

Speaker change detection using joint audio-visual statistics

Giridharan Iyengar and Chalapathy Neti

IBM T. J. Watson research Center

Yorktown Heights NY 10598

{giyengar,cneti}@us.ibm.com

Abstract

In this paper, we present an approach for speaker change detection in broadcast video using joint audio-visual scene change statistics. Our experiments indicate that using joint audio-visual statistics we achieve better recall without loss of precision as compared to purely audio domain approaches for speaker change detection.

1 Context

In the emerging multimedia and ubiquitous computing applications there is a general awareness of the need to perform joint audio-visual analysis. Examples of such applications include speechreading [1, 6, 7, 8, 12] and general multimodal communication[3].

One of the fundamental goals in this emerging discipline is to model audio-visual events for the purposes of understanding, indexing and managing multimedia content. There is significant value in generating automatic transcripts and summarization of such audio-visual content. An example of such would be generating a textual transcription/keywords automatically from the speech portions of the audio-visual content. This holds the potential of enabling automatic text-based indexing and retrieval. Good segmentation based on speaker identities, channel and environmental conditions is a core requirement for current state-of-the-art speech recognition and automatic transcription/retrieval systems. See citations [13, 14] for systems that use audio and video information for multimedia annotation, transcription and retrieval. We note here that in these systems, while both audio and video segmentation is performed, both these modalities are treated independently as opposed to our work.

Purely audio based speaker segmentation have received quite a bit of attention recently[2]. These techniques require either an explicit silence interval or need large audio samples to accurately detect segment boundaries. We hypothesize that there is much to be gained from joint audio-visual analysis. The performance of these audio-based techniques can be further improved by exploiting the joint statistics between the audio stream and its associated video. For example, take the scenario of a newscast. There is significant correlation between audio and video speaker changes. Frequently, the video scene change follows shortly after an audio change. In such a scenario, gathering the joint audio-visual statistics and leveraging from this to generate more accurate audio-segmentations (which in turn is desirable for accurate speech transcription and retrieval) seems to be of interest. For example, in 31 minutes of a television panel discussion that we analyzed, 66.7% of the audio speaker changes were immediately followed (within 4 seconds) by a corresponding video change. In addition, about 60% of the audio changes occurred before the video change and the remaining after the video change.

In the remaining parts of the paper, we briefly discuss one of the state of the art audio segmentation algorithm, discuss the video scene change detection algorithm we use, and discuss the different strategies for integrating the video information with audio information and present our experimental results.

2 BIC based audio change detection

Bayesian Information Criterion (BIC) is a minimum description length principle that is rather effective in detecting speaker changes in audio. Briefly, it looks at a window of audio data and evaluates if the window contains speech fragments from one or more speakers. In order to differentiate between multiple hypotheses, it uses the minimum description length principle. The BIC criterion is defined as

$$BIC(M) = \log L(X, M) - \frac{\lambda}{2} \#(M) \times \log(N) \quad (1)$$

where $X = x_i : i = 1, \dots, N$ is the data set we model, $M = M_i : i = 1, \dots, K$ are the candidate models that we are evaluating and assuming that we are maximizing the likelihood of the data set under the model, we obtain $L(X, M)$. $\#(M)$ is the number of model parameters (indicating model complexity) and λ is the penalty weight. We use the BIC criterion to detect speaker changes by evaluating the difference between the BIC values of the one speaker hypothesis versus two speakers in a given temporal window. The difference between the BIC values of the two models is indicative of the preferred model at a window. If the two-speaker model is preferred in a temporal window, then a segmentation is marked in that window. The exact location of the segmentation is found by the likelihood maximization in that window. For further details, please refer to[2].

We note here that we can change the penalty (λ) parameter to be other than unity to achieve different detection thresholds. While, we cannot strictly call λ values other than 1 as BIC, we use the term loosely to describe this algorithm for all λ values that we consider.

3 Histogram based video scene change detection

In both content-based retrieval and efficient video coding contexts, accurate identification and location of scene changes in video is important. This enables generation of key frames, storyboards, summaries [9, 10, 17, 15, 16] and also efficient compression, bandwidth allocation and related coding issues in the MPEG-2 and MPEG-4 contexts[5, 11].

