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. 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.

1. 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. 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 [2]. 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 speaker identification, a reject scenario is not defined and an unknown speaker is assigned to one of the N registered people. In the open-set identification, 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. The audio content also creates two different identification problems. We can refer these two problems as text-dependent and text-independent speaker identification. In the text-independent problem, identification is performed over a content free utterance of the targeted speaker, whereas in the text-dependent problem, each speaker is expected to utter a personalized secret phrase for the identification job. In the latter case, the system provides an additional level of access security. However a particular attention is needed in this case to handle impostor identity claims and the system has to be robust enough against unauthorized attempts to use the secret phrase of a registered speaker.

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 [3, 4] or from audio and lip information as in [5]. The speaker recognition (identification and/or verification) schemes proposed in [3, 5] are basically opinion fusion techniques that combine multiple expert decisions through adaptive or non-adaptive weighted summation of scores, whereas in [4], fusion is carried out at feature-level by concatenating the individual feature vectors so as to exploit the temporal correlations that may exist between audio and video signals.

In this paper, we propose a new adaptive classifier combining rule for fusion of multiple modalities for the text-dependent open-set speaker identification problem. The proposed rule uses an ordered cascade of classifiers each of which corresponds to a single modality or some linearly fused combination of modalities. The order of the classifiers in the cascade is adaptive and based on the reliability estimation of each modality so that the goal is at least not to fail whenever one of the classifiers gives the correct accept or reject decision. A new reliability measure is also proposed to assess decisions of a classifier under both reject and accept scenarios.

2. Speaker Identification

The speaker identification problem is often formalized by using a probabilistic approach: Given a feature vector \( f \) representing the sample data of an unknown individual, compute the a posteriori probability \( P(\lambda_n | f) \) for each class \( \lambda_n, n = 0, 1, ..., N \), i.e. for each speaker’s model. The sample feature vector is then assigned to the class \( \lambda_n \) that maximizes the a posteriori probability or equivalently the class-conditional probability:

\[
\lambda_n = \arg \max_{\lambda_n} P(f | \lambda_n)
\]  

(1)

In open-set speaker identification, a reject mechanism is also required due to possible impostor identity claims. A possible reject strategy is to refer a reject (impostor) class \( \lambda_0 \), so that a likelihood ratio \( \rho(\lambda_0) \) in logarithmic domain is used for accept or reject decision:

\[
\rho(\lambda_n) = \log \frac{P(f | \lambda_n)}{P(f | \lambda_0)} = \log P(f | \lambda_n) - \log P(f | \lambda_0)
\]  

(2)
Ideally, the impostor class model should be constructed by using all possible impostor observations for class \( n \), which is practically unfeasible to achieve. A common and effective approximation is to use the universal background model, which is estimated by using all available training data regardless of which class they belong to. The final decision strategy can then be stated as follows:

\[
\begin{align*}
\text{if } \rho(\lambda_n) \geq \tau & \quad \text{accept} \\
\text{otherwise} & \quad \text{reject}
\end{align*}
\]

where \( \tau \) is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate.

When more than one information source is available as in the case of multimodal speaker identification problem, the fusion of information from different sources can reduce overall uncertainty and increase the robustness of a classifier. One of the most generic way of computing joint ratios (or scores) can be expressed as a weighted summation:

\[
\rho(\lambda_n) = \sum_{p=1}^{P} \omega_p \rho_p(\lambda_n) \quad \text{for} \quad n = 1, 2, ..., N,
\]

where \( \omega_p \) values are weighting coefficients such that \( \sum_p \omega_p = 1 \). Then the fusion problem becomes finding the optimal choice of these coefficients. Most of the classifier fusion schemes existing in the literature [6] vary actually in the way they interpret the weighting coefficients in Eq. 4. On one side, there are hard-level combination techniques such as max rule, min rule and median rule [6], 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. 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 \( R_p \). 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 [3, 4, 5]. We will refer to this combination method as RWS (Reliability Weighted Summation) rule.

