ICASSP 2003, Hong Kong

AUDIOVISUAL-BASED ADAPTIVE SPEAKER IDENTIFICATION

Ying Li, Shrikanth Narayanan and C.-C. Jay Kuo

Integrated Media Systems Center and Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564

E-mail:{yingli,shri,cckuo}@sipi.usc.edu

ABSTRACT

An adaptive speaker identification system is presented in this paper, which aims to recognize speakers in feature films by exploiting both audio and visual cues. Specifically, the audio source is first analyzed to identify speakers using a likelihood-based approach. Meanwhile, the visual source is parsed to recognize talking faces using face detection/recognition and mouth tracking techniques. These two information sources are then integrated under a probabilistic framework for improved system performance. Moreover, to account for speakers' voice variations along time, we update their acoustic models on the fly by adapting to their newly contributed speech data. An average of 80% identification accuracy has been achieved on two test movies. This shows a promising future of the proposed audiovisual-based adaptive speaker identification approach.

1. INTRODUCTION

A fundamental task in video analysis is to organize and index multimedia data in a meaningful manner so as to facilitate user's access such as browsing and retrieval. This work proposes to extract an important type of information, the *speaker identity*, from feature films for the content indexing and browsing purpose.

So far, a large amount of speaker identification work has been reported on standard speech databases. A speaker detection approach based on likelihood ratio calculation was adopted in [1] to estimate target speaker segments using HUB4 broadcast news database. Johnson [2] addressed the problem of labeling speaker turns by automatically segmenting and clustering a continuous audio stream obtained from the 1996 Hub4 development data. Yu and Gish [3] reported their work on identifying speakers engaged in telephone dialogs obtained from the SWITCHBOARD corpus.

Recently, with the increase of the accessibility to other available media sources, researchers have attempted to improve the system performance by integrating the knowledge from all media cues. For instance, Tsekeridou and Pitas [4] proposed to identify speakers by integrating cues from both speaker recognition and facial analysis modules. This system is however impracticable for generic video types since it restricts the number of faces to be 1 in each shot. Similar work was also reported in [5] where TV sitcom was used as test sequence. In [6], Li *et al.* presented a speaker identification system for feature films where both audiovisual cues were employed. However, this system has certain limitations since it only identifies speakers in movie dialogs.

From the other point of view, most existing work in this field deals with supervised identification problem, where speaker models are not allowed to change once they are pre-trained. Two drawbacks arise when this approach is applied to feature films. First, we may not have sufficient training data. Because a speaker's voice can have distinct variations along time, especially in feature films. Thus, a model built with limited training data cannot model a speaker well for the entire sequence. Second, since we have to go through the movie at least once to collect and transcribe the training data before the actual identification process can be started, it wastes time and decreases system efficiency.

An adaptive speaker identification system is proposed in this work with the goal to offer a better solution for identifying speakers in movies. Specifically, after building coarse models for target speakers during system initialization, we will continuously update them on the fly by adapting to speakers' newly contributed data. It is our claim that, by adapting models to new speech data, we can achieve higher identification accuracy as they can better capture speakers' voice variations along time. Both audio and visual sources will be exploited in the identification process, where the audio source is analyzed to recognize speakers using a likelihood-based approach, and the visual source is parsed to find talking faces using face detection/recognition and mouth tracking techniques.

2. ADAPTIVE SPEAKER IDENTIFICATION

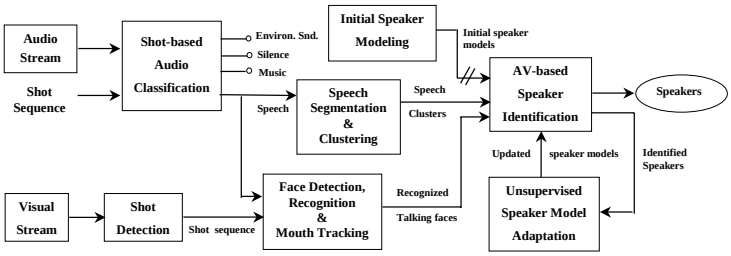


Fig. 1. The block diagram of the proposed adaptive speaker identification system.