Video scene change algorithms have been a topic of active research [4, 15]. Recent techniques reported in literature[4] report good scene change detection performance. Most scene change techniques compare adjacent frames and if the difference exceeds a threshold, it is considered a scene change. This is a computationally expensive process since scene changes in video occur much less frequently. For example, even if there is a scene change one every second, there are 30 frames per second in NTSC rate video, making 29 comparisons non-informative. A simple sub-sampling based scheme reduces the number of comparisons but at the cost of accuracy, since the frame at which the scene boundary occurs can be easily missed in sub-sampling. In this paper, we outline a new algorithm for scene change detection via hierarchical temporal sub-sampling. This algorithm can be applied to any scene change detection scheme. While the details of the scene change detection algorithm are the topic of another paper, we note here that in our experiments, the proposed algorithm achieves similar scene change performance at approximately 15% computational complexity of the current scene change detection techniques.

3.1 Hierarchical Scene Change Detection

The proposed hierarchical scene change algorithm can be used with any frame-by-frame scene change detection technique. As in sub-sampling based schemes, it starts with a large temporal separation between frames (algorithm parameter N) that are compared. Once a scene change

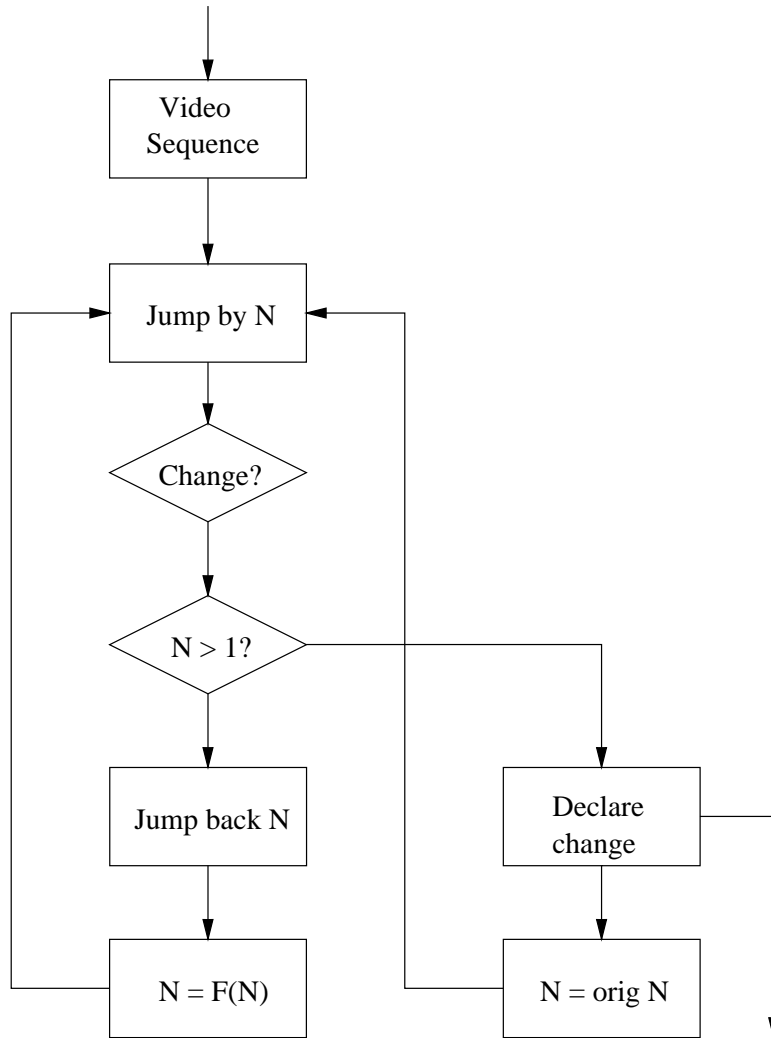


Figure 1: Top-level algorithm for Video Scene Change detection

is detected, it refines the separation N by a function $F(N)$. This function is left to implementation with the only restriction that it should reduce N and finally stop at $N = 1$. Once the precise location of the scene change is determined (when $N = 1$), the algorithm records the scene change and resets N back to its original value. Figure 1 illustrates the top-level algorithm employed by our hierarchical scene change detector.

We implemented this algorithm in the MPEG framework, exploiting the Group-of-Pictures (GOP) structure that is present in the MPEG stream. We start by comparing I-frames that are several GOPs away, reducing the temporal separation by half each time till the frames that are compared are the I-frames of adjacent GOPs. At this point, we compare every pair of adjacent frames within a GOP to precisely locate a scene change. We handle each type of scene change differently in our algorithm. For abrupt scene changes, the scene change algorithm detects only one frame in the sequence where the change occurs. This frame is marked as a key-frame and its time index is recorded. For gradual scene changes, the scene change algorithm flags a set of frames at which the frame-difference is exceeded and we record the time index of the midpoint of this set as the scene change and record the starting and the ending frames as key-frames.

