IDENTIFICATION OF SPEAKERS IN MOVIE DIALOGS USING AUDIOVISUAL CUES

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

The problem of identifying speakers from a movie dialog scene is addressed in this paper. While most previous work on speaker identification has been carried out using pure audio data, more robust results could be obtained by integrating the knowledge from multiple media sources such as visual and audio information when they are available. In this work, we first identify and isolate speech segments from background by applying video shot detection, audio classification and adaptive silence detection techniques, then a decision is made based on the calculated likelihood between the incoming speech data and pre-trained speaker/background models. Moreover, to verify the effectiveness of the adaptive silence detector, we have compared it with a statistically trained silence model. Experimental results show that the proposed algorithm can achieve approximately 84% identification accuracy by integrating multiple media cues.

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. Most of current approaches to this task rely on organizing video into shots, scenes or events, yet very little attention has been paid to video semantics extraction. This is partly due to the fact that understanding video semantics is a very difficult task since low-level features are no longer sufficient, and knowledge from multiple media modals is needed to form a concrete conclusion. In this work, we approach this issue by identifying speakers engaged in movie dialogs using speaker identification technology.

There has been a considerable amount of work on speaker identification based on pure speech data. In [1], a speaker detection algorithm based on a likelihood ratio calculation was applied to estimate the target speaker segments from a HUB4 broadcast news database. Acceptable results were reported in the one-target-speaker case, yet the system performance degraded dramatically in the two-target-speaker case. Johnson [2] addressed the problem of labeling speaker turns by automatically segmenting and clustering a continuous audio stream. A frame-based clustering approach was also proposed, and an accuracy of 70% was obtained on the 1996 Hub4 development data. Yu and Gish [3] reported their work on identifying speakers engaged in telephone dialogs obtained from the SWITCHBOARD corpus, where some speaker clustering techniques were explored. Based on the fact that only one speaker is bound to be talking between two silence frames, Tsekeridou and Pitas [4] proposed to isolate speech data with a silence detection scheme. Their test data were TV news with high-quality audio. In all above work, the frame-based speaker segmentation and clustering approach can easily introduce errors since only acoustic similarity is considered. Thus, given a multimedia data stream such as movies, knowledge from all available media sources should be effectively integrated to achieve more robust identification results.

Our current work focuses on identifying the speakers engaged in a movie dialog by using audiovisual cues. In particular, to solve the problem of "who speaks when", we propose an adaptive silence detection algorithm to help isolate individual speech segments from the background noise. GMMs (Gaussian Mixture Models) of all enrolled speakers are pre-trained and used to compute the frame-based likelihood given the incoming speech data. The speaker whose model gives the largest likelihood will then be identified as the target one.

The rest of paper is organized as follows. Section 2 gives the framework of the proposed speaker identification system. Section 3 briefly discusses the speaker modeling process. In Section 4, we elaborate on the proposed "adaptive silence detection algorithm". Preliminary experimental results are reported in Section 5. Finally, concluding remarks and future work are given in Section 6.

2. PROPOSED SPEAKER IDENTIFICATION SYSTEM

Figure 1 gives the framework of our proposed speaker identification system, which consists of the following 4 major modules.

1. Shot-based audio classification module. Given a video input, we first segment it into a series of cascaded shots with each containing a set of contiguously recorded image frames. Then, audio of each shot is classified into one of the following 4 categories: speech, music, silence and environmental sound [5].

2. Event detection module. Based on integrated audiovisual cues, this module extracts 3 major types of movie events: hybrid, 2-speaker dialog and multiple-speaker dialog scenes [6].

3. Silence detection module. This module performs an adaptive silence detection for every speech shot within a 2-speaker dialog. The output of this module is well-isolated

0-7803-7402-9/02/$17.00 ©2002 IEEE

silence and speech segments. The speech segments will then be fed into the speaker identification module.

4. Speaker identification module. This module carries out the actual speaker identification task and makes the final decision.

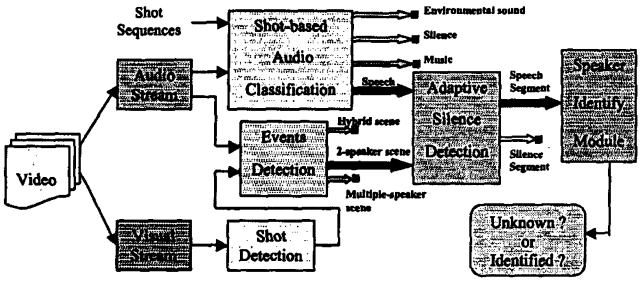

Fig. 1. The framework of the proposed speaker identification system.