Figure 1 shows the proposed system framework that consists of the following six major modules: (1) shot detection and audio classification, (2) face detection, recognition and mouth tracking, (3) speech segmentation and clustering, (4) initial speaker modeling, (5) audiovisual (AV)-based speaker identification, and (6) unsupervised speaker model adaptation. As shown, given a video input, shot detection is first carried out, followed by an audio classification process. Next, with non-speech shots being discarded, the speech shots are further processed by the speech segmentation/clustering module to generate homogeneous speech clusters.

Meanwhile, a face detection/recognition and mouth tracking process is performed to recognize talking faces in speech shots. Next, based on either initial or updated speaker models, the AV-based identification module identifies the target speakers by integrating both speech and face cues. Finally, we use the detected speaker identities to guide an unsupervised speaker model adaptation process. The updated speaker models will become effective in the next round.

Due to the space limit, we will mainly focus on the last three modules. For the details of other modules, please refer to our previous work [7].

2.1. Shot Detection and Audio Classification

The first step towards visual content analysis is shot detection. In this work, a color histogram-based approach is employed to carry out this task.

In the second step, we analyze the audio content of each shot and classify it into one of the following four classes: *silence*, *speech* (including speech with music), *music*, and *environmental sound*, based on five different audio features including short-time energy function, short-time average zero-crossing rate, short-time fundamental frequency, energy band ratio and silence ratio [8].

2.2. Face Detection, Recognition and Mouth Tracking

2.2.1. Face Detection and Recognition

The face detection and recognition library used in this work is mainly designed to detect upright frontal faces or faces rotated by plus or minus 10 degrees from the vertical. To speed up the process, we only carry out face detection on speech shots as shown in Figure 1. Also, to facilitate the subsequent recognition process, we organize detection results into a set of face sequences, where all frames within each sequence contain the same number (nonzero) of human faces.

The face database used for face recognition is constructed as follows. During the system initialization, we first ask users to select their N interested casts (also called *target speakers*) by randomly choosing video frames containing the casts' faces. These faces are then detected, associated with the names of the corresponding movie characters, and added into the face database.

During the face recognition process, each detected face in the first frame of each face sequence is recognized. The result is returned as a face vector $\vec{f} = [f_1, \ldots, f_N]$, where f_i is a value in [0, 1] which indicates the confidence of being target cast i.

2.2.2. Mouth Detection and Tracking

In this step, we aim to detect and track the mouth for each detected face sequence. Note that if more than two faces are present in the sequence, we will virtually split it into several sub-sequences with each focusing on one face. However, if a talking face is detected for one of the sub-sequences, others will no longer be processed.

Now, given the eye position information output from the face detector, we first exploit the facial mirror-symmetry and biometric analogy principles to locate the coarse mouth center, and subsequently expand it into a rectangular mouth search area. Then, a weighted block-matching process is carried out to locate the target mouth area based on the criterion that it should present the largest color difference from the skin color.

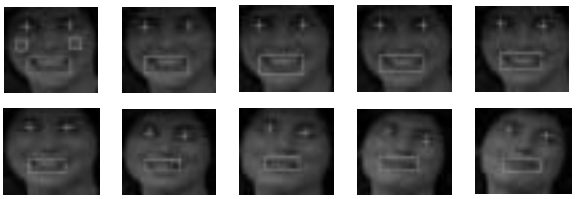


Fig. 2. Mouth detection and tracking results from 10 consecutive video frames.

To track the detected mouth for the subsequent frame, we derive its new centriod from that of the previous frame as well as from the newly obtained eye positions. Figure 2 shows some face detection and tracking results on a face sequence containing ten consecutive frames. Detected and tracked mouths are marked by rectangles while eyes are indicated by crosses. As we can see, encouraging results have been achieved.

Finally, we apply a color histogram-based approach to determine if the tracked mouth is talking. Particularly, if the normalized accumulated histogram difference in the mouth area of the entire or part of the face sequence exceeds a certain threshold, we label it as a talking mouth; and correspondingly, we mark this sequence as a talking face sequence.

