AUDIO-VISUAL SPEAKER RECOGNITION USING 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 a time-varying, context dependent information fusion methodology for multi-stream authentication based on audio and video data collected simultaneously during a user's interaction with a system. Scores obtained from the two data streams are combined based on the relative local richness, as compared to the training data or derived model, and stability of each stream. The results show that the proposed technique outperforms the use of video or audio data alone as well as the use of fused data streams (via concatenation). Of particular note, is that the performance improvements are achieved for clean, high quality speech, whereas previous efforts focused on degraded speech conditions.

1. INTRODUCTION

User authentication based on a speaker's interaction with any system is based on the information that the system collects. In general, it is possible and desirable to exploit multiple forms of information obtained by different transducers operating simultaneously. In particular, a significant benefit can be obtained by analyzing the correlation in the multiple data streams. More significantly, however, any subset of the data streams form a context for the analysis of any other subset of streams, allowing the formulation of a robust time-varying fusion methodology. In this exposition, audio and video data streams are used.

The focus is to develop a time varying approach to multi-stream analysis where in general the fusion of data, scores, and decisions occur locally in time at the stream element level and relative to a local context that includes a measure of data richness and data consistency. In the sequel, we will sometimes refer to a combination of these two properties as "reliability." It is believed that the reliability of each data stream varies with time as the interaction proceeds and performance gains can be achieved by using this knowledge intelligently.

As we show in this paper, even in the clean speech case, there exists a subset of the population for which audio based authentication is problematic and inconsistent. We provide and study methods for predicting the reliability of the data streams. It may turn out for any point in time that only audio, only video, or a combination of the two streams are used. And this combining process is time varying in that the reliability is modeled locally as a function of time and other signal properties.

That multi-stream analysis would be beneficial is no surprise given Doddington's study [2] where it is shown that there are speakers, termed "goats," who are difficult to recognize based on their voice. Speakers who are readily recognized based on voice are termed "sheep." In light of this, one may choose to make an a priori decision as to the efficacy of audio vs. video data for each individual and subsequently use only the data corresponding to the most effective modality. Using this approach however, does not leverage the fact that data quality for the two streams can vary independently over time, and the more different types of data collected, the better the chance that good data will be collected in at least one form. Another option is to model the joint statistics of the data streams. This approach uses both forms of the data, but since the joint statistics are used, if one data stream is noisy or otherwise poor, it can adversely affect the overall performance. The most flexible option is to create models independently for each data modality and combine scores and decisions from both. Previous studies [4] have focused on score combination at the test utterance level in the degraded speech case. We propose, and provide evidence for the hypothesis that the best approach is to model the two data streams independently and make an intelligent, time varying decision as to which model to use and/or how to combine scores for each point in time. The results presented herein are significant in that improvement in speaker recognition performance is obtained for the high quality, clean speech case (as evidenced by overall audio performance) by adding video stream data and performing a time-varying analysis.

2. FEATURE STREAMS

Simultaneous recordings of audio and video data are used to produce the three vector streams of interest: $\mathbf{X}^a = \{\mathbf{x}_t^a\}$

(audio), $\mathbf{X}^v = \{\mathbf{x}_t^v\}$ (video), and $\mathbf{X}^{av} = \{\mathbf{x}_t^{av}\}$ (vector-wise concatenation of audio and video streams), consisting for the audio of mean normalized, 23 dimensional, Mel frequency cepstral coefficient (MFCC) vectors (no C0) computed using 24 filters. Visual features are extracted using an appearance based technique [5]. For each video frame, a statistical face tracking algorithm is used to define a region of interest to which a 2-D, separable, discrete cosine transform (DCT) is applied. The 24 highest energy (over all training data) DCT coefficients are retained and mean normalization is applied to compensate for lighting variations. No delta parameters are used. The audio and video vector coefficients are further processed via short-time Gaussianization [7] which attempts to mitigate the effects on the mean and variance parameters of linear channel and additive noise distortions by locally mapping the features to the standard normal distribution.

