_{\mathrm{F}})$. Using this and comparing (24) and (28), we conclude that $P_{\mathrm{CASC}}(\mathrm{TA}) > P_{\mathrm{RWS}}(\mathrm{TA})$. The assumption given by (30) can also be regarded as another criterion that a reasonable reliability measure has to meet since otherwise the RWS rule would hardly be better than the conventional product rule, i.e., than the uniformly weighted summation in log domain.

The same analysis on true reject probabilities, $P_{\mathrm{MAX}}(\mathrm{TR})$, $P_{\mathrm{RWS}}(\mathrm{TR})$ and $P_{\mathrm{CASC}}(\mathrm{TR})$ yields a similar conclusion. In the reject scenario, the expressions for the max and adaptive cascade rules remain the same as in (23) and (25). For the RWS rule, the expression changes slightly, but the value of the expression remains the same since τ was set so as to give an equal error rate such that $P_{\mathrm{RWS}}(\mathrm{TR}) = P_{\mathrm{RWS}}(\mathrm{TA})$.

For the error analysis, we have not yet incorporated the RWS combinations into the cascade rule. When the RWS combinations are also included in the decision cascade, $P_{\mathrm{CASC}}(\mathrm{TA})$ is expected to improve. This improvement becomes critical if we generalize the analysis to arbitrary N, i.e. $N > 2$. In this case, it becomes difficult to meet the criterion stated in (30) since Δ_{T} and Δ_{F} of (29) are no longer identical to the likelihood score differences of (10). Let us now examine how the value of $P_{\mathrm{CASC}}(\mathrm{TA})$ changes when the RWS combinations are incorporated. Since $P = 2$, there is only one possible RWS combination and $P_{\mathrm{CASC}}(\mathrm{TA})$ can be written as a conditional probability depending on whether the decision of the RWS classifier is true or not, i.e. in terms of $P_{\mathrm{RWS}}(\mathrm{TA})$. Following a similar analysis that led us to (28), one can write the following inequality (with proper choice of τ_1 and τ_2):

$$\begin{aligned} P_{\mathrm{CASC}}(\mathrm{TA}) > P_{\mathrm{RWS}}(\mathrm{TA})\,&(1 - P(R_{\mathrm{F}} > R_{\mathrm{RWS}})P(R_{\mathrm{F}} > R_{\mathrm{T}})) \\ &+ (1 - P_{\mathrm{RWS}}(\mathrm{TA}))\,P(R_{\mathrm{T}} > R_{\mathrm{RWS}})P(R_{\mathrm{T}} > R_{\mathrm{F}}) \end{aligned} \tag{31}$$

where R_{RWS} denotes the reliability of the RWS combination. In (31), $(1 - P(R_{\mathrm{F}} > R_{\mathrm{RWS}})P(R_{\mathrm{F}} > R_{\mathrm{T}}))$ and $P(R_{\mathrm{T}} > R_{\mathrm{RWS}})P(R_{\mathrm{T}} > R_{\mathrm{F}})$ are the respective conditional probabilities that a true classifier appears as the foremost classifier in the cascade. In the former term, one can substitute $P(R_{\mathrm{F}} > R_{\mathrm{RWS}})$ by $P(R_{\mathrm{F}} > R_{\mathrm{T}})$ since R_{RWS} is the reliability of a true classifier by the given condition. Likewise, $P(R_{\mathrm{T}} > R_{\mathrm{RWS}})$ can be substituted $P(R_{\mathrm{T}} > R_{\mathrm{F}})$. Thus, we have

$$\begin{aligned} P_{\mathrm{CASC}}(\mathrm{TA}) > P_{\mathrm{RWS}}(\mathrm{TA})\,&\left(1 - P^2(R_{\mathrm{F}} > R_{\mathrm{T}})\right) \\ &+ (1 - P_{\mathrm{RWS}}(\mathrm{TA}))\,P^2(R_{\mathrm{T}} > R_{\mathrm{F}}). \end{aligned} \tag{32}$$