2.3. Speech Segmentation and Clustering

For each speech shot, the two major speech processing tasks are speech segmentation and clustering. In the segmentation step, all individual speech segments are separated from the background noise/silence. In the clustering step, we group speech segments of the same speakers into homogeneous clusters so as to facilitate successive processing [9].

The adaptive silence detector proposed in our previous work [6] is employed to perform the segmentation task. Specifically, given an incoming speech shot, this approach first calculates a proper threshold to distinguish speech signals from the background. Then, speech segments are extracted based on a 4-state transition diagram.

For the speech clustering task, we first apply the Bayesian Information Criterion (BIC) [9] to measure the similarity between two speech segments, and then group them together if the BIC distance is smaller than a certain threshold. The BIC distance between a speech segment X and a cluster C is computed as the weighted distance between X and all of C's component segments in this work [7].

2.4. Initial Speaker Modeling

To bootstrap the identification process, we need initial speaker models as shown in Figure 1. This is achieved by exploiting the inter-relations between the face and speech cues. Specifically, for each target cast A, we first find a speech shot that A is talking based on the face detection result. Then, we collect all of its speech segments and build A's initial model. The Gaussian Mixture Model (GMM) has been employed here for the modeling purpose. Note that at this stage, the initial model will only contain one Gaussian mixture with its mean and covariance computed as the global mean and covariance.

2.5. Likelihood-based Speaker Identification

At this stage, we aim to identify speakers based on pure speech information. Specifically, given a speech signal, we first decompose it into a set of overlapped audio frames; then 14 Mel-frequency cepstral coefficients (MFCC) are extracted from each frame to form an observation sequence X. Next, we calculate the likelihood $L(X; M_i)$ between X and all speaker models M_i based on the multivariate analysis. Finally, we get a list of speaker candidates sorted in the descending order of their likelihood values, with the top one being the most probable target speaker.

Now, based on this scheme, given any speech cluster C, we will assign it a speaker vector $\vec{v} = [v_1, \ldots, v_N]$, where v_i is a value in [0, 1] which indicates the confidence of being target speaker i.

2.6. Audiovisual Integration for Speaker Identification

This step aims at finalizing the speaker identification task for cluster C (in shot S) by integrating the audio and visual cues obtained in Sections 2.2, 2.3 and 2.5. Specifically, given cluster C and all recognized talking face sequences F in S, we examine if there is a temporal overlap between C and any sequence F_i. If yes, we assign F_i's face vector $\vec{f}$ to C if the overlap ratio exceeds a threshold. Otherwise, we set C's face vector to null. But if C is overlapped with multiple F_i due to speech clustering or talking face detection errors, we choose the one with the highest overlap ratio.

Now, we determine the speaker's identity in cluster C as

$$speaker(C) = \arg \max_{1 \le j \le N} (w_1 \cdot f[j] + w_2 \cdot v[j]),$$

where $\vec{f}$ and $\vec{v}$ are C's face and speaker vectors, respectively. w_1 and w_2 are two weights that sum up to 1.0. Currently we set them to be equal in the experiment.

2.7. Unsupervised Speaker Model Adaptation

Now, after we identify speaker P for cluster C, we will update his model using C's data in this step. Meanwhile, a background model will be either initialized or updated to account for all non-target speakers. Specifically, when there is no a priori background model, we use C's data to initialize it if the minimum of $L(C; M_i), i = 1, \ldots, N$ is less than a preset threshold. Otherwise, if the background model produces the largest likelihood, we denote the identified speaker as "unknown" and use C's data to update the background model.

The following three approaches are investigated to update the speaker model: *Average-based* model adaptation, *MAP-based* model adaptation, and *Viterbi-based* model adaptation.

2.7.1. Average-based Model Adaptation

In this approach, P's model is updated in the following three steps.

Step 1: Compute BIC distances between cluster C and all of P's mixture component b_i. Denote the component that gives the minimum distance d_{min} by b_0.