3. SPEAKER MODELS

Speaker modeling is based on the Gaussian Mixture Model [6] (GMM) framework and the transformation based enhancements described in [1] which use feature space optimizations on top of the initial feature sets. These optimizations, via the Maximum Likelihood Linear Transformation (MLLT) [3], are conditioned on the models which must therefore be built before the optimization. For each data stream s, and speaker j, the N_s^j-component model, $M_s^{j,o}$, is parameterized, prior to the feature space optimization, by $\{\mathbf{m}_{s,i}^{j,o}, \mathbf{\Sigma}_{s,i}^{j,o}, p_{s,i}^j\}_{i=1,\ldots,N_s^j}$, consisting of the estimates of the mean, covariance, and mixture weight parameters. Restriction to diagonal covariance models occurs in a transformed feature space where an MLLT transformation $\mathbf{T}_s^j$ is chosen, via a gradient descent, to minimize the loss in likelihood that results from the restriction [3]. Consequently, the new model parameterization is $M_s^j = \mathbf{T}_s^j M_s^{j,o} \equiv \{\mathbf{m}_{s,i}^j, \mathbf{\Sigma}_{s,i}^j, p_{s,i}^j\}_{i=1,\ldots,N_s^j}$, where $\mathbf{m}_{s,i}^j = \mathbf{T}_s^j \mathbf{m}_{s,i}^{j,o}$ and $\mathbf{\Sigma}_{s,i}^j = diag(\mathbf{T}_s^j \mathbf{\Sigma}_{s,i}^{j,o} \mathbf{T}_s^{j,\top})$. Note that the feature space optimization is carried out independently for each speaker model and each feature stream. As a result, each speaker model has its own associated feature space.

4. DISCRIMINANTS WITH TIME-VARYING CONTEXT-DEPENDENT PARAMETERS

We use a modified likelihood based discriminant function that takes into account the added transformation. Given a set of vectors $\mathbf{X}^s = \{\mathbf{x}_t^s\}$ in R^n from some stream s, the base discriminant function for any individual stream dependent target model M_s^j is

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

where the index i runs through the mixture components in the model M_s^j and $p(\cdot)$ is a multi-variate Gaussian density. The multi-stream input to the identification system is $\mathbf{X} = \{\mathbf{X}^a, \mathbf{X}^v\}$, a set of two streams with N vectors in each.

4.1. Generalized Discriminant

We define the general discriminant with time varying parameters for an N frame input as ($t \in \{1, ..., N\}$ and $s \in \{a, v\}$)

$$D(\mathbf{X}|j) = \sum_t \sum_s \left[\Phi_t^s(j) + \Psi_t^s(j)\right] \eta_s d_s(\mathbf{x}_t^s | M_s^j), \quad (2)$$

or

$$D(\mathbf{X}|j) = \sum_t \sum_s \Phi_t^s(j) \Psi_t^s(j) \eta_s d_s(\mathbf{x}_t^s | M_s^j). \quad (3)$$

where $\Phi_t^s(j)$ and $\Psi_t^s(j)$ are time, stream, and model dependent parameters that measure the local congruence of the test data with the model and the stability of the score stream, and η_s normalizes the scale of the scores. Note that there is a product and sum form of the combination. The important point is that Φ measures the match of the test data to the models and Ψ measures the predictability of the score stream. They are the normalized parameters, as defined in section 4.1.3, derived from $\phi_t^s(j)$ and $\psi_t^s(j)$.

4.1.1. Coverage: $\phi_t^s(j)$

To determine $\phi_t^s(j)$, which is a measure of the coverage of the model by the test data, we invert the roles of the test and training data and compute what we call an "inverse" likelihood. That is, for a time t, we model a neighborhood (in time) of the test vector by a GMM, $M_{s,t}^{test}$, and measure the likelihood of the model parameters M_s^j, and/or training data (for M_s^j), w.r.t. the test data model in computing the parameter. In its generalized form, the equation is:

$$\phi_t^s(j) = \sum_i \alpha_{s,i}^j d_s(\mathbf{m}_{s,i}^j | \mathbf{T}_s^j M_s^{test}) \quad (4)$$

where $\mathbf{T}_s^j M_{s,t}^{test}$ denotes transformation of the test model in to M_s^j's feature space and $\alpha_{s,i}^j$ is a constant proportional to $p_{s,i}^j$ and $|\mathbf{\Sigma}_{s,i}^j|$, the determinant of the diagonal matrix in the optimal model feature space. In the sum, i ranges over all the components of M_s^j. Figure 1 shows the behavior of this parameter, where "normalized" indicates the normalization in section 4.1.3. The major trend over the models can be seen to be relatively consistent for the two utterances, which indicates the relative richness of the training data for the 25 models used. However, there are a fair number of models where the values diverge, indicating the variable relative richness of the test data (in the two utterances). The power of this measure lies in the fact that ϕ is not symmetric with respect to the roles of the training and

