SPEAKER SEGMENTATION OF INTERVIEWS USING INTEGRATED VIDEO AND AUDIO CHANGE DETECTORS

Mathieu Lagrange[†], *Luis Gustavo Martins*[*], *Luis F. Teixeira*[*], *George Tzanetakis*[†]

[†]University of Victoria {lagrange,gtzan}@uvic.ca, [*]INESC Porto {lmartins,lfpt}@inescporto.pt

ABSTRACT

In this paper, we study the use of audio and visual cues to perform speaker segmentation of audiovisual recordings of formal meetings such as interviews, lectures, or courtroom sessions. The sole use of audio cues for such recordings can be ineffective due to low recording quality and high level of background noise.

We propose to use additional cues from the video stream by exploiting the relative static locations of speakers among the scene. The experiments show that the combination of those multiple cues helps to identify more robustly the transitions among speakers.

1. INTRODUCTION

The indexing of audio recordings of meetings, interviews, lectures, or courtroom sessions is an important application which usually comprises the task of segmenting the different speakers, followed by the transcription of each individual speaker. The speaker segmentation is usually performed using only the audio recording [1, 2], some times using a multiple microphone setup [3, 4]. However, in many of these applications there is only a single microphone available for recording. Furthermore, the use of low quality recording equipment and the presence of background noise and other acoustic events will have a significant negative impact in the robustness and performance of the speaker segmentation system. In such scenarios, the use of visual cues can help improving segmentation performance, even when using low resolution video acquisition devices, nowadays relatively cheap and commonly available.

In this paper we propose the use of acoustic and visual cues for speaker turn detection. Such a combination has been proposed for other application scenarios, such as scene [5-10] and commercial detection [11] in broadcast recording and ambulatory recordings [12].

Closer to our approach, Cutler et al. propose exploiting the correlation between video and audio [13] and Fisher et al. perform speaker motion and speech association using signal-level (early) fusion [14].

In this work, we perform a late fusion of audiovisual cues by combining two speaker change detectors. The assumptions are the following: one camera and one microphone are used and the speakers are always visible, with a relatively static location. Furthermore, the algorithm is robust to gesture or facial movements, luminosity change as well as acquisition quality.

The acoustical cues are processed by an algorithm based on a causal two-stage approach, where both metric and model-based criteria are used for unsupervised speaker turn detection. Concerning the visual cues, we exploit the above described assumptions. This allows us to set up the video acquisition device in fixed and predefined positions and to use a simple segmentation scheme to efficiently detect changes among speakers. Moreover, the number of speakers is usually known beforehand. All these constraints, although not always present in typical video content analysis scenarios, are reasonable for the application scenarios that we are interested in.

The paper is organized as follows. In Section 2, we describe the algorithm used for speaker segmentation using the acoustic signal. In Section 3, we introduce the visual segmentation scheme used to detect moving regions in the scene. We then describe the combination of the boundaries detected by the two algorithms in Section 4. The recording methodology used to record the two datasets used for evaluation, and the corresponding performance results of the different speaker change detectors and the proposed combination schemes are discussed in Section 5. Discussion on future work follows in Section 6.

2. ACOUSTIC SPEAKER SEGMENTATION

The algorithm used in this paper for the speaker segmentation based on the acoustic speech signal, assumes no prior knowledge about the number of speakers or their identities. It presumes that the audio input data contains only speech, or that non-speech audio segments were already filtered by a previous audio segmentation module [15, 16]. The system consists of two main processing stages. In a first step, a metric-based approach is implemented for coarse speaker segmentation using Line Spectral Pairs (LSP) [17]. Subsequently, the Bayesian Information Criterion (BIC) is used to validate the potential speaker change points previously detected [18, 19]. The algorithm tries to recognize speaker turn points in a causal manner (i.e. without having access to the whole speech stream), with the shortest possible delay.

