INFORMATION ACCESS USING SPEECH, SPEAKER AND FACE RECOGNITION

M. Viswanathan, H. S. M. Beigi

IBM T.J. Watson Research Center
P.O. Box 218,
Yorktown Heights, NY 10598, USA
E-mail: maheshv@watson.ibm.com,
(914) 945-1754 (tel), (914) 945-4490 (fax)

F. Maali

Signal Recognition Corporation
P.O. Box 7010,
New York, NY 10128, USA

ABSTRACT

We describe a scheme to combine the results of audio and face identification for multimedia indexing and retrieval. Audio analysis consists of speech and speaker recognition derived from broadcast news video clip. The video component is analyzed to identify the persons in the same video clip using face recognition. When applied individually both speaker and face recognition schemes have limitations on conditions under which they perform reasonably well. By integrating the match-score results of both audio and video analysis, we find that the two techniques can complement each other. We discuss the system architecture for such a combined system, and discuss how decision fusion is applied to disparate match-scoring systems to yield the final speaker identity.

1. INTRODUCTION

Our multimedia content based indexing and retrieval system requires analysis of both textual and speaker content in a broadcast news audio or video clip. A system that performs such analysis must have the following distinct capabilities in addition to the indexing and retrieval functions: (1) automatic transcription of speech into text; (2) automatic speaker-based segmentation of the incident audio; and (3) identification of the speakers of the detected segments. We have already prototyped and reported such a system [1]. This system retrieves video clips or frames much like audio or text, using only the audio component of the video to build a combined text- and speaker-based index.

This paper is concerned with introduction of video modality to such a system for the specific purpose of reinforcing confidence in the speaker identity which can be derived from a purely audio analysis. Face detection and identification processes used here have been reported in [2, 3]. Here we focus on how to incorporate the face recognition into our existing audio-only architecture. We also describe the decision fusion process that integrates the results of the audio-based and the video-based classifiers.

2. SYSTEM OVERVIEW

The media source is an MPEG1 encoded digital stream. The audio and the video elementary streams must first be extracted and fed to the audio signal processing front-end and the face recognition module respectively. The overall system, at this time, is expected to run on two PC's with one PC being dedicated to the face recognition task. This decision is forced by the compute-intensive nature of current face recognition algorithms. This forces the need for an anchoring of the time basis that relates the audio and video frames that belong together.

The functional block diagram of the overall system is depicted in Figure 1. The audio front-end converts the incident audio into mel-cepstral feature vectors (or frames) and provides them simultaneously to the transcription, segmentation and identification engines. The output of the transcription engine is time aligned text while the output of the segmentation en-

SUBMITTED: INT'L CONFERENCE ON MULTIMEDIA AND EXPO, NEW YORK CITY, NY, JULY 30 - AUGUST 2, 2000

gine is turns (segment boundaries). These are used to display the spoken segments on the screen for the user to view. But more importantly the time aligned words and turns are dumped into files for subsequent indexing.

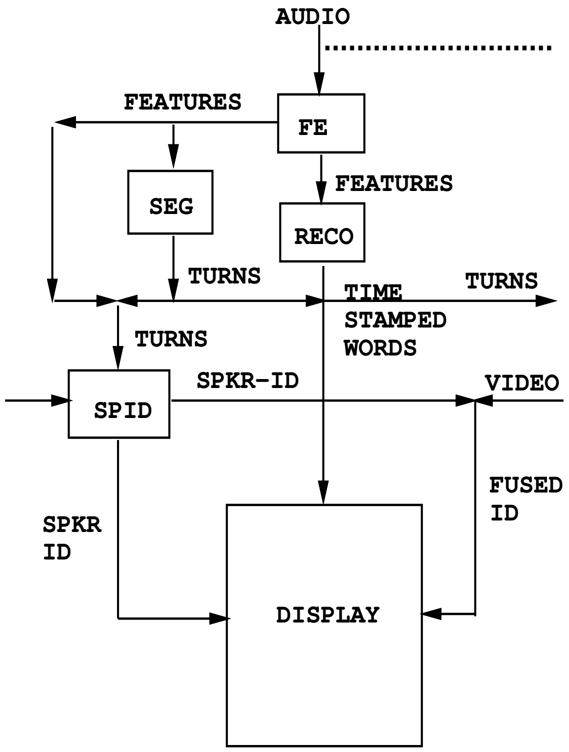

Figure 1: Architecture of our system combining speech, speaker, and face recognition - audio portion only