Figure 2 shows a detailed block diagram of the speaker identification module. The input of this scheme is a speech signal that consists of a sequence of speech frames. The front-end analysis module extracts audio features from each frame and forms the feature vector $\vec{X} = \{x_1, x_2, ..., x_m\}$. The likelihood values $P_i(\vec{X}|M_i)$ between this vector $\vec{X}$ and all pre-trained speaker models M_i will then be calculated and normalized by using a background model. Finally, the likelihoods over all speech frames with respect to each model are summed, and the speaker whose model produces the maximum likelihood will be claimed as the identified speaker.

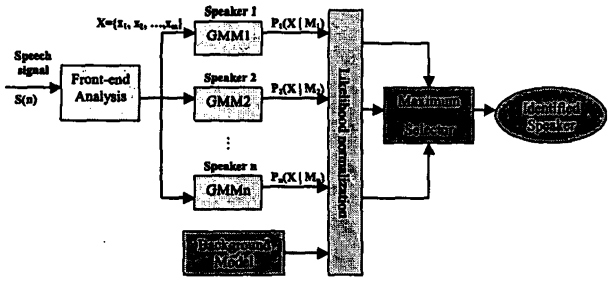

Fig. 2. The blockdiagram of the speaker ID module.

3. SPEAKER MODELING

Although there are no exclusively speaker distinguishing speech features, the speech spectrum has been shown to be very effective for speaker identification applications. In this work, we have chosen to use 14-dimensional MFCCs (Mel-Frequency Cepstral Coefficients) to represent the short-time speech spectra. A cepstral mean normalization is also performed so as to help reduce the effect caused by various kinds of background sounds that constantly exist in a movie. To model enrolled speakers, we have employed the Gaussian Mixture Model (GMM) due to its successful application in the speaker identification area.

3.1. Likelihood Calculation and Normalization

Let M_i be the GMM model corresponding to the i^{th} enrolled speaker and let X be the observation sequence consisting of T frames $\vec{x}_t$, $t = 1, \ldots, T$. Assuming that all observation frames are independent, the average log likelihood between X and M_i can be computed as

$$P(X|M_i) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|M_i)$$

where $p(\vec{x}_t|M_i)$ is the probability of $\vec{x}_t$ belonging to M_i. With these probabilities, an unknown speaker can be identified from a set of N speakers by choosing the one with the maximum likelihood.

Likelihood normalization, while proved to be very necessary for speaker verification, is usually not needed in a speaker identification system, since decisions made based on the likelihood from a single utterance require no inter-utterance likelihood comparisons [7]. In this work, however, we do need this normalization scheme since we have to deal with an "open-set" identification scenario, where not all speakers are known to our database.

To accommodate all uninterested speakers in the movie, we have built a background model M_b, which is trained using 40-second speech segments from all unenrolled speakers. The normalized version of $p(\vec{x}_t|M_i)$ can then be computed as

$$p_{norm}(\vec{x}_t|M_i) = \frac{p(\vec{x}_t|M_i)}{p(\vec{x}_t|M_b)}$$

4. ADAPTIVE SILENCE DETECTION SCHEME

Given the audio data of a speech shot, one natural assumption is that it is either from a known speaker, or by an unknown. However based on our observation, we find that this assumption is not always true. It is actually quite common that people take turns to speak within one single shot. In this case we need to first isolate the individual speech segments, then do the recognition job.

Recently, the "who speaks when" problem has attracted a lot of researcher's interest, and most of them relies on the speaker segmentation and clustering technique. However this approach tends to introduce errors since all processing is carried out at a low syntactic level [1], [3]. In this work, we propose an adaptive silence detection algorithm that performs at a higher semantic level, thus producing better results. In particular, this approach attempts to partition incoming audio signals into separate speech and silence segments so that only speech segments need to be processed by the identification module. The assumption used here is that a speech segment bounded by two silence segments should be attributed by one single speaker. This is a reasonable assumption, and has been well supported by all test data we have considered. Also, this algorithm is developed based on the observation that, although the background noise and

foreground speech levels are constantly changing over the time, they are typically quasi-stationary within one single shot.