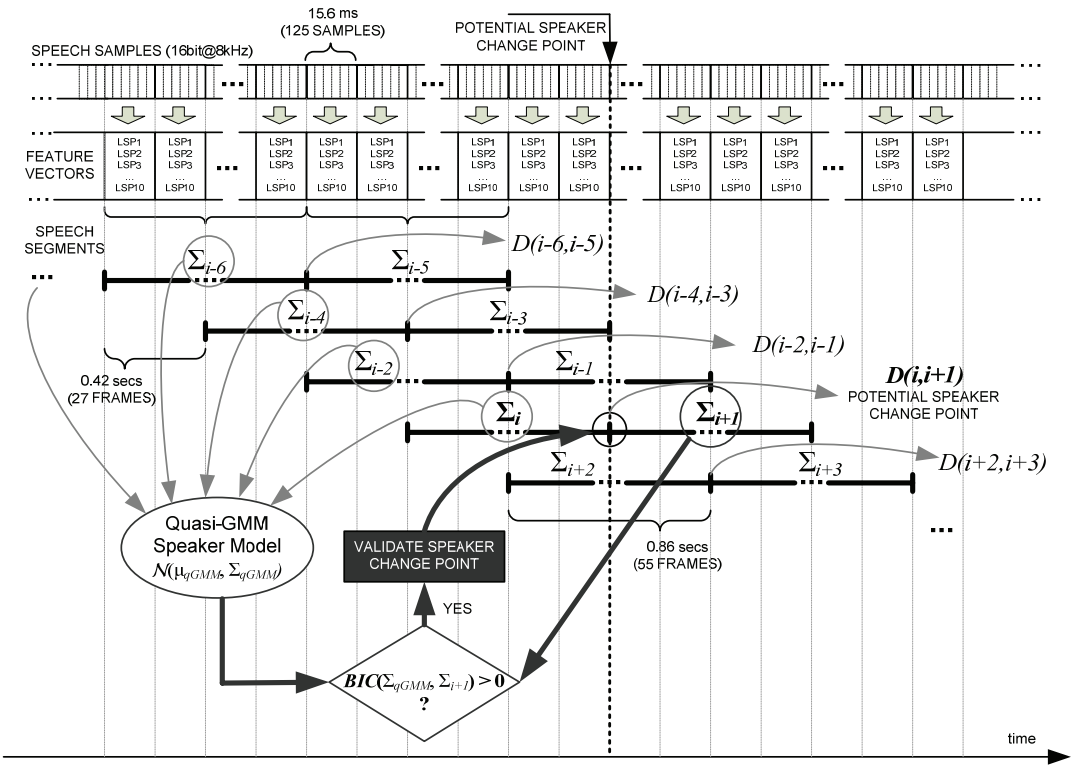

Figure 1 – Two-stage metric and model-based acoustic speaker segmentation.

A summarized system overview of the algorithm is presented below, and shown schematically in Figure 1. The interested reader can find a more detailed description in [20, 21], as well as a comparative evaluation of this algorithm's performance, where it achieved comparable results to other more computationally demanding and non-causal approaches.

The algorithm starts by down-sampling the input speech audio to 8 kHz, 16 bits mono audio format, and applies a pre-emphasis filter (using a first order FIR filter with frequency response $H(z) = 1 - 0.93z^{-1}$). The speech stream is then divided into analysis frames of about 16 ms duration, without overlap. From each audio frame 10-order LSP features are extracted [17].

In the first stage, speaker change detection is coarsely performed using a metric-based approach to calculate the distance between consecutive and non-overlapping speech segments (see Figure 1). Each speech segment includes 55 speech frames (corresponding to about 1 sec.), which is the minimum number of frames that prevents an ill-conditioned covariance matrix for 10^{th}-order LSPs.

Assuming that the LSPs are Gaussian distributed, each speech segment can be represented by the covariance matrix Σ of a multivariate Gaussian model. The Kullback-Leibler (K-L) divergence shape distance [17] is then used to estimate the distance between two subsequent speech segments (represented as *D* in Figure 1).

A potential speaker turn point is detected between two segments (*i, i+1)* whenever its divergence distance *D*(*i, i+1)* is a prominent local maximum (i.e. when compared with its preceding and following segment distances). In this paper, the parameter α was set to 0.8, as defined in [22].