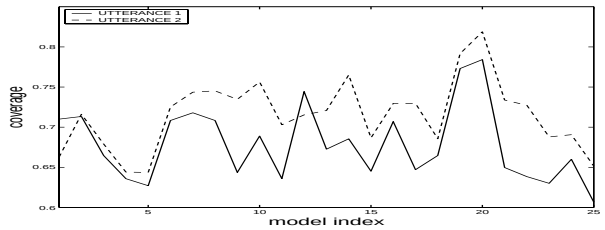

Figure 1: Normalized coverage (Φ_t^a) for 25 models over 2 different utterances from one speaker.

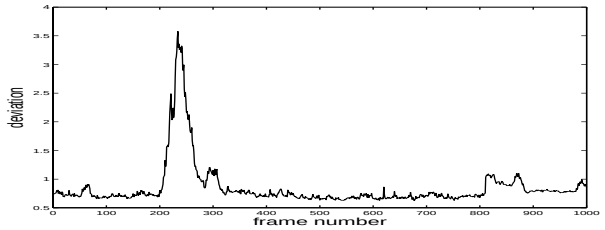

Figure 2: Video stream deviation ($\psi_t^v(j)$) for a 1000 frame section.

test data. A high concentration of test data near one model component can yield a high likelihood, yet when the roles are reversed, and the model is built on the test data, the training data means and/or vectors could have a very low likelihood, indicating that the training data is much richer in comparison to the test data. One can associate ϕ with the fraction of the model covered by the test data.

4.1.2. Deviation: $\psi_t^s(j)$

The parameter $\psi_t^s(j)$ is computed using the deviation of the score at time t from a point estimate of the score at time t, based on a neighborhood (in time) of test vectors $\mathbf{X}_{nbhd}^s$ (the size of this neighborhood is in general independent of that used for determining ϕ). It is a measure of the relative instability of the data stream at time t.

$$\psi_t^s(j) = \beta_{s,t}^j \frac{d_s(\mathbf{x}_t^s|M_s^j) - \mu[d_s(\mathbf{x}^s|M_s^j); \mathbf{x}^s \in \mathbf{X}_{nbhd}^s]}{\sigma[d_s(\mathbf{x}^s|M_s^j); \mathbf{x}^s \in \mathbf{X}_{nbhd}^s]} \tag{5}$$

Notice that $\beta_{s,t}^j$ should ensure that $\psi_t^s(j)$ is positive. The deviation shown in figure 2 typically hovers closely to a constant value, there are a number of sections where the deviation becomes quite large. The ψ parameter is related to the ϕ factor in that an unstable score stream can be the result of the differing richness of the train and test data. However, as one does not necessarily imply the other, it is advantageous to use both parameters.

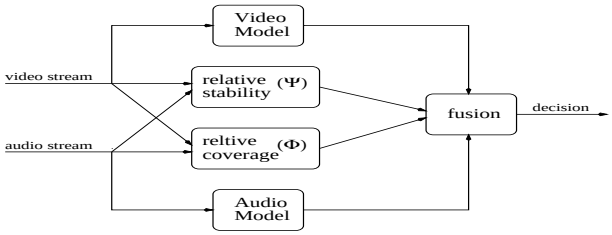

Figure 3: Decision parameters and data flow.

4.1.3. Normalization: $\Phi_t^s(j)$, $\Psi_t^s(j)$, and η_s

Φ and Ψ are normalized parameters based on ϕ and ψ:

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

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

This induces a context dependence since the weights on one stream depend on the other. The reciprocal is used in computing Ψ because we want the factor to be inversely proportional to the deviation. The η_s parameter, incorporating the normalization for the scale differences in the score streams, is set (based on empirical performance) to $1/\mu_{s,global}$, which is the reciprocal of the mean value of the stream elements to be combined, taken over a large sample of data.

