

# INFORMATION FUSION AND DECISION CASCADING FOR AUDIO-VISUAL SPEAKER RECOGNITION BASED ON TIME-VARYING STREAM RELIABILITY PREDICTION

*Upendra V. Chaudhari, Ganesh N. Ramaswamy, Gerasimos Potamianos, and Chalapathy Neti*

IBM T.J. Watson Research Center  
Rt. 134, Yorktown Heights, NY 10598  
Email: {uvc,ganeshr,gpotam,cneti}@us.ibm.com

## ABSTRACT

We examine techniques for multi-modal biometric information fusion for verification and identification of speakers, where the reliability of each data stream, either audio or video, is modeled with parameters that are time-varying and depend on the context created by its local behavior. The complementary nature and the time dependent relative reliability of audio and video data is studied in the context of verification and identification, on data collected during a user's interaction with an automated system. Of significance is that this data is not corrupted artificially. Particular focus is directed to verification and its ability to refine identification decisions, by indicating a level of confidence in the system decisions. Results show more striking effects for verification, when using time-dependent fusion, than for identification.

## 1. INTRODUCTION

As systems become increasingly automated with a continuously growing number of users, monitoring of that usage becomes critical to their secure and efficient operation. Secure operation requires that a user is who they say they are and efficient operation is aided by knowing the identity of a user even when an identity claim is not necessary. Personalization in a non-secure environment is one example of the latter case. Since automated agents are becoming more ubiquitous, they are becoming situated in more varied environments that can change both slowly and rapidly with time, for example as with a kiosk. Precise and accurate identification and verification of users requires that this time-varying nature of the environment be taken into account.

Not only are systems becoming more automated, but they are also becoming more complex with a variety of input modalities. Audio and video are perhaps the two most common and complementary modes of input and are the focus of this investigation where verification and identification are studied in the framework of a user sitting in front of a computer which is collecting both audio and video data simultaneously.

The main focus is verification, in isolation and on its ability to refine identification decisions. Typically, the best match from a database of users is chosen as the id for a test sample. If subsequently, a verification of the returned id is attempted, there is a the possibility of rejecting the decision, thus allowing decisions to be inconclusive and to act as flags for further processing or out of population test data.

The information fusion methodology is dependent on properties of the test data in each stream alone, as well as on its relationship to training data for the models against which it is evaluated. These properties will be captured parametrically in the form of a measure of test data deviation and test data coverage. Both of these are measured in a local neighborhood for a point in time. Together, they indicate the reliability of the test data for any given stream. Thus, by comparing the relative values across different streams, a dynamic weighting procedure can be developed.

In exploring the above, we highlight the complementary nature of audio and video as well as to show that a benefit can be gained by considering multiple streams of data, even when they are not complementary, as when one stream is temporarily corrupted. That is, in general, it is reasonable to assume that one modality of operation, say audio only, may not sufficient to uniquely identify an individual in all environments. Furthermore, it is reasonable to assume that any given modality of operation will vary in quality as the environment changes over time. The complementarity of audio and video addresses the uniqueness point, and the ability to trade off between the two addresses the robustness to a time varying environment.

## 2. MULTIPLE DATA STREAMS

Since the purpose is to study both the complementary and correlative nature of the audio and video streams, we define three streams of interest:  $\mathbf{X}^a = \{\mathbf{x}_t^a\}$  (audio),  $\mathbf{X}^v = \{\mathbf{x}_t^v\}$  (video), and  $\mathbf{X}^{av} = \{\mathbf{x}_t^{av}\}$  (A/V vector-wise concatenation). The audio data consist of 23 dimensional MFCCs with mean subtraction applied. No delta parameters are

used, and C0 is ignored. The visual data is derived from an appearance based technique. Features are extracted from a 2-D, separable, discrete cosine transform (DCT) applied to a region of interest defined for each video frame, by a statistical face tracking algorithm [5]. The 24 DCT coefficients with the highest energy, as determined during a training phase, are retained. Mean normalization is applied to compensate for lighting variations. As with audio, no delta parameters are used. Short-Time Gaussianization [8] is used to process both the audio and video features. It attempts to mitigate the effects that linear channel and additive noise distortions have on the data streams. This is achieved by mapping the features to the standard normal distribution in localized windows.