The impostor model, i.e. \( P(F|\mu)\lambda_n \), is a mere approximation of what it is in 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. In this paper a sigmoid normalization is used as in [3], which maps likelihood ratios to the \((0, 1)\) interval by normalizing the likelihood ratio \( \rho \) using the function \( g(\rho) \):

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

where \( \mu \) and \( \sigma \) are the mean and the standard deviation of the likelihood ratio \( \rho \) over the accept subjects, respectively.

### 3. Multimodal Decision Fusion

In this section, we start with the description of a slightly modified version of the so-called max rule that will give us an insight to develop our fusion scheme based on an adaptive classifier cascade. The proposed fusion scheme will eventually compromise the soft-level and the hard-level classifier combination strategies so as to yield a better assignment of the weighting coefficients in Eq. 4.

#### 3.1. Max Rule

The conventional max rule sets the coefficient \( \omega_p \) to 1 in Eq. 4 when the likelihood ratio of the \( p \)-th classifier is the maximum among all classifiers, and all other coefficients are set to zero.

The max rule may filter out some of the erroneous contributions in the final decision; however, the max rule is not appropriate to be used for the reject scenario, i.e. to detect impostors; it does not adequately take into account strong reject decisions and tends to yield a high false accept rate. The conventional max rule can be modified so as to handle the possible false identity claims. In this slightly modified scheme, the likelihood ratio is substituted with a more appropriate measure. Looking back to Eq. 3, once a threshold \( \tau \) is set in the log-likelihood ratio test, one can claim that if the best likelihood score \( \rho_p(\lambda_n^C) \) from modality \( p \) is much larger or much smaller than \( \tau \), respectively the confidence of the accept or reject decision is stronger. Hence the absolute difference between the likelihood score \( \rho_p(\lambda_n^C) \) and the threshold \( \tau \) can be used as a measure of accept-reject likelihood:

\[
C_p = |\rho_p(\lambda_n^C) - \tau|,
\]

where \( C_p \) will be referred to as the likelihood offset of the decision of the \( p \)-th classifier. Then the modified max rule uses the following assignment for the weighting parameters:

\[
w_p = \begin{cases} 
1 & \text{if } p = \arg \max_i C_i; \\
0 & \text{otherwise}
\end{cases}
\]

In this manner, a strong reject decision can also be taken into account and favored even though the corresponding likelihood score is not the maximum of the best likelihoods resulting from all available modalities.

#### 3.2. Adaptive Cascade Rule

In the RWS decision rule, when the numerical estimation of reliability values is not very accurate, the erroneous likelihood scores contribute to the joint score, corrupting correct partial decisions. Our objective is now to define a compromising rule that addresses all these problems associated with the max and RWS rules, inheriting the merits of each of them. 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 to be taken by only one classifier. 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. In this case, a better option is to incorporate reliability weighted modality combinations into the decision scheme.

The proposed rule will be referred to as adaptive cascade rule as it employs a cascade of classifiers in which all \( P \) classifier combinations are adaptively ordered based on their reliability estimates. Suppose for the moment that a reliability value \( R_p \), \( 0 \leq R_p \leq 1 \), is available for each classifier along with the likelihood offset \( C_p \) of that decision. 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 classifier. Then starting with the foremost, i.e. the most reliable classifier \( p_1 \), the cascade rule...
iteratively searches for a decision with a sufficiently high likelihood offset. As soon as a classifier \( p_1 \) with sufficiently high likelihood offset is encountered, the decision cascade is concluded with accept or reject decision. Note that the adaptive cascade rule uses \( P \) levels of decision, which creates a heavy computational load for determining the equal error rate for large \( P \). A reduction is obtained by adaptive selection of the most reliable \( \hat{P} \) classifiers depending on the reliability order that varies from one decision instant to another. Setting \( \hat{P} = 3 \) is usually sufficient for determination of the equal error rate and the three corresponding thresholds \( \tau, \tau_1 \) and \( \tau_2 \) with a reasonably low complexity. For \( \hat{P} = 3 \), the adaptive cascade algorithm can be described at three steps as follows:

i. A decision (accept or reject) is taken according to the classifier \( p_1 \) with the most reliable modality if the likelihood offset \( C_{p_1} \) is high enough (i.e. if \( C_{p_1} > \tau_1 \)).

ii. Else if \( C_{p_2} > \tau_2 \), a decision is taken according to the classifier with the highest likelihood offset among the classifiers with the two most reliable modalities \( p_1 \) and \( p_2 \).

iii. Else, a decision is taken according to the classifier with the highest likelihood offset among the classifiers with the three most reliable modalities.