4.2. Identification Decision

Speaker identification is carried out by computing equation 2 or 3 for each speaker j, and letting the decision be given by

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

5. EXPERIMENTS

Our experiments are based on an audio-visual database consisting of 304 speakers. The speech and audio were captured as the users read prompted text while in front of a computer that was equipped with a microphone and camera. For each speaker, approximately 120 seconds of speech was used for training and on average the test utterances were 6.7 seconds long with a standard deviation of 2.7. Experiments were conducted at both the utterance level and frame level. For the frame level experiments, a 100 speaker subset of the data was chosen to reduce computation and storage costs. The total number of tests for the full (All) and reduced (100spkr) sets are 19714 and 7307 respectively.

Config	Data Set	
	All	100Spkr
Audio, $\mathbf{X}^a$	98.2%	98.0%
Video, $\mathbf{X}^v$	75.4%	89.1%
Audio+Video, $\mathbf{X}^{av}$	69.1%	90.7%

Table 1: Identification rates on A/V multi-stream data.

	Config		
	Ψ ($\Phi = 0$)	Φ ($\Psi = 0$)	$\Psi\&\Phi$
100spkr	99.3%	99.1%	99.6%
	audio only		$\Psi\&\Phi$
changed spkrs	95.5%		99.5%

Table 2: Performance with frame-level time and context dependent weights.

5.1. Baseline

Results are given in table 1 for the cases where the three streams $\mathbf{X}^a$, $\mathbf{X}^v$, and $\mathbf{X}^{av}$ are used in isolation (there is no score combination or weighting). Recall that $\mathbf{X}^{av}$ is the vector-wise concatenation of $\mathbf{X}^a$ and $\mathbf{X}^v$. The discriminant in these cases is $D(\mathbf{X}|j) = \sum_t d_s(\mathbf{x}_t^s|M_s^j)$, where s is either a, v, or av. As can be seen from the results in table 1, vector-wise concatenation can be detrimental. It is evident that the speakers for whom good video data existed and still preserved the base audio error rate were chosen for the reduced 100spkr set. Also, for the sake of comparison, we give results for the case where the streams are weighted with a constant factor for all time, i.e.

$$D(\mathbf{X}|j) = \sum_t [\omega_a d_a(\mathbf{x}_t^a|M_a^j) + (1-\omega_a) d_v(\mathbf{x}_t^v|M_v^j)].$$

The identification performance on the 100spkr set is computed on a grid of weights with ω_a ranging from 0.0 to 1.0. The boundary error rates are the same as in table table 1. We observe that there is a monotonic increase in accuracy until the fraction of audio goes beyond 0.9, where the performance peaks at 98.9, showing some benefit of adding video to the audio system.

5.2. Time Varying Discriminants

Here we focus on the reduced 100spkr population and the frame-level combination experiments. In table 2 the effects of using the time and context dependent weights (Φ and Ψ) in isolation and together, using the sum form (2), are shown. Using either parameter in isolation is beneficial, but using both together clearly outperforms all cases. If we consider the speakers for whom at least one decision (for one test utterance) changed, we get 27 speakers whose tests account for 3131 trials. The improvement for these speakers (over the audio only case) is given in table 2.

6. CONCLUSION

Here we have presented a new method to combine information present in two, or more, streams of data. We propose that the quality of data and the richness of the testing data relative to the training data vary over time and in fact within the boundaries of an utterance. A notion of data reliability was developed incorporating the stability, or point consistency, of a score stream and the coverage of the model by the test data. Experiments showed that this decision method out-performed the use of audio alone, video alone, or a concatenation of the streams. The results are promising because they are obtained for the clean speech case, for which it was previously questioned whether adding video data could improve performance.

7. 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] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, Goats, Lambs and Wolves: An Analysis of Individual Differences in Speaker Recognition Performance", *NIST presentation at ICSLP98, Sydney, Australia, November 1998.*

[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 (MMSP99)*, 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 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.

[7] 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.