Given the input audio signal containing N frames with respect to one speech shot, we first sort frames into an array based on their energies precomputed in the dB scale. Then, for all frames whose energy values are greater than a preset threshold $Discard\_thresh$, we quantize them into $NumBin$ bins, where bin_1 has the lowest average energy and bin_{NumBin} has the highest average energy. Since we already know that this shot contains both silence and speech signals, obviously bin_1 gives the lower bound of the silence energy, and bin_{NumBin} has the upper limit of the speech energy. Thus, the threshold separating the speech from silence should be a value between these two extremes. In this work, the threshold value is adjusted based on the sum of the first 3 bins and the sum of the last 3 bins as well as a predefined value $Speech\_thresh$, which gives the minimum dB difference between the two signals.

The last step of the algorithm is to postprocess the obtained segmentation results which include: a) sparkling points within the segments should be removed so as to obtain completely isolated speech and silence fragments; b) too short speech segments should be discarded to avoid meaningless results.

Figure 3(a) gives a silence detection example. The audio signals shown in the figure are obtained from one particular shot, where two persons take turns to talk. The pulse curve gives the detected speech-silence results, where the speech segments correspond to the peaks. Two detected speech fragments, as indicated in the figure, are considered to be too short for further processing. Here, we see that all speech fragments have been precisely isolated from the impulse background noise, resulting in 100% precision and recall rates. Moreover, these speech fragments are guaranteed to be from one single speaker, although several sentences continuously spoken by one person will probably be separated into several segments due to intermittent silence.

To verify the effectiveness of the proposed adaptive silence detection algorithm, we have performed experiments on detecting silence using a silence model statistically trained with various kinds of background sounds. Figure 3(b) shows the detection result on the same audio clip, where 2 noise clips are falsely detected as speech segments, resulting in 80% precision rate. Besides, 3 originally separated speech fragments are now recognized to be within one segment, thus reducing the recall rate to 80%. We can also see that many detected speech fragments have imprecise segment boundaries due to the inclusion of background noise. The reason for this performance degradation is that, although the silence model can model the global distribution of the background noise, it may be too coarse to catch the local background variation within each scene or shot. In this case, a good solution is to adapt the silence model to every local video unit.

5. EXPERIMENTAL RESULTS

To evaluate the robustness and effectiveness of the proposed speaker identification system, study has been carried out by using a 2-hour long movie "Legend of the Fall", which

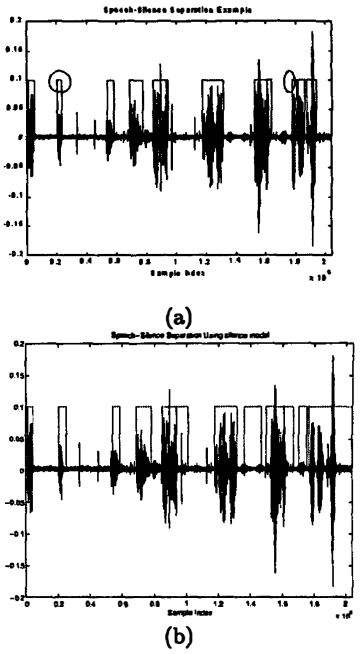

Fig. 3. Examples of silence detection by using (a) adaptive silence detection and (b) the silence model.

is a romantic, thrilling and tragic movie. Totally 4 speaker models are built corresponding to 4 key actors/actresses by using approximately 40-second training data per speaker with silence removed. Their GMM models are trained with the standard Expectation Maximization (EM) algorithm, and each model consists of 16 mixture components with a diagonal covariance matrix. A background model is also built as discussed earlier. Totally we have extracted 12 2-speaker dialogs, and identification results are reported in confusion matrix as shown in Table 1, where the 4 actors/actresses are indexed by A, B, C, D, and their respective characters in the movie are denoted by A', B', C' and D'. "Unknown" is used for any other uninterested speakers. The number in each grid, say grid (A', B), indicates the number of times that movie character A' is identified as actor B. Obviously the ideal case is that all off-diagonal grids are 0. Three parameters, namely, false acceptance (FA), false rejection (FR), and identification accuracy (IA), are calculated to evaluate the system performance. For a particular actor/character, FA = (sum of off-diagonal numbers in the corresponding row)/(sum of all numbers in the row), FR = (sum of off-diagonal numbers in the corresponding column)/(sum of all numbers in the column), and IA = $1 - FA$.