### 3. AUDIO-VISUAL SPEAKER MODELS

#### 3.1. Background Models

For both identification and verification, speaker modeling is based on the Gaussian Mixture Model [7] (GMM) framework. Consider first, the creation of background models which characterize the entire effective space of features, i.e. without specificity to any individual. For data stream  $s$ , a background model  $M_s^{BG}$  is trained on a large pool of data from many individuals. It is parameterized by  $\{\mathbf{m}_{s,i}^{BG}, \Sigma_{s,i}^{BG}, p_{s,i}^{BG}\}_{i=1,\dots,N_s^{BG}}$ , where  $\mathbf{m}$ ,  $\Sigma$ , and  $p^{BG}$ , are the maximum likelihood estimates of the mean, covariance, and mixture weight parameters respectively and  $N_s^{BG}$  indicates the number of Gaussian components. The same basic parameterization applies to all models. Note however, that for scoring, only the diagonal portion of the covariance matrices are used due to data constraints.

#### 3.2. Identification Models

For identification, each speaker model is first created as with the background model resulting in, for each speaker  $j$  and stream  $s$ :  $\{\mathbf{m}_{s,i}^j, \Sigma_{s,i}^j, p_{s,i}^j\}_{i=1,\dots,N_s^j}$ . It was observed, however, that using a subsequent feature space transformation improves performance [1]. The Maximum Likelihood Linear Transformation (MLLT) [3] represents a feature space optimization which depends on the initial parameterization of the model and seeks the best linear transform, in the maximum likelihood sense, in which to diagonalize the given model. A gradient descent procedure is used to search for an MLLT transformation  $\mathbf{T}_s^j$  (note the dependence on the speaker and stream) that minimizes the loss in likelihood [3] that would result when the input model covariances are diagonalized. The new parameterization of the model includes this transformation.  $M_s^j = \mathbf{T}_s^j M_s^j \equiv \{\mathbf{m}_{s,i}^j, \Sigma_{s,i}^j, p_{s,i}^j\}_{i=1,\dots,N_s^j}$ , where  $\mathbf{m}_{s,i}^j = \mathbf{T}_s^j \mathbf{m}_{s,i}^j$  and  $\Sigma_{s,i}^j = \text{diag}(\mathbf{T}_s^j \Sigma_{s,i}^j \mathbf{T}_s^{j,\top})$ .

### 3.3. Verification Models

For verification, the speaker models are created via Maximum A posteriori Probability (MAP) adaptation from the background model [6]. For speaker  $j$  and stream  $s$ ,  $M_s^{BG}$  is used as the background. Then, the training data for speaker  $j$  in stream  $s$  is used to update the model parameters. Sufficient statistics are calculated for the speaker training data and the new model parameters are a linear combination of the background model parameters and sufficient statistics. Again, note that diagonal models are used. Here, only the means are adapted. Then, an MLLT is computed for each adapted model as for id.

## 4. RECOGNITION DECISIONS

Both verification and identification involve scoring the test data against stored models. Denote the test data in stream  $s$  by  $\mathbf{X}^s = \{\mathbf{x}_t^s\}$ , with vectors in  $R^n$ .

#### 4.1. Identification Decisions

The basic score unit for identification is the maximum value of the likelihood of a vector computed over all components of the model being scored, say  $M_s^j$ :

$$d_{j,s,t}^{ID} = \max_i \left[ \log p(\mathbf{T}_s^j \mathbf{x}_t^s | \mathbf{m}_{s,i}^j, \Sigma_{s,i}^j, p_{s,i}^j) \right], \quad (1)$$

Identification involves selecting the model in a database that best matches a test sample, whereas verification naturally seen as an hypothesis test, where an identity claim is provided and the goal is to determine whether the claim is true or false.

#### 4.2. Verification Decisions

For verification, the basic score unit above must be modified:

$$d_{j,s,t}^{VER} = \max_i \left[ \log p(\mathbf{T}_s^j \mathbf{x}_t^s | \mathbf{m}_{s,i}^j, \Sigma_{s,i}^j, p_{s,i}^j) \right] - \max_i \left[ \log p(\mathbf{T}_s^{BG} \mathbf{x}_t^s | \mathbf{m}_{s,i}^{BG}, \Sigma_{s,i}^{BG}, p_{s,i}^{BG}) \right], \quad (2)$$

This can be seen as a normalization of the claimed model score with respect to the background.

## 5. DISCRIMINANTS

The final discriminants for identification are a function of  $\{d_{j,s,t}^{ID}\}$  where  $j$  ranges over all speakers,  $s$  ranges over all streams, and  $t$  ranges over the number of vectors in the test data. The final discriminants for verification are a function of  $\{d_{j,s,t}^{VER}\}$  where  $j$  is equal to the claimed id, and  $s$  and  $t$  are as for id.