Speaker identification follows speaker segmentation using the same mel-cepstral vectors and turns information along with a database of voice-prints, to identify the speakers of the successive segments. Without face recognition, the identity of the speaker segment would be displayed on the screen and stored in the same buffer as the turns. But with face recognition, a fusion of the audio and video identities determines the final identity of the speaker. Now the fused identity replaces the audio-only speaker identity for speaker indexing.

The video processing progresses concurrently but asynchronously with audio processing. The turns generated by the audio speaker segmentation are used to delineate the video frames whose face content need to be identified for subsequent decision fusion with the audio speaker identification. In reality, faces are detected, tracked and identified at every n frames without regard to the location of the turns. The turns are used to correlate a single audio speaker segment with the set of video frames associated with it (see Figure 2).

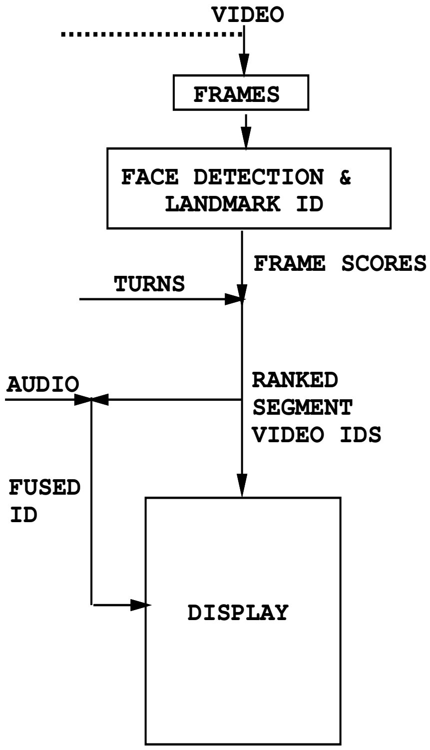

Figure 2: Architecture of our system combining speech, speaker, and face recognition - video portion only

3. AUDIO SPEAKER SEGMENTATION

The Bayesian Information Criterion (BIC) partitions the incident audio stream represented by mel-cepstral vectors into segments spoken by different speakers. A window W of N frames is examined to determine whether it includes a turn at its ith frame or not. Such a turn engenders two windows, W_1 and W_2. Then:

$$\Delta BIC(i) = -R(i) + \lambda P \text{ where,}$$

$$R(i) = -\frac{N}{2} log|\Sigma| - \frac{i}{2} log|\Sigma_1| - \frac{N-i}{2} log|\Sigma_2|$$

and $P = \frac{1}{2}(d + \frac{d(d+1)}{2}) logN$, i is the number of frames in W_1, $(N-i)$ is the number of frames W_2, and d is the dimension of the cepstral vectors. λ is a penalty function which for best results for 24-dimensional vectors has the value 1.3.

A turn is confirmed at frame i when i is not only the minimizer of ΔBIC but also drives it negative. If no turns are found, the window size is increased by an increment of N and the test is applied again. This process is repeated until a turn is found, but when no turn is found and the window size exceeds a predefined size without a segment boundary, the earliest frames are dropped from consideration but their statistics are merged with that of the new window. See [4] for details.