### 3.3. Estimation of Modality Reliability

One of the main approaches in the speaker identification literature for adaptive estimation of the reliability of a modality is to analyze directly the statistics and the rank correlation of the resulting likelihood scores. Reliability estimates based on this approach might have accuracy problems; but the estimates are computationally feasible and general, addressing all kinds of possible corruptions.

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 as in [5]. Let \( \rho_p(\lambda_+) \) and \( \rho_p(\lambda_+) \) denote respectively the best and the second best likelihood ratios resulting from the \( p \)-th classifier. Then the associated likelihood ratio difference, \( \Delta_p \), is defined as,

\[
\Delta_p = \rho_p(\lambda_+) - \rho_p(\lambda_+). \tag{8}
\]

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

\[
R_p = \frac{1}{\sum_i \gamma_i} \gamma_p, \tag{9}
\]

where

\[
\gamma_p = \left( e^{(\rho_p(\lambda_+)-\Delta_p)} - 1 \right) + \left( e^{(\rho_p(\lambda_+)-\Delta_p)} - 1 \right). \tag{10}
\]

The first and second terms in \( \gamma_p \) are associated with the true accept and true reject, respectively, and \( \kappa \) is a factor that sets the relative weight of the true reject contribution of the reliability measure \( R_p \). Hence the reliability measure \( R_p \) increases when there is an evidence of reliability either for true accept or true reject, otherwise stays at low levels. It will later be verified by experiments that when the proposed reliability measure is employed in a non-uniform weighted summation scheme for fusion of multiple modalities, it yields superior results as compared to equal weighting.

### 4. Experimental Results

The proposed multimodal speaker identification system has been tested on the audio-visual database MVGL-AVD [7]. 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 database is partitioned into two equal sets in two different ways, so that four different and independent training and testing sessions are deployed.

Audio, lip and face modalities are considered in the multimodal speaker identification system, where audio (A), lip (L) and audio-lip multi-stream (AL) modalities are characterized using HMM structures and face modality (F) is characterized using eigenface method as in [7]. 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 is composed of 13 cepstral coefficients using 26 mel frequency bins. The resulting audio feature vector of size 39 includes the MFCC vector together with the first and second delta MFCC vectors. Each lip stream is extracted by cropping \( 64 \times 40 \) lip frames to form the lip sequence of each secret phrase utterance. The gray scale lip stream is transformed into 2D-DCT domain and then each lip frame is represented by the first 60 DCT coefficients of the zig-zag scan excluding the dc-term. The stream weights are picked respectively as 0.7 and 0.3 for the audio stream and the lip stream in the multi-stream HMM structure.

Similarly, face image streams are extracted and the eigenface technique is implemented as in [7] with an eigenspace of dimension 20 using a collection of face images that includes two face images from each utterance in the training part of the MVGL-AVD database.

The identification results are shown in Table 1, 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, we have to point out that the images in the training and testing set have varying backgrounds and lighting; this is why the face-only identification performance may seem to be worse than expected. The lip and audio-lip modalities yield a decent equal error rate performance; but they are still not satisfactory to be used directly as a single modality. 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.

The performance of the reliability weighted summation (RWS) rule can be compared with the so-called product rule.
Table 1: Speaker identification results: equal error rates for different decision fusion techniques (Product rule +, RWS rule ⊕, Modified max rule *, Adaptive cascade rule •, Non-adaptive cascade rule ○ (fixed a priori reliability ordering, where leftmost being the most reliable modality)) at varying noise levels.