The particular scene change detection scheme we implemented for our experiments compares the KL divergence between the normalized color histograms of the two frames and flags a scene change if the difference exceeds a threshold. The KL divergence between two discrete

densities is defined as

$$D(p||q) = \sum_{i=0}^N p(i) \log \frac{p(i)}{q(i)} \quad (2)$$

where the densities p and q have N discrete bins. We note here that the KL divergence is not a true distance metric since it does not satisfy the triangle inequality. However, it is an adequate dissimilarity measure for our purposes.

We note a few advantages in our hierarchical temporal sub-sampling technique. First, take the case of gradual scene changes. A gradual scene change is usually implemented as a linear combination of two different sequence of images, where one scene fades and the second scene gradually replaces it. The intermediate images combine characteristics of both the scenes. At the other extreme, in an abrupt scene change the two adjacent frames that re compared belong to two different scenes. In a frame-by-frame comparison technique, a single threshold for both abrupt and gradual scene changes does not work as well [15, 9]. A separate lower threshold is usually adopted for handling gradual scene changes. Even so, gradual scene changes are often missed. In our scheme, since we begin by comparing frames that are temporally farther apart, both abrupt and gradual scene changes manifest similarly at higher levels in the hierarchy. We start with a high threshold at coarse comparisons and systematically lower the threshold as we do finer comparisons. Also, since it is known from coarse level comparisons that a scene change has occurred and only the precise location needs to be determined, we have found empirically that it is easier to catch gradual scene changes in our method.

In addition, in a frame-by-frame comparison method, events such as camera flashes can masquerade as a scene change since these methods rely on sudden change in image characteristics which also manifests during a flash event. Such events last one to two frames typically. Sub-sampling methods are robust to camera flashes since only a fraction of the image frames are considered and the likelihood that a flash event occurs in a subsampled frame is correspondingly low. Since our technique is similar to sub-sampling, this technique has similar robustness to flash events. we note here that our algorithm achieves similar scene change performance at approximately 15% computational complexity of the current scene change detection techniques.

4 Joint audio visual speaker change detection

Once the video scene changes in a particular audio-visual stream are estimated using the above algorithm, we proceed with the joint audio-visual speaker change detection. The basic speaker change algorithm that we use is the BIC-based technique mentioned earlier in section 2. We recall that the BIC-based speaker change detection scheme has a penalty term λ that can be thought as encoding our prior belief about a speaker change in the temporal window under consideration. Based on our formulation we can make the following observation: if λ is higher than unity, a one-speaker hypothesis is preferred and when λ is less than unity, a two-speaker hypothesis is preferred, prior to any likelihood estimation. In our fusion experiments, we manipulate this prior-belief, albeit in an ad-hoc manner at the time of this paper.

Based on the above observations, we formulate the following two audio-visual fusion strategies. First, we lower the λ parameter below unity whenever the temporal window under consideration contains a visual change and raise it above unity otherwise. The amount by which we lower or raise the λ term is determined empirically to be 0.5. A principled mechanism to determine the value of λ parameter would be equate it to the estimated *prior* probability of an audio change given a video change (prior estimated from training data). However, given the sparseness of our dataset (4 sequences, total approximately 1 hour of video) at the time of these experiments we cannot reliably estimate the *prior*.

Seq	Audio Changes	Video Changes
Cspan	58	49
CNN1	26	46
CNN2	39	83
CNN3	17	54

Table 1: Number of actual audio and video changes in the test data

In the second approach, we compute the joint audio-visual change histogram which can be thought of as $P(A(t)/V(t+k))$ i.e, the probability of an audio change given a video change has occurred k seconds after the audio change. This histogram is estimated from one of the sequences as training data. We then use this joint statistics to selectively change λ parameter. Given the temporal window under consideration and the location of a video change closest to the window, we compute the cumulative probability of an audio change occurring in a given window from the estimated joint pdf and use this to change the λ parameter. It can be seen that the first scheme is a simplified version of the second scheme where instead of an estimated (albeit crudely) joint pdf, we implicitly assume a uniform joint pdf.

In our final approach, we employ decision fusion to merge the audio and video speaker change statistics. The BIC criterion is estimated independently in the time window under consideration. This decision is then merged with an independent decision based on the joint audio-visual statistics and the visual scene change within the same window. The relative weights given to BIC and joint AV statistics are 67% and 33% respectively. We note that we did not experiment extensively to arrive at these weights. Further experimentation is needed to determine the optimal weighting scheme.