4. AUDIO-BASED SPEAKER IDENTIFICATION

Speaker enrollment is a prelude to audio-based speaker identification. Utterances, represented by mel-cepstral vectors, from each speaker of interest are modeled as a mixture of Gaussian distributions through a clustering process [5]. Let $\{M_i \mid i = 1..I\}$ denote the models pertaining to each of the enrollees. Each model M_i can have N_i distributions associated with it. Let ω_{ij} refer to the jth distribution of model I. Also let $\{\vec{x}_t \mid t = 1..T\}$ denote the frames representing the test utterance **z** whose speaker is sought. The run-time class assignment progresses as follows:

$$\text{assign } \mathbf{z} \longrightarrow \mathrm{M_q} \text{ iff } \mathrm{D_q} = \min_{\mathrm{i=1..I}} [\mathrm{D_i}], \text{ where}$$

$$D_i = \sum_{t=1}^{T} d(i,t), \; i = 1..I, \text{ and}$$

$$d(i,t) = -log[\sum_{j=1}^{N_i} P(\omega_{ij}) \, p(\vec{x}_t|\omega_{ij})],$$

with $P(\omega_{ij})$ being the prior of the jth distribution of model i, and $p(\vec{x}_t|\omega_{ij})$ being the conditional pdf of the tth frame of the test utterance conditioned on the jth component of model i.

Each class assignment is accompanied by a pseudo distance which expresses the respective confidence. For model i this is $D_i \times T$, where T denotes the number of frames in the respective test utterance.

5. FACE RECOGNITION

The video elementary stream is subjected to a frame-by-frame analysis whose object is to first segment the image by partitioning it into face and non-face regions and also isolating the faces from each other. It is the so-called face detection process [2] that brings about image segmentation in a video frame. With every detected face a track is initiated and an attempt is made to maintain the track temporally in the subsequent frames for that face. Face detection is attempted in any frame only when tracks cannot be maintained.

Within each face also some 29 landmarks are located. Face identification is based on landmark recognition [3]. Hence a face is assigned to any one of the number of prototype face classes when its landmarks exhibit the highest comparative similarity to the constituent landmarks of that given prototype face.

After face detection the scale and orientation of each landmark can be inferred. After being standardized to its nominal scale and orientation each landmark is represented by a Gabor jet – a feature vector representation of the local grey value distribution. Such representation for prototype landmarks are derived in the course of a prior enrollment. The detected landmarks are correlated to their prototype counterparts through the following similarity measure, and the mean of thusly computed similarity measures for all the detected landmarks yields the overall score or confidence in the respective face identity.

$$S(\mathbf{y}, \mathbf{Y}) = \frac{\sum_j y_j \; \sum_j Y_j}{\sqrt{\sum_j {y_j}^2 \; \sum_j {Y_j}^2}}$$

where $\mathbf{y}$ and $\mathbf{Y}$ are the Gabor jets for a given test and prototype landmark, and S denotes the inter landmark similarity measure. The mean of S for all detected landmarks constitutes the video score qualifying a class assignment to a given face.

6. DECISION FUSION

The audio-based speaker identification yields a single set of ranked identities for each audio segment. On the other hand, a single audio segment corresponds to multiple video frames with each video frame yielding a set of ranked face identities. Scores expressing the confidence in the respective class assignments are available with each ranked identity.

Let the number of video frames which survive face detection in a given speaker segment be n. As a first step these n sets of ranked identities have to be abstracted into a single set of ranked identities. This comprises of: 1) finding the most frequent face identity (the statistical mode) at each rank across all the video frames corresponding to the audio speaker segment; and 2) computing the median score for that rank and assigning it to thusly derived most frequent face identity. We now have two sets of ranked identities, one audio-based and the other video-based.

In the next step, the audio speaker scores are scaled to a 0-1 range and normalized by the standard deviation of the scores of the ranked identities for each segment. This is repeated for the video segment scores. This operation makes the video and the audio scores compatible for the subsequent integration.

Our decision fusion scheme is based on linear combination of the audio and the video class assignments. Hence we are immediately faced with the issue of weight selection. The weights assigned to the audio and the video scores affect the influence of their respective scores in the ultimate outcome. One approach is to use fixed weights, say 0.5, for each. Another approach is to let the weights be derived from the data itself. The rest of the formulation used here uses this latter approach.

