COST278 and ISCA Tutorial and Research Workshop
on Robustness Issues in Conversational Interaction
University of East Anglia, Norwich, UK
August 30-31, 2004

Multimodal Speaker Identification using Adaptive Decision Fusion with Reliability Weighted Summation

Engin Erzin, Yücel Yemez, A. Murat Tekalp

Multimedia, Vision and Graphics Laboratory
College of Engineering, Koç University, Sarıyer, Istanbul, 34450, Turkey
eerzin,yyemez,mtekalp@ku.edu.tr

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, the so called product rule with a novel adaptive reliability based weighting structure is employed. The proposed adaptive product rule is more robust in the presence of unreliable modalities, provided that the employed reliability measure is effective in assessment of classifier decisions. The proposed reliability measure, that genuinely fits to the open-set speaker identification problem, is used to assess more robust accept and reject 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 decision fusion system with a new adaptive reliability measure for the text-dependent open-set speaker identification problem. The new reliability measure 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 $\boldsymbol{f}$ representing the sample data of an unknown individual, compute the a posteriori probability $P(\lambda_n|\boldsymbol{f})$ for each class λ_n, $n = 0, 1, ..., N$, i.e. for each speaker's model. The sample feature vector is then assigned to the class λ_* that maximizes the a posteriori probability or equivalently the class-conditional probability:

$$\lambda_* = \arg\max_{\lambda_n} P(\boldsymbol{f}|\lambda_n) \quad (1)$$

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 $\lambda_{\bar{n}}$, so that a likelihood ratio $\rho(\lambda_n)$ in logarithmic domain is used for accept or reject decision:

$$\rho(\lambda_n) = \log \frac{P(\boldsymbol{f}|\lambda_n)}{P(\boldsymbol{f}|\lambda_{\bar{n}})} = \log P(\boldsymbol{f}|\lambda_n) - \log P(\boldsymbol{f}|\lambda_{\bar{n}}) \quad (2)$$

Ideally, the impostor class model should be constructed by using all possible impostor observations for class n, which is practically infeasible 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:

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

where τ 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 classification system. One of the most generic way of computing joint ratios (or scores) can be expressed as a weighted summation:

$$\rho(\lambda_n) = \sum_{p=1}^{P} \omega_p \rho_p(\lambda_n) \quad \text{for } n = 1, 2, ..., N, \quad (4)$$

where ω_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. $P(\boldsymbol{f}_p|\lambda_{\bar{n}})$, 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 ρ using the function $g(\rho)$,

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

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

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 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 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, Δ_p, is defined as,

$$\Delta_p = \rho_p(\lambda_*) - \rho_p(\lambda_{**}). \quad (6)$$

However in the presence of a reject class, Δ_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 Δ_p are evidences of a true accept decision, and alternatively a low likelihood ratio and a low Δ_p are evidences of a true reject decision. We propose a new reliability measure R_p based on these evidences as,

$$R_p = \frac{1}{\sum_i \gamma_i} \gamma_p, \quad (7)$$

where

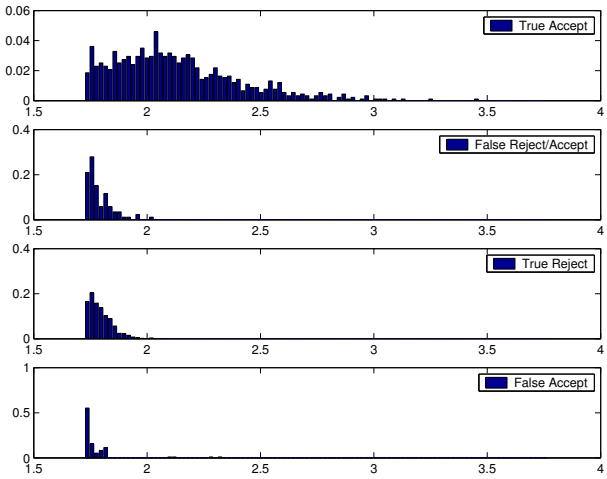
$$\gamma_p = (e^{(\rho_p(\lambda_*)+\Delta_p)} - 1) + (e^{(\kappa-\rho_p(\lambda_*)-\Delta_p)} - 1). \quad (8)$$

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 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. Figure 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. 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.

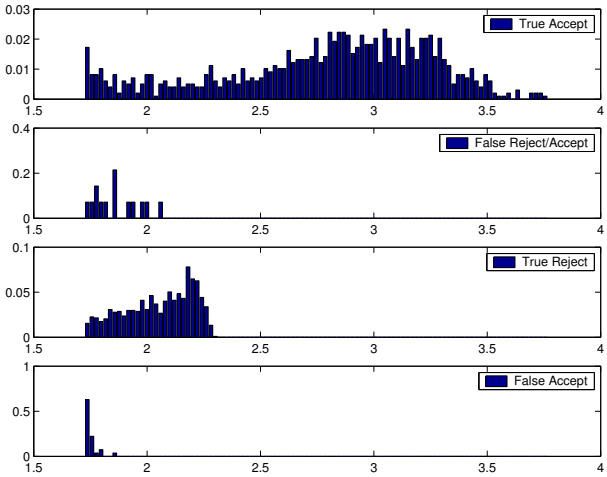
4. Experimental Results

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 720×576 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



(a) Audio only classifier with EER 6.1%.