We note here that the performance of the first scheme is not comparable to either the joint pdf or the decision fusion scheme and hence is omitted from our discussions.

Without access to the video scene change information, the only possibility is to change the BIC penalty (λ) parameter uniformly for the entire sequence. We note that such a scheme would sacrifice recall for precision or vice versa as can be seen from our experiments.

5 Experiments

As baseline experiments, we perform the audio speaker change detection using the BIC alone at $\lambda = 1$ and $\lambda = 0.5$. Table 1 below shows the actual number of audio and video scene changes in the data for the 4 Television news sequences that we analyzed. The *Cspan* sequence is a panel discussion and the other 3 *CNN* sequences are broadcast news segments. *Cspan* is 31 minutes long and *CNN1* through *CNN3* are between 8 and 14 minutes long.

Tables 2 and 3 show the results of these experiments. The BIC only results are in Table 2 and the two fusion strategies are detailed in Table 3. For each experiment, we list the precision and recall values.

In the Table 3, *Decis* refers to the decision fusion technique and *Joint pdf* is the scheme with joint pdf estimation technique. We note here that the last sequence (CNN3) was atypical in that it had much fewer scene changes compared to the rest even though the sequence durations were comparable. Each scene in CNN3 was much longer than either the training set or the test set. The ROC curve for BIC at $\lambda = 1$ and the decision fusion is shown below in Figure 2. Clearly, we see the advantage in using the joint AV statistics for decision fusion over pure audio.

These experiments are at a preliminary stage with a small number of test sequences. How-

Seq.	BIC $\lambda = 1$		BIC $\lambda = 0.5$	
	P	R	P	R
Cspan	95.0	65.5	20.4	86.2
CNN1	86.76	83.87	41.09	96.77
CNN2	89.74	87.5	31.4	95.0
CNN3	88.23	83.33	23.61	94.44

Table 2: Precision/Recall Performance of the BIC-only audio scene change criterion

Seq.	Joint		Decis.	
	P	R	P	R
Cspan	97.50	67.24	95.71	72.41
CNN1	84.84	90.32	86.5	93.54
CNN2	68.4	97.5	89.23	95.0
CNN3	55.17	88.88	89.7	94.44

Table 3: Precision/Recall Performance of the 2 fusion techniques

ever, at this early stage we do note encouraging improvements over plain audio-only speaker change detection. Using these joint statistics, we note that we gain in recall without loss of precision.

6 Conclusions

In this paper, we presented two approaches for integration of audio and video information for speaker change detection. While there are other papers in the literature where audio and video modalities are considered, we note that to our knowledge this work for speaker change detection using joint statistics is the first of its kind.

From our preliminary experiments we note the advantage of using visual information to perform better speaker change detection. Changing the thresholds of BIC alone does increase recall but at a significant loss of precision. However, using the joint statistics for decision fusion results in a higher recall at no degradation of precision. This is a clear advantage that we gain from using the joint statistics. While the initial results are encouraging, the data set we used is relatively small and that needs to be addressed. In addition, a more principled mechanism for integrating the audio and video information needs to be formulated.

References

- [1] C. Bregler, S. Manke, H. Hild, and A. Waibel. Bimodal sensor integration on the example of speech-reading. In *Intl. Conference on Neural Networks*, pages 667–671. IEEE, May, 1993.
- [2] Scott S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. *Intl. Conf. On Acoust., Sp., and Sig. Proc.*, 1998.
- [3] Tshuan Chen and Ram R. Rao. Audio-visual integration in multimodal communication. *Proc. IEEE*, 86(5):837–852, May 1998.