Table 1(a) gives the identification result where no adaptive silence detection is applied to incoming audio signals. Instead, a simple silence detection scheme based on a fixed energy threshold is applied. Besides, this case assumes there is only one speaker per shot. By observing this ta-

ble, we find that a couple of movie characters tend to be mis-identified as "unknown" speaker, which mainly occurs when the shot contains speeches from both speakers. In this case, since no single speaker can be properly claimed as the target one, identifying it to be "unknown" is perhaps the best solution we could offer. This table presents an average 60.7% IA and 39% FA, which cannot be considered satisfactory.

Table 1(b) gives experimental results with the proposed adaptive silence detection. As shown, the numbers on the diagonal have been increased significantly, since now speech segments can be well isolated. Although we do still have some cases where some characters are mis-recognized as "unknown", these are mainly caused by the failure of identifying some too-short speech segments. This table presents an average 83.3% IA, which is quite encouraging. Also, the FA and FR are as low as 16.7% and 15.4%, respectively. Actually we expect to get even better results if we could have more speech data from speaker C.

The identification results obtained from the silence model are tabulated in Table 1(c). As shown, a lot of false alarms have occurred due to incorrect isolation and detection of the speech segments. Only an average of 64.4% IA is achieved. The FR is 31%, which is much larger than that of case (b).

Another experiment was designed to use pure audio information without considering shot boundaries. The results are shown in Table 1(d) where an overall 74.5% accuracy is obtained. The major reason for this performance degradation is that the proposed adaptive silence detection approach is no longer always correct since it is now carried out at the scene range instead of the shot range. Because a typical movie scene will usually cover tens of shots, we can no longer assume a constant background noise level, especially when some shots have loud music background. However, it still presents better FA and FR than those in cases (a) and (c).

To summarize, better identification results could be achieved if we apply the adaptive silence detection scheme and integrate it with audiovisual cues.

Table 1. Speaker identification results: (a) w/o using adaptive silence detector, (b) using adaptive silence detector, (c) using silence model, and (d) using pure audio cue.

	A	B	C	D	Unkn	FA	IA
A'	11	0	3	2	11	59%	41%
B'	0	20	0	0	3	13%	87%
C'	1	1	4	0	2	50%	50%
D'	1	1	1	17	6	35%	65%
Unkn	0	2	2	1	32		
FR	18%	20%	60%	15%			

(a)

	A	B	C	D	Unkn	FA	IA
A'	18	1	1	0	2	18%	82%
B'	0	27	1	0	2	10%	90%
C'	0	1	9	0	2	25%	75%
D'	1	1	1	38	3	14%	86%
Unkn	0	3	2	1	40		
FR	5%	18%	36%	2.5%			

(b)

	A	B	C	D	Unkn	FA	IA
A'	15	0	1	0	0	6%	94%
B'	3	22	2	8	24	63%	37%
C'	3	0	8	2	8	62%	38%
D'	1	0	2	32	1	11%	89%
Unkn	8	0	2	2	28		
FR	50%	0%	47%	27%			

(c)

	A	B	C	D	Unkn	FA	IA
A'	17	3	1	1	5	37%	63%
B'	0	38	0	0	1	3%	97%
C'	0	1	9	0	3	31%	69%
D'	1	4	6	35	5	31%	69%
Unkn	0	13	4	0	42		
FR	5.6%	36%	55%	2.8%			

(d)

6. CONCLUSION AND FUTURE WORK

A robust speaker identification scheme based on audiovisual cues was proposed in this work. In our future work, we will explore the feasibility of building the speaker model on the fly as more speech data coming in an unsupervised mode. Another research direction is to perform a key speaker spotting, where only key speakers' speech data will be continuously collected and used to update their models.

7. REFERENCES

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

[2] S. E. Johnson, "Who spoke when ? - automatic segmentation and clustering for determining speaker turns," Proc. of 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 Trans. on Circuits and Systems for Video Technology, vol. 11, no. 4, pp. 522–535, 2001.

[5] Ying Li and C.-C. Jay Kuo, "Real-time segmentation and annotation of MPEG video based on multimodal content analysis I & II," Technical Report, University of Southern California, 2000.

[6] Ying Li, S. Narayanan, W. Ming, and C.-C. Jay Kuo, "Automatic movie index generation based on multimodal information," Proc. of SPIE, vol. 4519, pp. 42–53, August, Denver, 2001.

[7] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol. 17, no. 1-2, pp. 91–108, 1995.