To accomplish time-varying fusion, the total discriminant has the following form (for identification and verification respectively):

$$D^{ID}(\mathbf{X}|j) = \sum_t \sum_s [\Phi_t^s(j) + \Psi_t^s(j)] \eta_s d_{j,s,t}^{ID}. \quad (3)$$

$$D^{VER}(\mathbf{X}|j \equiv \text{claim}) = \sum_t \sum_s [\Phi_t^s(j) + \Psi_t^s(j)] \eta_s d_{j,s,t}^{VER}, \quad (4)$$

### 5.1. Time-Context Dependent Parameters

$\Phi_t^s(j)$  and  $\Psi_t^s(j)$  are parameters that depend on the time, stream, and model. They encapsulate the reliability based tradeoff under investigation. They are normalized versions of  $\phi_t^s(j)$  and  $\psi_t^s(j)$ : The parameter  $\phi_t^s(j)$  is a time dependent measure of coverage, or how well data in a local time window predicts the parameters of the model being tested [2]. The parameter  $\psi_t^s(j)$  measures the deviation of the score stream at time  $t$  from the target value which is taken to be a point estimate of the score at time  $t$  [2].  $\Phi$  and  $\Psi$  are defined as follows:

$$\Phi_t^s(j) = \phi_t^s(j) / \sum_{q \in \{a,v\}} \phi_t^q(j),$$

$$\Psi_t^s(j) = (1/\psi_t^s(j)) / \sum_{q \in \{a,v\}} (1/\psi_t^q(j)).$$

This normalization achieves the context dependence that is sought. The relative values of the parameters are important in determining the time-dependent weighting. Since  $\Psi$  should be inversely proportional to the deviation, the reciprocal of  $\psi_t^s(j)$  is used. The  $\eta_s$  parameter normalizes for the scale differences in the score streams.

The identification decision is given by computing equation 3 for each speaker  $j$ , and choosing

$$id = \arg \max_j D^{ID}(\mathbf{X}|j).$$

The verification decision is binary, either accept or reject the claim. It is made by computing equation 4 for the claimed speaker, say  $j$ , and comparing the value against a threshold. The claim is accepted if the following condition is met:

$$D^{VER}(\mathbf{X}|j \equiv \text{claim}) > \tau.$$

Errors can occur as false accepts (alarms) or false rejects (misses) and both are a function of  $\tau$ . Results are thus typically presented as an ROC (receiver operating characteristic) curve.

## 6. INCONCLUSIVE DECISIONS

In many respects, Verification and Identification are complementary. Verification can be used to gauge the confidence of an identification decision. Assume that  $j_0 =$

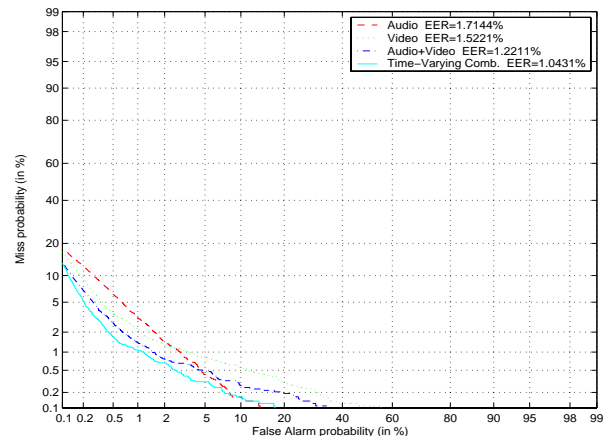


Figure 1: ROC curve.

$\arg \max_j D^{ID}(\mathbf{X}|j)$  for test data  $\mathbf{X}$ . Then  $j_0$  can be set to the claimed id for a subsequent verification decision. If  $D^{VER}(\mathbf{X}|j_0) > \tau$  then the identification decision is accepted, otherwise the decision is considered to be inconclusive. Thus, instead of simply a correct identification rate, one can in addition report an inconclusive decision rate. There is a tradeoff between these two rates, as well as the error rate. For a given identification system, the goal is to increase the inconclusive decision rate, while keeping the correct id rate constant (it cannot increase with this procedure).