As an attempt to reduce the false alarm rate that results from the metric-based approach (which is tuned for lower miss detections as opposed to few false alarms), BIC [18, 19] is used in the second stage to validate any potential speaker change point detected by the coarse segmentation procedure. However, BIC is well known for suffering from insufficient model estimation traits when dealing with small amounts of data, as happens from just using data from two speaker segments. In an attempt to circumvent this, as new speech segments are received, and while no potential speaker change is detected, all the arriving data is used to incrementally update the current speaker model. At the presence of a potential speaker change, this will allow better model estimates, potentially increasing the accuracy of the BIC validation. In order to implement such an approach, we make use of a solution based on quasi-GMM modeling, a non-iterative technique that allows causal operation with a reasonable accuracy, and originally proposed in [23]. In this paper we set the BIC penalty parameter λ to 0.6, as used in [22].

3. VISUAL SCENE SEGMENTATION

Most of existing research for the segmentation of video content using visual cues uses intensity or colorimetric changes [7]. Here such cues are not useful as the camera is supposed to be static. A very common approach relies on face detection to identify potential speakers [24]. Alternatively, we propose to exploit the relative static locations of speakers within the scene avoiding complex face and mouth/lips detection algorithms.

In this paper, we limit ourselves to scenarios with only two speakers, such as interviews or lectures. However, the proposed video scene segmentation can easily be generalized to a larger number of speakers, provided that the locations of each speaker are previously given.

In order to identify the person speaking we will be using motion estimation. Moreover, since we are considering only sequences with two persons facing the camera it should be safe to assume that there are two distinct visual regions associated with each person (Figure 2).

The separation of these regions is defined by a boundary β. For simplicity this is kept as a vertical straight line splitting the image in two halves. It should however be noted that the two regions could be arbitrarily complex and defined, for example by binary masks. This scheme can be generalized to several speakers, by previously selecting several areas of interest.

The motion was estimated using a traditional block matching algorithm [25]. For all test sequences, 4x4 blocks were used, with 1x1 shifts and a 4x4 maximum search range. Based on the motion, two features were then computed: motion centroid and motion balance. A dense motion map could also be estimated based on optical flow techniques but, given the usually poor image quality, it would not add relevant information.

The motion centroid is calculated using the absolute value of each component in each pixel as a weight. It can be defined as follows:

$$(C_x, C_y) = \left(\frac{\sum_I x|\Delta_x|}{\sum_I |\Delta_x|}, \frac{\sum_I y|\Delta_y|}{\sum_I |\Delta_y|} \right) \quad (1)$$

where I is a set containing all pixels of the image, and Δ_x and Δ_y are horizontal and vertical components of the motion, respectively. To avoid excessive jitter, a median filter of order 5, *i.e.* the current and the past four estimates, is applied to estimate a more accurate centroid.

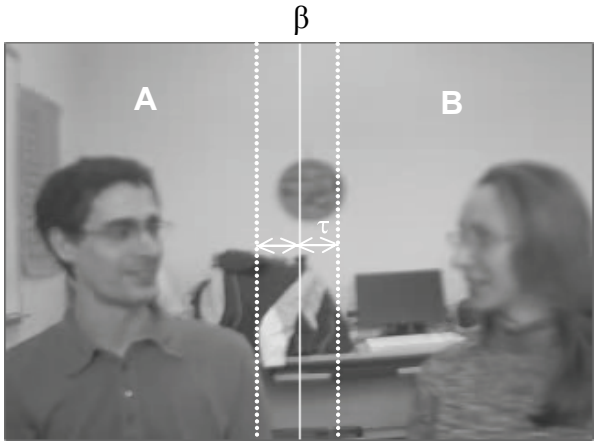

Figure 2 – Definition of visual regions.

In a first approach, given the motion centroid for each new frame, a simple classifying algorithm can be applied. It consists of a threshold matching the previously defined boundary but with added hysteresis as shown in Figure 2. Instead of marking a frame as a speaker change when the centroid crosses the boundary, it is only marked as such if it crosses a new boundary displaced by a tolerance τ. The tolerance was set, for all sequences, as 5% of the frame width.

3.1. Adaptive Thresholding

Even if the speaker is always moving while speaking, the amount of movement depends on each speaker, and may change across time. To adapt to the characteristic of each speaker and to allow change of dynamics over time, we propose to use an adaptive thresholding scheme.