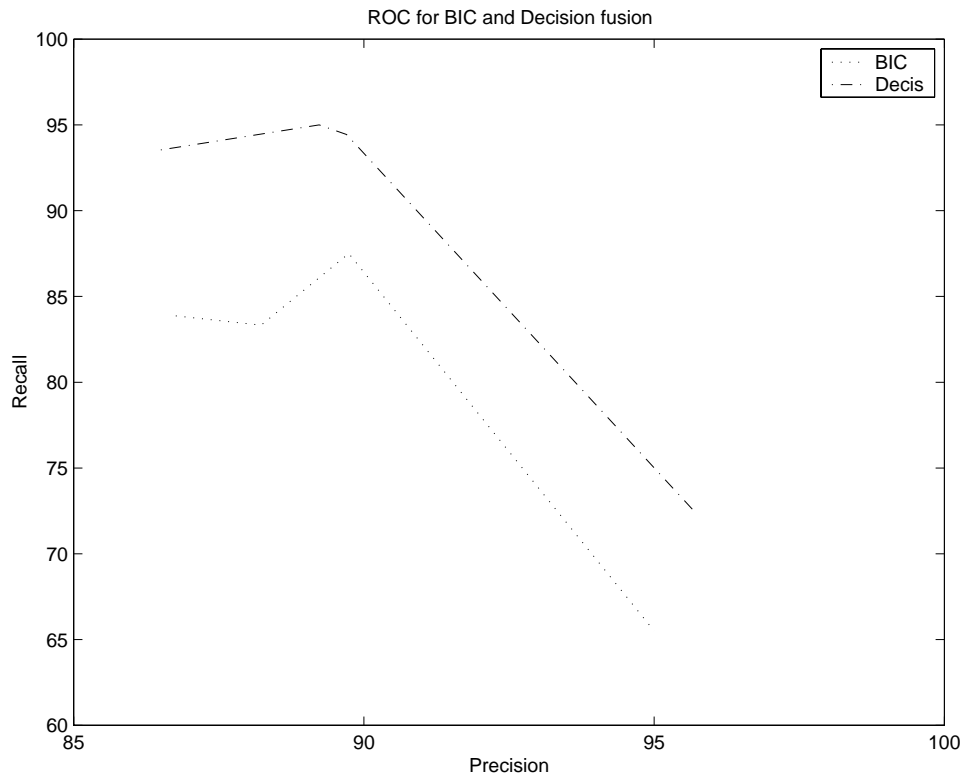


Figure 2: ROC curve for decision fusion and BIC

- [4] U. Gargi, R. Kasturi, and S. Antani. Performance characterization and comparison of video indexing algorithms. In *Intl. Conf. on Computer Vision and Pattern Recognition*. IEEE, Jun. 1998.
- [5] Luo Li-Jun, Zou Cai-Rong, and He Zhen-Ya. A new algorithm on MPEG-2 target bit-number allocation at scene changes. *IEEE Transactions on Circuit and Systems for Video Technology*, 7(5):815–819, October 1997.
- [6] J. Luetttin, N. A. Thacker, and S. W. Beer. Speechreading using shape and intensity information. In *Intl. Conf. on Speech and Lang. Proc.*, pages 58–61. IEEE, 1996.
- [7] U. Meier, W. Hurst, and P. Duchnowski. Adaptive bimodal sensor fusion for automatic speechreading. In *Intl. Conf. On Acoust., Sp., and Sig. Proc.*, pages 833–836. IEEE, 1996.
- [8] Eric Petajan and Hans Peter Graf. Robust face feature analysis for automatic speechreading and character animation. In *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 357–362. IEEE, 1996.
- [9] Behzad Shahraray. Scene change detection and content-based sampling of video sequences. In *Digital Video Compression: Algorithms and Technologies*. SPIE, 1995.
- [10] Behzad Shahraray and David C. Gibbon. Automated authoring of hypermedia documents of video programs. In *Multimedia*. ACM, 1995.
- [11] Chong Song, Li San-qi, and J. Ghosh. Predictive dynamic bandwidth allocation for efficient transport of real-time vbr video over atm. *IEEE Journal on Selected areas in Comm.*, 13(1):12–29, January 1995.

- [12] David G. Stork and Marcus E. Hennecke. Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques. In *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 16–26. IEEE, 1996.
- [13] M. Viswanathan, H. S. M. Beigi, S. Dharanipragada, and A. Tritschler. Retrieval from spoken documents using content and speaker information. In *Intl. Conf. on Document Analysis and Retrieval*, pages 567–572. IEEE, 1999.
- [14] Lynn Wilcox and John S. Boreczky. Annotation and segmentation for multimedia indexing and retrieval. In *Thirty first Hawaii International Conference on Systems Science*, volume 2, pages 259–266. IEEE, 1998.
- [15] B.-L. Yeo and Bede Liu. Rapid scene analysis on compressed video. *IEEE Transactions on Circuit and Systems for Video Technology*, 5, December 1995.
- [16] B.-L. Yeo and M. M. Yeung. Retrieving and visualizing video. *Communications of the ACM*, pages 43–52, December 1997.
- [17] Minerva M. Yeung, Boon-Lock Yeo, and Bede Liu. Extracting story units from long programs for video browsing and navigation. In *International Conf. on Multimedia Computing and Systems*. IEEE, Jun. 1996.