Step 2: If d_{min} is less than an empirically determined threshold, we consider C to be acoustically close to b_0, and will use C's data to update it. Specifically, let $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ be C and b_0's Gaussian models, respectively, we update b_0's mean and covariance as

$$\mu_2' = \frac{N_1}{N_1 + N_2}\mu_1 + \frac{N_2}{N_1 + N_2}\mu_2, \quad (1)$$

$$\Sigma_2' = \frac{N_1}{N_1 + N_2}\Sigma_1 + \frac{N_2}{N_1 + N_2}\Sigma_2 + \frac{N_1 N_2}{(N_1 + N_2)^2}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T, \quad (2)$$

where N_1 and N_2 are the numbers of feature vectors in C and b_0, respectively.

Otherwise, if d_{min} is larger than the threshold, we will initialize a new mixture component for P with its mean and covariance equaling to μ_1 and Σ_1. However, once the total number of P's components reaches a certain value (which is set to 32 in this work), only component adaptation is allowed. This is adopted to avoid having too many Gaussian components in each model.

Step 3: Update the weight for each of P's mixture component.

2.7.2. MAP-based Model Adaptation

MAP adaptation has been widely and successfully used in speech recognition, yet it has not been well explored in speaker identification. In this work, due to the limited speech data, only Gaussian means will be updated. Specifically, given P's model M_p, we update the component b_i's mean $\mu_i (i = 1, \ldots, m)$ via

$$\mu_i' = \frac{L_i}{L_i + \tau}\bar{\mu} + \frac{\tau}{L_i + \tau}\mu_i, \quad (3)$$

where τ defines the "adaptation speed" and is currently set to 10.0. L_i gives the occupation likelihood of the adaptation data $\vec{x}_t (t = 1, \ldots, T)$ to component b_i, and is defined as $L_i = \sum_{t=1}^{T} p(i|\vec{x}_t, M_p)$, where $p(i|\vec{x}_t, M_p)$ is the *a posteriori* probability of $\vec{x}_t$ to b_i. Finally, $\bar{\mu}$ gives the mean of the observed adaptation data [7].

Unlike the previous method, this MAP adaptation is applied to every component of P based on the principle that every feature vector has a certain possibility of occupying every component. Thus, MAP adaptation provides a soft decision on which feature vector belongs to which component.

2.7.3. Viterbi-based Model Adaptation

Similar to the MAP-based approach, this approach also allows different feature vectors belonging to different components. Nevertheless, while the MAP approach provides a soft decision, this approach implies a hard decision, *i.e.* for any one particular feature vector $\vec{x}_t$, it can either occupy component b_i or not. Therefore, the probability function $p(i|\vec{x}_t, M_p)$ is now replaced by an indicator function which is either 0 or 1. Now, given any feature vector $\vec{x}_t$, the mixture component it occupies will be determined by

$$m_0 = \arg \max_{1 \le i \le m} p(i|\vec{x}_t, M_p). \quad (4)$$

Finally, Equations (1) and (2) are used to update P's components after we assign every feature vector to its component. As one can see, this approach is actually a compromise between the previous two methods.

3. EXPERIMENTAL RESULTS

To evaluate the system performance, we tested our algorithms on two 1.5-hour long movies. Experimental results on the first movie is reported here.

Three target speakers (casts), denoted by A, B, and C, were chosen for Movie1. Totally, 952 speech clusters were generated. The identification results for all these clusters are reported in Table 1 with respect to all three adaptation approaches. Three parameters, namely, *identification accuracy* (IA), *false rejection* (FR), and *false acceptance* (FA) are calculated to evaluate the system performance. However, since $IA = 1 - FR$, we only report results for IA and FA.

Overall, acceptable results have been achieved considering the unsupervised nature of the proposed system. The MAP-based adaptation approach performs slightly better than the average-based approach, yet at the cost of a higher computational complexity. The Viterbi-based approach gives the best result, which may imply that, for speaker identification, a hard decision would be good enough.

By carefully studying the results, we found two major factors that degrade the system performance: (a) imperfect speech segmentation and clustering, and (b) inaccurate facial analysis results. Due to the various sounds/noises existing in movies, it is extremely difficult to achieve perfect speech segmentation and clustering. Besides, incorrect facial data can result in mouth detection and tracking errors, which will further affect the identification accuracy.