We initialized a threshold for each region (τ_l and τ_r) as τ. To update those thresholds, we propose to divide the evolution of the x-location of the centroid into left and right sections whose boundaries $b_l(k)$ and $b_r(k)$ respectively follow:

$$C_x(b_l(k)-1) > \beta \wedge C_x(b_l(k)) < \beta \quad (2)$$

and

$$C_x(b_r(k)-1) < \beta \wedge C_x(b_r(k)) > \beta. \quad (3)$$

In order to obtain more relevant boundaries, a median filter of order 30 is applied to the observed data. At each right boundary $b_r(k)$, the right threshold is updated as follows:

$$\tau_r = \tau + (1-\alpha)\cdot(\tau_r - \tau) + \alpha \cdot m_r / 2 \quad (4)$$

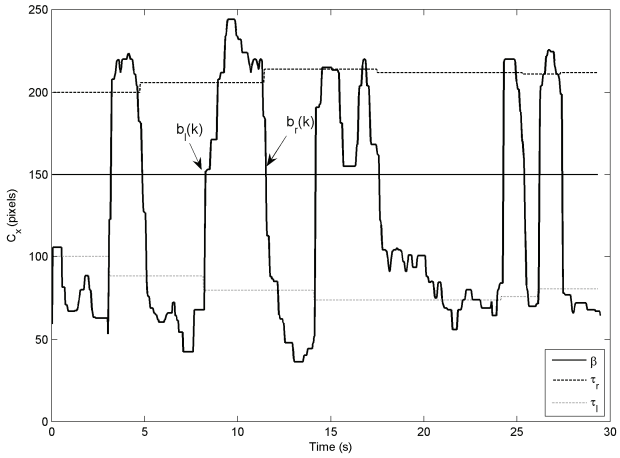

Figure 3 – Centroid-based decision with adaptive thresholding.

where α=0.5 and $m_r = \max_{i\in[b_l(k),\ldots,b_r(k)]}(C_x(i)-\beta-\tau)$. Right-speaker change detection is triggered if $C_x(i) > \beta + \tau_r$ and if the last detection was a left-speaker detection. Figure 3 shows the evolution of *Cx* and the decision boundaries over time for one of the test sequences. The left-speaker detector is triggered similarly. One can notice that a speaker labelling can easily be achieved by considering which detector is triggered.

4. MULTIMODAL SEGMENTATION ALGORITHM

While the acoustic and video speaker change detectors could be used independently, performing a combination will help improve the performance. We perform "decision in-decision out" fusion [26] of the audio and video cues by combining the speaker changes boundaries detected using the two algorithms presented in the previous sections.

The audio and video modalities can be considered to observe the same behavior of the audiovisual scene, but in a different way. Indeed, people tend to move their bodies, arms and lips before producing any sounds and the first sounds produced are usually non speech vocalizations such as breath, etc. The speech segments happen lastly, usually with half a second delay.

The visual change detector is then more likely to be fired before the audio one, as can be observed on Figure 4. Also, this last audio detector is more likely to detect the correct boundary but with a higher false alarm rate due to the presence of non speech sounds and background noise.

We aim at designing a causal combination algorithm that takes these two constraints into account. At a given time t advanced at video frame rate (30 fps), we seek for audio and video speaker change boundaries, respectively noted $s_a(k)$ and $s_v(k)$, such that $abs(s_{(a,v)}(k)-t)$ is below a given threshold δ, set to 1.6 seconds. Only those boundaries are now considered. We then seek for the combination

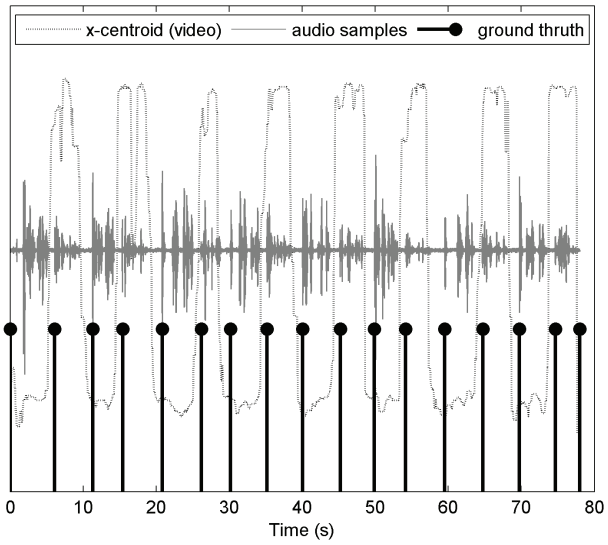

Figure 4 – Properties of the different modalities.