## 7. EXPERIMENTS

Identification and Verification experiments were based on an audio-visual database consisting of data captured by a microphone and camera with users situated in front of a computer and reading prompted text. The database consists of 304 speakers, of which 100 were selected as targets, with the rest being used to model the background space for verification. Each speaker model was created using 2 minutes of training data with test data typically ranging from 3 to 10 seconds. All experiments were conducted at the frame level. The total number of tests was 7307 for identification and 730700 for verification, where every speaker excluding the correct one was used as an imposter for every cell.

### 7.1. Verification

Figure 1 shows the verification results in the form of an ROC curve for the cases where the three streams  $\mathbf{X}^a$ ,  $\mathbf{X}^v$ , and  $\mathbf{X}^{a,v}$  are used in isolation (there is no score combination) as well as when using the time and context dependent weights ( $\Phi$ & $\Psi$ , Time Varying Comb). Recall that  $\mathbf{X}^{a,v}$  is the concatenation of  $\mathbf{X}^a$  and  $\mathbf{X}^v$ .

Config	Error
Audio, $\mathbf{X}^a$	2.01%
Video, $\mathbf{X}^v$	10.95%
Audio+Video, $\mathbf{X}^{av}$	9.28%
$\Psi \& \Phi$	0.40%

Table 1: Original identification error rates on A/V multi-stream data.

Config	Identification Rate Type		
	Correct	Inconclusive	Error
Audio, $\mathbf{X}^a$	97.94%	0.23%	1.83%
Video, $\mathbf{X}^v$	88.74%	0.85 %	10.41%
Audio+Video, $\mathbf{X}^{av}$	90.61%	0.38%	9.01%
$\Psi \& \Phi$	99.25%	0.40 %	0.35%

Table 2: Identification followed by Verification.

The time varying combination has an equal error rate (EER, where the false alarm probability equals the miss probability) of 1.04% which is a 39% relative improvement over verification with audio alone. Also, it is interesting to note that the concatenation (Audio+Video) also does very well, with a 29% relative improvement.

## 7.2. Verifying Identification Decisions

Results given in tables 1 and 2 compare identification results with and without the post id verification step.

Here, the tradeoff between the error rates is seen clearly. Note that for identification, different models were used than for verification (see Section 3), leading to relative performance differences. Consider the performance for audio, where the original error rate was 2.01%. After verification, the error rate is 1.83% with a 0.23% inconclusive decision rate. But test data that is tagged with an inconclusive decision can be further processed, or more data could be collected to augment it before a final decision is made. Thus the error rate is achievable. Looking at the relative improvements, the biggest are for the audio alone and the time dependent combination.

## 8. CONCLUSION

We have investigated the complementary nature of audio and video data streams in the context of speaker recognition, focusing on verification and its ability to refine identification decisions. When verification performance alone is considered, the effects of the time-varying discriminants are more dramatic than for the identification performance

alone. This is all the more significant in that we use real world data, that has not been artificially corrupted. Combining the two reveals that the error rate can be reduced, by converting some errors to inconclusive decisions. However, there is an associated reduction in identification accuracy, the extent of which can be controlled by choosing different thresholds for the verification decision. Experiments show that this reduction can be less than the reduction in the error rate.

## 9. REFERENCES

- [1] U.V. Chaudhari, J. Navrátil, and S.H. Maes, Transformation Enhanced Multi-grained Modeling for Text-Independent Speaker Recognition. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, October 2000.
- [2] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Audio-Visual Speaker Recognition using Time-Varying Stream Reliability Prediction", to appear in *Proc. ICASSP*, Hong Kong, April 2003.
- [3] R. A. Gopinath, "Maximum Likelihood Modeling With Gaussian Distributions For Classification", *Proc. ICASSP*, Seattle, May 1998.
- [4] B. Maison, C. Neti, and A. Senior, "Audio-Visual Speaker Recognition for Video Broadcast News: Some Fusion Techniques", *IEEE Multimedia Signal Processing (MMSPP99)*, Denmark, Sept., 1999.
- [5] G. Potamianos, C. Neti, G. Iyengar, A. W. Senior, and A. Verma, "A Cascade Visual Front End for Speaker Independent Automatic Speechreading", *International Journal of Speech Technology*, 4(3-4):193-208, 2001.
- [6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Speaker Models", *Digital Signal Processing*, Vol. 10, Nos. 1-3, January/April/July 2000.
- [7] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, January 1995.
- [8] B. Xiang, U.V. Chaudhari, J. Navrátil, G. N. Ramaswamy, and R. A. Gopinath, "Short-Time Gaussianization for Robust Speaker Verification", *Proc. ICASSP*, Orlando, May 2002.