Looking at (32), we observe that $P_{\mathrm{CASC}}(\mathrm{TA}) > P_{\mathrm{RWS}}(\mathrm{TA})$ whenever the following condition is satisfied

$$P_{\mathrm{RWS}}(\mathrm{TA}) < \frac{1}{\left(\frac{1-P(R_{\mathrm{T}}>R_{\mathrm{F}})}{P(R_{\mathrm{T}}>R_{\mathrm{F}})}\right)^2 + 1}. \tag{33}$$

The above condition is much easier for a good reliability measure to meet as compared to the condition in (30). Note that when $P(R_{\mathrm{T}} > R_{\mathrm{F}}) = 1$, which means that the reliability measure is perfect, the inequality is always satisfied since the right-hand side of the inequality becomes 1. Moreover the left-hand side, which is also dependent on the value of $P(R_{\mathrm{T}} > R_{\mathrm{F}})$, is expected to decrease more rapidly than the expression on the right-hand side as the value of $P(R_{\mathrm{T}} > R_{\mathrm{F}})$ decreases within the range of values close to 1.

REFERENCES

[1] N. Ratha, A. Senior, and R. M. Bolle, "Automated biometrics," *Proc. ICAPR*, pp. 445–474, May 2001.

[2] J. Campbell, "Speaker recognition: a tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[3] Y. Y. J. Zhang and M. Lades, "Face recognition: eigenface, elastic matching, and neural nets," *Proc. IEEE*, vol. 85, no. 9, pp. 1423–1435, Sep. 1997.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cog. Neurosci.*, vol. 3, no. 1, pp. 586–591, Sep. 1991.

[5] D. D. Zhang, *Automated Biometrics*. Norwell, MA: Kluwer, 2000.

[6] A. Verma, T. Faruquie, C. Neti, and S. Basu, "Late integration in audio-visual continuous speech recognition," *Automatic Speech Recognition and Understanding (ASRU)*, 1999.

[7] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[8] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognit.*, vol. 36, no. 2, pp. 293–302, Feb. 2003.

[9] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1065–1074, Sep. 1999.

[10] V. Chatzis, A. G. Borş, and I. Pitas, "Multimodal decision-level fusion for person authentication," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 29, no. 6, pp. 674–680, Nov. 1999.

[11] R. Brunelli and D. Falavigna, "Person identification using multiple clues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 955–966, Oct. 1995.
[12] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction," in *Proc. Int. Conf. Multimedia & Expo 2003 (ICME2003)*, vol. 3, Jul. 2003, pp. 9–12.
[13] ——, "Audio-visual speaker recognition using time-varying stream reliability prediction," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing 2003 (ICASSP 2003)*, vol. V, 2003, pp. 712–715.
[14] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Process.*, vol. 11, no. 3, pp. 169–186, Jul. 2001.
[15] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognit. Lett.*, vol. 18, no. 9, pp. 853–858, 1997.
[16] M. R. Civanlar and T. Chen, "Password-free network security through joint use of audio and video," in *Proc. SPIE Photonic*, Nov. 1996, pp. 120–125.
[17] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. L. les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacretaz, *Lecture Notes in Computer Science*. Heidelberg, Germany: Springer-Verlag, Aug. 2003, vol. 2688, pp. 845–853.
[18] R. Frischholz and U. Dieckmann, "BioID: a multimodal biometric identification system," *J. IEEE Comput.*, vol. 33, no. 2, pp. 64–68, Feb. 2000.
[19] L. A. Alexandre, A. C. Campilho, and M. Kamel, "Combining independent and unbiased classifiers using weighted average," in *Proc. 15th Int. Conf. Pattern Recognition*, vol. 2, 2000, pp. 3–7.
[20] I. Bloch, "Information combination operators for data fusion: a comparative review with classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 26, no. 1, pp. 52–67, 1996.
[21] H. Altincay and M. Demirekler, "Undesirable effects of output normalization in multiple classifier systems," *Pattern Recognit. Lett.*, vol. 24, pp. 1163–1170, 2003.
[22] ——, "An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification," *J. Speech Commun.*, vol. 30, pp. 255–272, 2000.
[23] K. Al-Ghoneim and B. V. K. V. Kumar, "Unified decision combination framework," *Pattern Recognit.*, vol. 31, no. 12, pp. 2077–2089, 1998.
[24] J. A. Benediktsson and P. H. Swain, "Consensus theoretic classification methods," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 4, pp. 688–704, Apr. 1992.
[25] N. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 158–166, Mar. 2002.
[26] A. Rogozan, P. Deleglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," in *Proc. AVSP Workshop, ESCA*, 1997, pp. 61–64.
[27] J. Kittler and F. M. Alkoot, "Sum versus vote fusion in multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 1, pp. 110–115, Jan. 2003.
[28] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[29] E. Erzin, Y. Yemez, and A. M. Tekalp, "Joint audio-video processing for robust biometric speaker suidentification in car," in *DSP for In-Vehicle and Mobile Systems*, H. Abut, J. H. L. Hansen, and K. Takeda, Eds. Berlin/Heidelberg/New York: Springer-Verlag, 2004.
[30] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins summer 2000 workshop," in *Proc. Workshop Multimedia Signal Processing*, 2001, pp. 619–624.

