

AUDIO-VISUAL SYNCHRONY FOR DETECTION OF MONOLOGUES IN VIDEO ARCHIVES

G. Iyengar, H. J. Nock, C. Neti

IBM TJ Watson Research Center
Yorktown Heights, NY 10598
USA

ABSTRACT

In this paper we present our approach to detect monologues in video shots. A monologue shot is defined as a shot containing a talking person in the video channel with the corresponding speech in the audio channel. Whilst motivated by the TREC 2002 Video Retrieval Track (VT02), the underlying approach of synchrony between audio and video signals are also applicable for voice and face-based biometrics, assessing of lip-synchronization quality in movie editing, and for speaker localization in video. Our approach is envisioned as a two part scheme. We first detect occurrence of speech and face in a video shot. In shots containing both speech and a face, we distinguish monologue shots as those shots where the speech and facial movements are synchronized. To measure the synchrony between speech and facial movements we use a mutual-information based measure. Experiments with the VT02 corpus indicate that using synchrony, the average precision improves by more than 50% relative compared to using face and speech information alone. Our synchrony based monologue detector submission had the best average precision performance (in VT02) amongst 18 different submissions.

1. INTRODUCTION

This paper is motivated by the TREC 2002 Video Retrieval Track (VT02) problem of monologue detection in digital video archives, defined as the detection of video segments which “contain(s) an event in which a single person is at least partially visible and speaks for a long time without interruption by another speaker” [1]. Our approach to monologue detection involves detection of synchrony between audio and video in addition to detection of a face in the video and speech in the audio track. We hypothesize that using audio-visual synchrony we can disambiguate between instances of narrations where there is unrelated speech in the audio track along with a face in the video and instances of monologues where the face on screen is “responsible” for the speech in the audio track.

We note that synchrony detection has wide applicability. Firstly, detecting synchrony between speech and face can be used to determine dominant speakers in applications such as meeting transcription where we assume that we have access to both audio and visual data. In addition, in such a setting, we can use synchrony to perform speaker localization in a video track [2, 3]. Such localization enables the possibility of using noise-robust audio-visual speech recognition. Secondly, reliable metrics for assessing quality of lip-synchronization would be useful for movie editing and dubbing into multiple languages. Likewise, Speaker voice- and face-based

biometrics systems [4] can benefit from techniques that assess correspondence between speech and facial movements.

Hershey [5] assumes audio and video signals to be individually and jointly Gaussian random variables and estimates the mutual information between them as a measure of synchrony. In [3], we suggest an extension to this approach by relaxing the single Gaussian assumption and allowing the audio and video signals to be locally Gaussian. Fisher et al [6, 7] learn linear projections from audio and visual feature spaces to a joint subspace where the mutual information is maximized. A similar approach is suggested by Slaney [8] using Canonical Correlation Analysis on training data to find a linear projection of audio and video data onto a single axis that maximizes the correlation between the projected variables. We suggest use of empirical distributions for evaluating synchrony between audio and video using vector quantization (VQ) codebooks are used to estimate empirical distributions of audio, video, and joint distributions and their corresponding mutual information[3]. In addition, we also suggest “strong model-based” approach where a word hypothesis is generated by performing automatic speech recognition (ASR) on the audio signal and the likelihood of the joint audio-visual signal for the word hypothesis is evaluated[3]. We note here that this technique not only evaluates synchrony but also “plausibility” (i.e. lip movements that correspond to speech and not just being synchronized with it). Cutler [2] trains a time-delay neural network that captures the relationship between audio and video features for a given speaker. They then use this to locate the speaker in the video using the synchrony between audio and video.

We now outline our approach to monologue detection. Prior to processing, a video sequence is broken into contiguous segments called shots. A shot is defined as a single camera action. This shot change detection is performed automatically. In the case of VT02, NIST provided a standard shot segmentation for the corpus as part of the data. Our algorithm for monologue detection in video shots proceeds in two steps. For each video shot, we perform speech and face detection to evaluate whether the shot contains speech and has a face. For shots containing speech and face, we further evaluate the synchrony between the face and speech using mutual information. The combined scores of speech, face, and synchrony is used to rank all shots in the corpus.

The rest of the paper is organized as follows: In section 2, we detail our approach to speech detection, face detection and synchrony detection between audio and video. In section 3, we detail our monologue detection algorithm. In section 4, we present the results of our monologue detector, evaluated by NIST as part of VT02 benchmarking activity and follow with conclusions.

2. FACE, SPEECH AND SYNCHRONY DETECTION

2.1. Face Detection

The face detector we use is the likelihood ratio between two Gaussian Mixture models (GMM), one trained on frontal faces and one trained on non-face images. Specifically, on a training corpus of frontal faces, we mark facial feature points (eyes, nostrils, mouth etc). Some example annotated faces are shown in Figure 1. From this annotated corpus we extract images that are normalized for orientation and size, resulting in a normalized size of 11×11 pixels. We perform a 2D separable Discrete Cosine Transform (DCT) and retain the top 50 coefficients as the feature representation for these normalized faces. These feature vectors are used to train a 64-mixture, diagonal covariance GMM model representing faces. Similarly, we extract arbitrary 11×11 pixel regions from the same corpus, ensuring that these regions are far away from the annotated face. These non-face regions are used to build a similar 64-mixture, diagonal covariance non-face GMM model.