of an audio boundary $s_a(k_{\min})$ such that $abs(s_a(k)-t)$ is minimal and a video boundary such that

$$s_v(k) - s_a(k_{\min}) < \delta/2$$

and

$$abs(s_v(k) - s_a(k_{\min}) + \delta/2) < abs(s_v(k) - s_a(k_{\min}+1)).$$

If those two conditions are met, $s_a(k_{\min})$ is used as a combined boundary and the two boundaries are discarded. An example of the resulting boundaries is plotted on Figure 5.

5. EXPERIMENTAL RESULTS

The performance of the acoustic, visual and multimodal segmentation algorithms was evaluated in a dataset created for this task and described in the next subsection. The experiments consist of running the different algorithms and combinations of algorithms for each sequence of the dataset and comparing the resulting segmentation using a given set of metrics.

5.1. Experimental Setup

In order to assess the performance of the proposed system, an audiovisual dataset was created by recording two-speaker conversations using a consumer web camera equipped with a microphone (refer to Figure 2 for an example of the quality of the images captured). The dataset comprises 14 audiovisual sequences of 5 different speakers, with a mean duration of 60 seconds. It is divided in two smaller datasets with different scenarios, one formal and the other closer to a real interview setting.

In the first scenario, the speakers are asked to alternatively read 8 poetry sentences. The total number of speaker

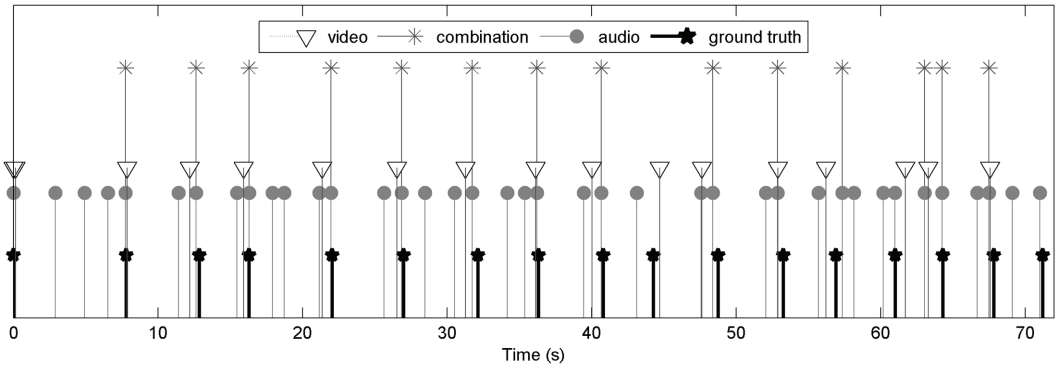

Figure 5 – Combined speaker boundaries.

changes for each sequence is therefore 15, plus the 2 edges corresponding to the start and end of the speech sequence (see Figure 5). The speaker segments in this dataset have a mean duration of 4.14 seconds, with a standard deviation of 1.21 seconds. In the second one, the speakers are asked to improvise an interview without any prior preparation. The total number of speaker changes for each sequence is around 8, plus the 2 edges. The speaker segments in this dataset have a mean duration of 3.53 seconds, with a standard deviation of 6.68 seconds. Interested researchers can obtain the dataset by emailing the authors.

In this work we are interested in detecting speaker turn points, based on the acoustical speaker changes. So the boundaries that we are interested in correspond to those of the audio. The two datasets have therefore been hand-labeled using the audio stream as the reference.

Few software platforms enable the combination of multimodal information in a flexible and efficient manner. The proposed implementation is a first attempt to achieve this goal. Both the audio and video based detection algorithms were implemented using an experimental version of Marsyas, an open-source C++ software framework[1] [27].