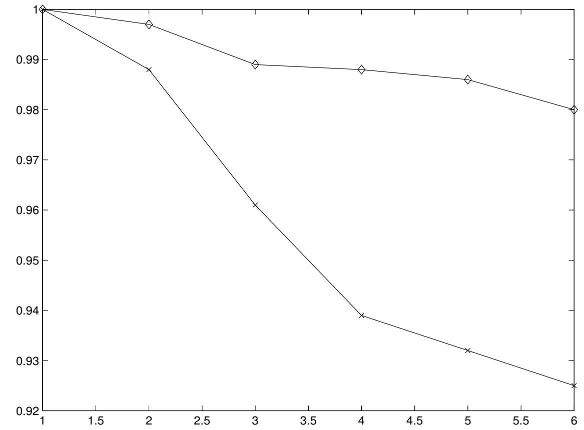

Figure 3: Ranked id scores for audio (above) and video (below) for o ne audio segment with six rankings.

Let $\{(rank_r, audioScore_r) \mid r = 1..maxRank\}$ denote the scores for ranked identities for audio speaker class assignments in a rank-score coordinate system, where $rank_r$ represents the rank and $audioScore_r$, the audio score of the rth point. In the same manner, let $\{(rank_r, videoScore_r) \mid r = 1..maxRank\}$ denote the corresponding data set for the video identities. Both audio-based and video-based vary monotonically along the rank axis (Figure 3). We impose a linear variation on the rank-score data by: (1) removing outliers using the Hough transform; and (2) fitting the surviving points set to a line using the least mean squares error method. Thereby the ranked identities output by the audio and video identification systems are reduced to two straight lines defined by $audioScore = m_1 \times rank + b_1$ and $videoScore = m_2 \times rank + b_2$. The line with higher slope is assumed to convey more discriminative information. The normalized slopes of the two lines are used as the weight of the respective results when combining the scores from the audio-based and video-based speaker analysis.

With w_1 and w_2 representing audio and the video channels respectively we compute the fused score, FS_k, for each speaker as follows:

$$w_1 = \frac{m_1}{m_1 + m_2} \text{ and } w_2 = \frac{m_2}{m_1 + m_2}.$$

$$\begin{aligned} FS_k &= w_1[m_1 \times rank_k + b_1] \\ &\quad + w_2[m_2 \times rank_k) + b_2] \end{aligned}$$

where $rank_k$ is the rank for speaker k.

The above expression is computed over the collection of audio and video speaker identities derived from the last step, and later sorted by score to obtain a new set of fused ranked identities. These fused identities may be displayed to the user as the identified speaker result, and buffered alongside the turns for subsequent use in the speaker indexing.

7. RESULTS

The above fusion scheme was applied to a broadcast news video clip with eight speakers. Figure 4 illustrates the result and the impact of multi channel decision fusion compared with audio-based speaker identification.

RANK	AUDIO		VIDEO		FUSED	
1	UM	1.0	JW	1.0	OH	.99
2	OH	.997	OH	.988	AK	.966
3	UF	.989	AK	.961	GB	.941
4	AK	.988	SF	.939	JM	.936
5	JM	.986	GB	.932	JW	.808
6	GB	.98	JM	.925	SF	.759

Figure 4: Fused results for six top ranks. Note that the true speaker, OH, was not identified on either channels in the top position, but the integrated decision yielded the correct result (The speaker was indeed OH).

8. REFERENCES

[1] M. Viswanathan, H.S.M. Beigi, S. Dharanipragada, and A. Tritschler, "Retrieval from Spoken Documents Using Content And Speaker Information," *Proc., Intl. Conf. on Doc. Anal. and Retr. (ICDAR99)*, Bangalore, India, 1999, pp. 567–572.

[2] A. Senior, "Face and Feature Finding for Face Recognition System," *2nd Intl. Conf. on Audio- and Video-based Biom. Person Authent.*, Washington D.C., March 1999.

[3] A. Senior, "Recognizing Faces in Broadcast Video," *Proc. IEEE International. Workshop on Recog., Anal., and Tracking of Faces and Gestures in Real-time Systems*, Kerkyra, Greece, 1999, pp. 105-110.

[4] S. S. Chen and P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," *Proc., DARPA Workshop*, 1998, pp. 127–132.

[5] H.S.M. Beigi, S. Maes, U.V. Chaudhari and J.S. Sorenson, "IBM Model-Based and Frame-By-Frame Speaker-Recognition," *Proc., Speaker Recog. and its Commercial and Forensic Appl.*, Avignon, France, 1998.