<table>
<thead>
<tr>
<th>Source Modality</th>
<th>Noise Level (dB SNR)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clear 25 30 35 40 45 50</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>2.4 2.5 3.7 6.1 11.0 18.9 26.5</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>18.9</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>13.5 13.8 13.8 13.8 14.9 15.3 15.4</td>
<td></td>
</tr>
<tr>
<td>A + F</td>
<td>1.9 2.0 2.6 3.8 6.6 8.3 11.9</td>
<td></td>
</tr>
<tr>
<td>A ⊕ F</td>
<td>0.5 1.0 1.1 1.2 1.2 2.2 3.2</td>
<td></td>
</tr>
<tr>
<td>A ⋆ F ⊕ AL</td>
<td>1.1 1.1 1.1 1.5 1.9 2.5 2.8</td>
<td></td>
</tr>
<tr>
<td>A + F ⊕ AL</td>
<td>0.7 0.7 0.7 1.0 1.4 2.2 3.3</td>
<td></td>
</tr>
<tr>
<td>A ⊕ F ⊕ AL</td>
<td>1.8 1.8 2.6 4.9 6.4 7.0 9.0</td>
<td></td>
</tr>
<tr>
<td>A + F ⊕ AL</td>
<td>2.3 2.2 2.5 3.3 5.3 7.4 12.6</td>
<td></td>
</tr>
<tr>
<td>A ⊕ F ⊕ AL</td>
<td>1.4 1.5 1.7 1.9 2.5 3.9 6.3</td>
<td></td>
</tr>
<tr>
<td>A ⋆ F ⊕ AL</td>
<td>0.9 0.9 0.8 1.2 2.5 4.3 7.5</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>0.7 0.7 0.7 1.0 1.4 2.5 3.5</td>
<td></td>
</tr>
<tr>
<td>M1 = (A ⊕ F)</td>
<td>0.5 1.0 1.1 1.2 3.2 5.2 8.2</td>
<td></td>
</tr>
<tr>
<td>M2 = (F ⊕ AL)</td>
<td>2.9 3.0 3.2 3.4 3.6 3.8 3.9</td>
<td></td>
</tr>
<tr>
<td>M3 = M1 ⊕ M2 ⊕ A F</td>
<td>0.4 0.6 0.6 1.0 1.6 1.9 2.6</td>
<td></td>
</tr>
<tr>
<td>M4 = A ⋆ F</td>
<td>0.3 0.3 0.4 0.5 1.2 2.1 4.5</td>
<td></td>
</tr>
<tr>
<td>M5 = M1 ⋆ A F</td>
<td>0.2 0.2 0.3 0.4 1.1 2.1 4.3</td>
<td></td>
</tr>
<tr>
<td>M6 = M1 ⋆ A F</td>
<td>0.2 0.2 0.3 0.6 1.0 1.6 2.3</td>
<td></td>
</tr>
<tr>
<td>M7 = M1 ⋆ A F</td>
<td>0.2 0.2 0.3 0.6 1.0 1.6 2.3</td>
<td></td>
</tr>
</tbody>
</table>

in the second split of Table 1. RWS rule assumes the proposed reliability measure in Eq. 9 with an optimal weighting factor $\kappa = 0.65$, which is found experimentally to minimize the average EER figure. 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 multi-stream audio-lip modalities.

Adaptive classifier cascade sets a reliability order for each modality using the proposed modality reliability with an optimal weighting factor $\kappa = 1.25$. Max, non-adaptive cascade and adaptive cascade rules are evaluated in the first three rows of the third split of Table 1 by combining audio, face and audio-lip streams. The performance of the adaptive cascade clearly out performs both max and fixed reliability ordered non-adaptive 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. 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 as expected as the multi-stream 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 $AL$ 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 as it also yields a poor reliability estimation. This finding leads us to examine the RWS modality combinations 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, $M_2$, $M_4$ and $M_6$ are considered in Table 1. Once these combined modalities are adaptively cascaded with relatively reliable unimodal streams, i.e. audio and face (the last row of Table 1), 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.

5. Conclusions

We have presented an adaptive classifier cascade fusion for multimodal (audio-lip-face) speaker identification system that improves the identification performance over traditional fusion schemes, such as the product rule and the max rule. Also, a novel modality reliability estimation scheme is proposed, that is based on the likelihood ratio stream and that differentiates the best likelihood ratio score from the rest of the scores, creating a relative assessment measure on the reliability of true accept and true reject decisions. We experimentally show 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. 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 are presented are encouraging for robust multimodal speaker identification systems.

6. References