5.2. Evaluation Methodology

Two pairs of metrics have been used to assess the performance of the different speaker change detection implementations presented in this paper. On the one hand, one may use the false alarm rate (*FAR*) and the miss detection rate (*MDR*) defined as:

$$FAR = \frac{FA}{GT + FA} \qquad MDR = \frac{MD}{GT} \tag{5}$$

where *FA* denotes the number of false alarms, *MD* the number of miss detections, and *GT* stands for the actual number of speaker turns, i.e. the ground truth. A false alarm occurs when a speaker turn is detected although it does not exist, a miss detection *MD* occurs when the process does not detect an existing speaker turn. On the other hand, one may employ the precision (*PRC*) and recall (*RCL*) rates given by:

$$PRC = \frac{CFC}{DET} \qquad RCL = \frac{CFC}{GT} \tag{6}$$

where *CFC* denotes the number of correctly found changes and *DET* is the number of the detected speaker changes. For the latter pair, another common metric is the F_1 measure

$$F_1 = \frac{2.PRC.RCL}{PRC + RCL} \tag{7}$$

that admits a value between 0 and 1. Higher values of F_1 indicate that a better performance is obtained. Additionally, and as a reference, the following relationships between the pairs (*FAR, MDR*) and (*PRC, RCL*) hold:

$$\begin{aligned} MDR &= 1 - RCL \\ FAR &= \frac{RCL \cdot FA}{DET \cdot PRC + RCL \cdot FA} \end{aligned} \tag{8}$$

To compute those metrics, we used a tolerance between the detected boundary and the hand-labeled one of 0.8 seconds.

5.3. Experiments

The experimental results for the poetry reading dataset are summarized on Table 1. The audio segmentation algorithm parameterized as described in Section 2, performs as expected, with a low MDR and a high FAR, leading to an F_1-Measure around 50%. Concerning the use of the video cues, two detectors are considered. The first uses a constant threshold (presented as "video" in the tables) and the second uses the adaptive scheme proposed in Section 3.1 (presented as "adaptive video" in the tables). Their performance characteristics are of a balanced FAR and MDR. The adaptive scheme outperforms the constant one and lead to an improvement of 7% in terms of F_1-Measure.

Compared to the respective performance of the two detectors, the combined one significantly decreases the FAR, while averaging the MDR, leading to an improvement of

[1] `http://marsyas.sourceforge.net`

	FAR		MDR		Recall		Precision		F_1-Measure	
	mean	*stddev*	*mean*	*stddev*	*mean*	*stddev*	*mean*	*stddev*	*mean*	*stddev*
audio	63.52	4.43	7.63	10.18	92.37	10.18	34.6	5.36	50.21	6.75
video	43.62	6.41	43.82	14.64	56.18	14.64	41.94	10.57	47.83	11.83
adaptive video	34.81	6.04	37.16	12.39	62.84	12.39	53.66	11.16	57.88	11.72
audio+video	36.68	6.84	30.02	7.65	69.98	7.65	54.72	8.06	61.17	6.95
audio+adaptive video	25.05	7.73	25.4	17.3	74.6	17.3	68.01	13.3	71.03	14.83

Table 1 – Performance results for the poetry reading scenario

	FAR		MDR		Recall		Precision		F_1-Measure	
	mean	*stddev*	*mean*	*stddev*	*mean*	*stddev*	*mean*	*stddev*	*mean*	*stddev*
audio	53.38	12.18	20.71	24.06	79.29	24.06	39.63	10.16	51.4	12.63
video	40.3	15.89	24.54	22.06	75.46	22.06	52.18	12.92	59.23	12.83
adaptive video	28.01	10.9	39.73	31.28	60.27	31.28	56.96	19.57	56.7	25.16
audio+video	36.26	10.77	35.26	24.99	64.74	24.99	51.84	12.08	56.04	15.76
audio+adaptive video	21.05	13.24	43.38	24.68	56.62	24.68	66.02	17.7	58.32	21.32

Table 2 – Performance results for the interview scenario

20% over the audio performance in terms of F_1-Measure. In this formal setting, the proposed detection scheme performs reasonably well.