(b) RWS rule for audio, face, and audio-lip modalities with EER 0.7%.

Figure 1: Histograms of the reliability measure R_p for different classifiers for accept (top two rows) and reject (bottom two rows) scenarios.

varying levels of additive noise during the testing sessions to simulate adverse environmental conditions. The additive acoustic noise is picked to be car 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 64×40 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 unimodal identification results are shown in Table 1, where we observe the equal error rates at varying levels of acoustic car 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.

Table 1: *Unimodal and audio-lip speaker identification results: equal error rates at varying noise levels.*

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	20	10	0	-5	-10	-15
A	2.4	2.4	2.5	3.7	5.6	12.1	29.5
F	8.4						
L	18.0						
AL	13.6	13.6	13.6	13.8	14.0	14.8	15.3

Table 2: *Speaker identification results: equal error rates for for product rule ($+$) and RWS rule ($\oplus$) at varying noise levels.*

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	20	10	0	-5	-10	-15
A + F	1.9	1.9	2.1	2.6	3.7	5.6	13.0
A $\oplus$ **F**	0.4	0.4	0.4	0.9	1.4	3.0	8.6
A + F + AL	1.0	1.0	1.0	1.1	1.2	1.5	3.9
A $\oplus$ **F** $\oplus$ **AL**	0.6	0.6	0.6	0.7	1.0	1.2	3.5

The performance of the reliability weighted summation (RWS) rule can be compared with the so-called product rule, that is the summation of log-likelihoods with uniform weighting, in Table 2. RWS rule assumes the proposed reliability measure in Eq. 7 with an optimal weighting factor $\kappa = 0.65$, which is found experimentally to minimize the average EER figure. 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, 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 multi-stream audio-lip modalities. Note that, when the multi-stream audio-lip modality is fused with the (**A** $\oplus$ **F**) combination by RWS rule, that is **A** $\oplus$ **F** $\oplus$ **AL**, the performance significantly improves over all SNR range except for high SNR conditions. This performance improvement is expected as the multi-stream audio-lip modality sustains fairly robust performance, which is partially uncorrelated from the audio-only performance under noisy conditions.

The RWS modality combinations that have better EER performances than the unimodal streams, are expected to estimate better reliability measures than the unimodal reliability estimates. Three such combined modalities, $\mathbf{M}_0$, $\mathbf{M}_1$ and $\mathbf{M}_2$ are considered in Table 3. Once these combined modalities are adaptively fused with relatively reliable unimodal streams, i.e. audio and face (the last row of Table 3), 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 RWS decision fusion.

Table 3: *Speaker identification results: equal error rates for for RWS rule over combined modalities at varying noise levels.*

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	20	10	0	-5	-10	-15
$\mathbf{M}_0 = (\mathbf{A} \oplus \mathbf{F} \oplus \mathbf{AL})$	0.6	0.6	0.6	0.7	1.0	1.2	3.5
$\mathbf{M}_1 = (\mathbf{A} \oplus \mathbf{F})$	0.4	0.4	0.4	0.9	1.4	3.0	8.6
$\mathbf{M}_2 = (\mathbf{F} \oplus \mathbf{AL})$	3.2	3.2	3.2	3.3	3.4	3.7	3.9
$\mathbf{M}_0 \oplus \mathbf{M}_1 \oplus \mathbf{M}_2 \oplus \mathbf{A} \oplus \mathbf{F}$	0.4	0.4	0.4	0.5	0.8	1.3	3.0

5. 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 RWS decision fusion for multimodal (audio-lip-face) speaker identification system improves the identification performance over traditional fusion schemes. The novel modality reliability estimation is based on the likelihood ratio stream and it 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 RWS rule outperforms the product rule, provided that the employed reliability measure is effective enough in assessment of classifier decisions. Another important feature of this work is the use of combined modalities in the decision fusion scheme. Some modality combinations, obtained via RWS rule, may achieve much better EER performances than the single modalities; such combined modalities can be considered as additional reliable sources for boosting the decision fusion. The experimental findings support that the adaptive RWS fusion 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. Acknowledgments

This work has been supported by TUBITAK under the project Cost278-EEEAG-101E026 and by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).

7. References

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

[2] Y. Y. J. Zhang and M. Lades, "Face recognition: Eigenface, elastic matching, and neural nets," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1423–1435, September 1997.

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

[4] 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," *Proc. of the Int. Conf. on Multimedia & Expo 2003 (ICME2003)*, vol. 3, pp. 9–12, July 2003.

[5] T. Wark and S.Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, vol. 11, no. 3, pp. 169–186, July 2001.

[6] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[7] E. Erzin, Y. Yemez, and A. M. Tekalp, *DSP in Mobile and Vehicular Systems*. Kluwer Academic Publishers, forthcoming 2004, ch. Joint Audio-Video Processing for Robust Biometric Speaker Identification in Car.