Table 1. Adaptive speaker identification results.

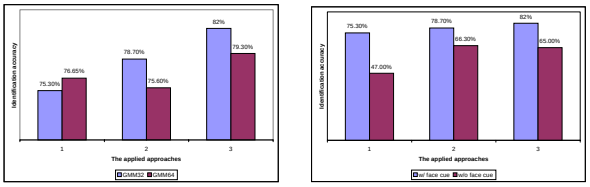
Method	IA			FA		
	A	B	C	A	B	C
AVG-based	74%	75%	77%	20%	19%	28%
MAP-based	78%	80%	78%	27%	14%	22%
VTB-based	80%	84%	82%	22%	14%	24%

Further, to examine the robustness of the three set of speaker models (denoted by AVG, MAP and VTB) obtained from the three adaptation approaches, we also carried out a supervised speaker identification using these models. Identification results are shown in Table 2, and a slightly degraded system performance is observed. This indicates that fixed models are not suitable for a long movie sequence no matter how many training data are used. Nevertheless, this table still presents acceptable results, especially for the Viterbi-based approach, where we have an averaged value of 80% IA and 23% FA. These results are comparable to those reported by other supervised identification work [1], [3].

Table 2. Supervised speaker identification results.

Model Set	IA			FA		
	A	B	C	A	B	C
AVG	68%	74%	70%	23%	21%	33%
MAP	71%	81%	72%	25%	21%	26%
VTB	74%	90%	76%	24%	17%	28%

To determine the optimal upper limit for the number of model components, we have examined the average identification accuracy in terms of 32 GMMs and 64 GMMs for all three adaptation methods and plotted them in Figure 3(a). As shown, except for the average-based method where a similar performance is observed, the use of 32 GMMs produces a better performance.



(a) (b)

Fig. 3. Comparison of identification accuracy for average-based, MAP-based and Viterbi-based approaches in terms of (a) GMM32 vs. GMM64, and (b) w/ face cues vs. w/o face cues.

Finally, the average identification accuracies in terms of using or without using face cues are compared in Figure 3(b). Clearly, without the assistance of the face information, the system performance has been significantly degraded, especially for the average-based adaptation method. This indicates that the face cue plays an important role in guiding the model adaptation.

4. CONCLUSION AND FUTURE WORK

An adaptive speaker identification system was proposed in this work, which employs both audio and visual cues to identify target speakers for feature films. For future work, we attempt to exploit the inter-relation between audio and visual cues in a more effective manner as well as work with more target speakers.

5. REFERENCES

[1] I. M. Chagnolleau, A. E. Rosenberg, and S. Parthasarathy, "Detection of target speakers in audio databases," *ICASSP'99*, Phoenix,1999.

[2] S. E. Johnson, "Who spoke when ? - automatic segmentation and clustering for determining speaker turns," *Eurospeech'99*, 1999.

[3] G. Yu and H. Gish, "Identification of speakers engaged in dialog," *ICASSP'93*, pp. 383–386, 1993.

[4] S. Tsekeridou and I. Pitas, "Content-based video parsing and indexing based on audio-visual interaction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 522–535, 2001.

[5] D. Li, G. Wei, I. K. Sethi, and N. Dimitrova, "Person identification in TV programs," *Journal of Electronic Imaging*, vol. 10, no. 4, pp. 930–938, 2001.

[6] Ying Li, S. Narayanan, and C.-C. Jay Kuo, "Identification of speakers in movie dialogs using audiovisual cues," *ICASSP'02*, Orlando, May 2002.

[7] Ying Li and C.-C. Jay Kuo, "Unsupervised real-time speaker identification for daily movies," *Proc. of SPIE*, vol. 4862, pp. 151–162, Boston, August 2002.

[8] T. Zhang and C.-C. Jay Kuo, "Audio-guided audiovisual data segmentation, indexing and retrieval," *Proc. of SPIE*, vol. 3656, pp. 316–327, 1999.

[9] S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.