The results while using the informal interview setting are summarized on Table 2. The audio detector achieves similar performances with a higher MDR rate. The adaptive thresholding scheme fails to adapt nicely and performs globally worse than the standard one. Yet, it achieves a lower MDR, which explains why the combination detector that makes use of the adaptive scheme still outperforms the one with the standard video detector.

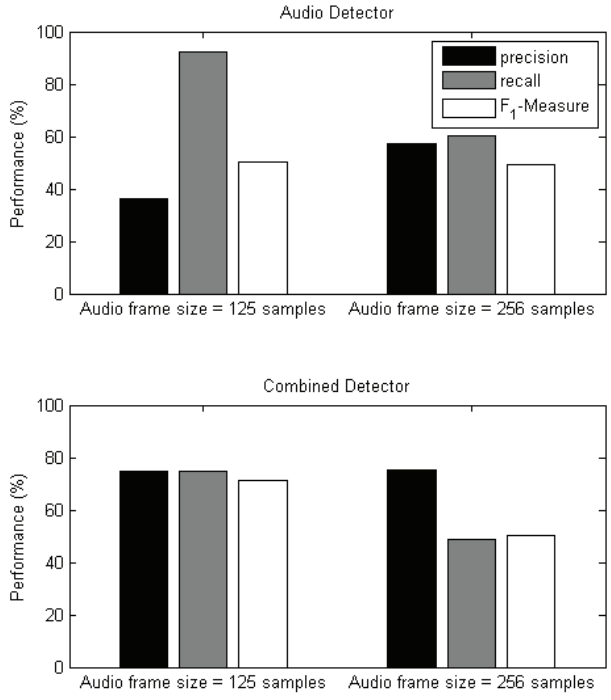

Figure 6 – Performance results for two configuration settings of the audio speaker change detector

From this, we can conclude that the proposed combination scheme takes advantage of the complementary characteristics of the performance statistics of the two combined detectors. This statement can be further asserted using the following experiment. We tested the combination scheme using an audio detector scheme with longer frame sizes (i.e. 256 samples instead of the 125 samples per frame, as defined in Section 2) in order to achieve a more balanced ratio between MDR and FAR performance. The comparative results for the poetry reading case are plotted on Figure 6. We can see at the bottom plot that, although the use of the proposed combination scheme still improve the F_1-Measure over the audio detection, the improvement is no more significant when the two detectors both have balanced statistical performance properties, i.e. their respective MDR/FAR≈1. To achieve better results in this aspect, the combined detector should be built on two stages. First, the sets of highly reliable audio and video boundaries should be selected. This would allow us to handle the case where only one detector is relevant, such as when the speaker is out of scope, resulting in a lower MDR. Next, an algorithm similar to the one proposed in the previous section could be used to handle non trite cases.

6. FUTURE WORK

These experiments show that the proposed scheme is effective for the speaker segmentation of two-speaker interviews. We plan to work on more diverse setups such as meetings and court case recordings where the video can help even more due to the diversity of speaker levels.

We also intend to incorporate confidence indicators of the detected boundaries. This approach would be more versatile and would allow us to use detectors with similar statistical performance characteristics.

We also plan on using more sophisticated video detector algorithms, where the each speaker's relevant image areas would be automatically detected (e.g. mouth and chin) in order to correlate their motion activity with the speech signal.

7. CONCLUSION

The segmentation of audiovisual interviews among different speakers has been studied. A segmentation of the video scene is performed using motion cues to assist an acoustic speaker change detector. Those two detectors are combined using a scheme that takes advantage of their complementary statistical performance characteristics.

It was shown that the use of video cues is helpful to overcome potential problems and also provide an automatic labeling without any need for any speaker recognition process.

8. ACKNOWLEDGMENTS

The authors want to thank Jennifer Murdoch, Randy Jones, Graham Percival, Adam Tindale, and Manj Benning for their readings and useful comments.

9. REFERENCES

[1] L. Lu and H. Zhang, "Real-time unsupervised speaker change detection," in *16th International Conference Pattern Recognition*, Quebec City, Canada, 2002.

[2] L. Lu and H.-J. Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," in *10th ACM International Conference on Multimedia*, 2002, pp. 602 – 610.

[3] T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," 2001, pp. 107-110.