Engin Erzin (S'88–M'96) received the B.Sc., M.Sc., and Ph.D. degrees from Bilkent University, Ankara, Turkey, in 1990, 1992, and 1995, respectively, all in electrical engineering.

During 1995–1996, he was a postdoctoral Fellow in the Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, where he was with Consumer Products for one year as a Member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he has been with the Electrical and Electronics Engineering Department and Computer Engineering Department, Koç University, Istanbul, Turkey. His research interests include speech signal processing, pattern recognition, and adaptive signal processing.

Yücel Yemez (M'03) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1989, and the M.Sc. and Ph.D. degrees from Boğaziçi University, Istanbul, Turkey, in 1992 and 1997, respectively, all in electrical engineering.

From 1997 to 2000, he was a postdoctoral Researcher in the Image and Signal Processing Department, Télécom Paris (Ecole Nationale Supérieure des Télécommunications), Paris, France. Currently he is an Assistant Professor of the Computer Engineering Department at Koç University, Istanbul, Turkey. His current research is focused on various fields of computer vision and three-dimensional computer graphics.

A. Murat Tekalp (S'80–M'84–SM'91–F'03) received the M.S. and Ph.D. degrees in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, in 1982 and 1984, respectively.

From December 1984 to August 1987, he was with Eastman Kodak Company, Rochester, NY. He joined the Electrical and Computer Engineering Department, University of Rochester, in September 1987, where he is currently a Distinguished Professor. Since June 2001, he has also been with Koç University, Istanbul, Turkey. His research interests are in the area of digital image and video processing, including video compression and streaming, motion-compensated video filtering for high-resolution, video segmentation, object tracking, content-based video analysis and summarization, multicamera surveillance video processing, and protection of digital content. His group has contributed technology to the ISO/IEC MPEG-4 and MPEG-7 standards. He is author of *Digital Video Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1995) and holds six U.S. patents.

Dr. Tekalp was named a Distinguished Lecturer by the IEEE Signal Processing Society in 1998, and awarded a Fulbright Senior Scholarship in 1999. He received the TUBITAK Science Award (highest scientific award in Turkey) in 2004. He has chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (January 1996–December 1997). He has served as an Associate Editor for the IEEE TRANS. ON SIGNAL PROCESSING (1990–1992), IEEE TRANS. ON IMAGE PROCESSING (1994–1996), and the Kluwer *Journal Multidimensional Systems and Signal Processing* (1994–2002). He was an Area Editor for the Academic Press *Journal Graphical Models and Image* Processing (1995–1998). He was also on the editorial board of the Academic Press *Journal Visual Communication and Image Representation* (1995–2002). He was appointed as the Technical Program Chair for the 1991 IEEE Signal Processing Society Workshop on image and Multidimensional Signal Processing, the Special Sessions Chair for the 1995 IEEE International Conference on Image Processing, the Technical Program Co-Chair for IEEE ICASSP 2000 in Istanbul, Turkey, and the General Chair of IEEE International Conference on Image Processing (ICIP) in 2002. He is the founder and first Chairman of the Rochester Chapter of the IEEE Signal Processing Society. He was elected as the Chair of the Rochester Section of IEEE in 1994–1995. Currently, he is the Editor-in-Chief of the EURASIP journal *Signal Processing: Image Communication* (published by Elsevier).