[4] R. Prasad, L. Nguyen, R. Schwartz, and J. Makhoul, "Automatic transcription of courtroom speech," in *7th International Conference on Spoken Language Processing (ICSLP2002)*, Colorado, USA, 2002, pp. 1745-1748.

[5] 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, pp. 522-535, April 2001.

[6] Z. Liu and Y. Wang, "Major Cast Detection in Video Using Both Speaker and Face Information," *IEEE Transactions on Multimedia,* vol. 9, pp. 89-101, January 2007.

[7] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia Content Analysis - Using Both Audio and Visual Clues," *IEEE Signal Processing Magazine,* vol. 17, pp. 12-36, November 2000.

[8] H. Sundaram and C. Shih-Fu, "Video scene segmentation using video and audio features," in *IEEE International Conference on Multimedia and Expo (ICME2000)*, New York, USA, 2000, pp. 1145-1148.

[9] J. Nam and A. H. Tewfik, "Combined Audio and Visual Streams Analysis for Video Sequence Segmentation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany, 1997.

[10] Y. Li, S. Narayanan, and C.-C. J. Kuo, "Content-Based Movie Analysis and Indexing Based on Audiovisual Cues," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 14, pp. 1073-1085, August 2004.

[11] M. Covell, S. Baluja, and M. Fink, "Advertisement Detection and Replacement using Acoustic and Visual Repetition," in *International Workshop on Multimedia Signal Processing (MMSP06)*, Victoria, Canada, 2006.

[12] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP99)* Phoenix, USA, 1999, pp. 3037-3040.

[13] R. Cutler and L. Davis, "Look Who's Talking: Speaker Detection using Video and Audio Correlation," in *IEEE International Conference on Multimedia and Expo*, New York, USA, 2000.

[14] J. W. F. III and T. Darrell, "Speaker Association With Signal-Level Audiovisual Fusion," *IEEE Transactions on Multimedia,* vol. 6, June 2004.

[15] E. Scheirer and M. Slaney, "Construction And Evaluation Of A Robust Multifeature Speech/music Discriminator," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, 1997.

[16] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *Journal of the Audio Engineering Society,* vol. 52, pp. 724-739, July/August 2004.

[17] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE,* vol. 85, September 1997.

[18] S. Chen and P. S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[19] P. Delacourte and C. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication,* vol. 32, pp. 111-126, 2000.

[20] M. Kotti, E. Benetos, C. Kotropoulos, and L. G. Martins, "Speaker Change Detection using BIC: A comparison on two datasets.," in *2006 Second International Symposium on Communications, Control and Signal Processing*, Marrakech, Morocco, 2006.

[21] M. Kotti, L. G. Martins, E. Benetos, J. S. Cardoso, and C. Kotropoulos, "Automatic Speaker Segmentation Using Multiple Features and Distance Measures: A Comparison of Three Approaches," in *ICME 2006 - IEEE 2006 International Conference on Multimedia & Expo*, Toronto, Canada, 2006.

[22] M. Kotti, L. G. Martins, E. Benetos, J. S. Cardoso, and C. Kotropoulos, "Automatic Speaker Segmentation Using Multiple Features and Distance Measures: A Comparison of Three Approaches," in *IEEE 2006 International Conference on Multimedia & Expo (ICME2006)*, Toronto, Canada, 2006.

[23] L. Lu and H. Zhang, "Real-time unsupervised speaker change detection," in *Proc.16th Int. Conf. Pattern Recognition*, Quebec City, Canada, 2002.

[24] J. M. Rehg, K. P. Murphy, and P. W. Fieguth, "Vision-Based Speaker Detection Using Bayesian Networks," in *Computer Vision and Pattern Recognition*, 1999, pp. 110-116.

[25] M. Bierling, "Displacement estimation by hierarchial blockmatching," in *SPIE Conference Visual Communications and Image Processing*, 1988, pp. 942-951.

[26] B. V. Dasarathy, "Sensor Fusion Potential Exploitation - Innovative Architectures and Illustrative Applications," *Proceedings of the IEEE* vol. 85, pp. 24-38, January 1997.

[27] G. Tzanetakis and P. Cook, "MARSYAS: a framework for audio analysis," *Organized Sound, Cambridge University Press* vol. 4, 2000.