Fig. 1. Example ground-truth marked-up faces

To search for faces in an image, we extract 11×11 pixel regions from the image at various scales and at all possible translations. For each such extracted region, we perform the likelihood ratio test and retain the likelihood ratio as the score for the region. Equation 1 shows the score for a region. The top-ranked N (where N is the maximum number of faces we want to detect) regions with positive scores are returned as face candidates. For monologue detection, we retain only the top face candidate.

$$S(i) = \frac{P(D(i)|F)}{P(D(i)|\bar{F})} \quad (1)$$

where $S(i)$ is the score for region i and $D(i)$ is the feature vector containing the top-50 DCT coefficients for the region. P is the likelihood of the observation given the models; F and \bar{F} represent the face and non-face GMMs, respectively¹.

From every shot in the VT02 corpus, we extract the frame at the mid-point of the shot description and use this as the key frame for the shot. We perform face detection on this key frame and score each shot in the corpus. Figure 2 shows the results of the face detector on the VT02 corpus. We note here that this face detector has an average precision (AP) of .35². Average precision

¹In the IBM VVAV corpus which is captured under clean conditions [9], we detect faces with an accuracy of 99.8% using the GMM face detector.

²Average Precision is the evaluation metric used by NIST to encode the entire precision recall performance by a single number. It corresponds to the area under the ideal precision recall curve. The metric credits both precision and recall.

is shown in Equation 2

$$AP = \frac{\sum_{i \in RR} p(i)}{N} \quad (2)$$

where RR is the retrieved-relevant set and N is the total number of correct documents in the dataset and $p(i)$ is the precision of the i th retrieved-relevant document. As can be seen from the results, the GMM face detector has high precision at the top 20 shots and detects mostly frontal faces. We note that this detector is not optimized for detecting non-frontal faces.



Fig. 2. The top 20 candidates of the GMM face detector on the VT02 search set

2.2. Speech Detection

We begin with an annotated audio training set where pure audio concepts are manually annotated. By pure concepts we imply instances of audio with only one concept (such as speech, music, silence) in the audio track. Specifically, we manually annotate speech, music, silence, explosion, and traffic sounds in the development set of the VT02 corpus. Regions corresponding to each concept are segmented from the audio and low-level features are extracted. One obvious modeling scheme uses these features to train a GMM for each concept. However, this ignores the duration properties of the audio events; use of these GMMs to label new (or even training) videos (by assigning each frame in the new data to the most likely generating concept) may yield implausibly short events. One scheme for incorporating duration modeling is as follows: An HMM is used to model each audio concept; each state in a given HMM has the same observation distribution, namely the GMM trained in the previous scheme³. This can be viewed as im-

³It is closely related to the speech vs. non-speech segmentation scheme of IBM-Spine2, see Kingsbury et al. [10]

position of a minimum duration constraint on the temporal extent of the atomic labels.

Given a set of HMMs one for each audio concept, during testing (labeling new videos) we use the following scheme to compute the confidences of the different hypotheses. We use the HMMs to generate an N-best list at each audio frame and then average these scores over the duration of the shot. We notice that there are variations in the absolute values of these scores due to variations in the shot lengths and the thresholds chosen for generating the N-best list etc. For example, a lower threshold allows for more hypotheses at any one time but also allows a hypothesis to be valid for a longer duration. To counter these variations, we *normalize* these scores by dividing each concept score with the sum of all the concept scores in a particular shot. The scores are now indicative of the relative strengths of the different hypotheses in a given shot rather than their absolute values.

We use this approach on 5 hours of validation data derived from the 30 hours of VT02 feature development set to evaluate the performance of speech detection. On the validation set, comprising 2295 shots, we detect speech with an average precision (AP) of .99. On the feature test set, comprising 1849 shots we have a similar performance. Of the top 1000 returned shots, we correctly detect 990 of them as containing speech.

2.3. Detection Of Synchrony Between Audio And Video

Based on experimental results of a variety of synchrony detection techniques [3], we choose the scheme that models audio and video features as locally Gaussian distributions. For details of the experiments, please refer [3]. Given a video shot, we extract all the video frames corresponding to it and the associated audio. From the audio signal, we derive MFCC coefficients that are standard in speech recognition systems. The audio features are segmented into locally Gaussian segments using a model selection based segmentation scheme [11]. For each such locally Gaussian audio segment, we evaluate the mutual information between every pixel in the video frames and the audio features. This experiment is related to [5] but differs mainly in the local Gaussian assumption using model selection. On a corpus of 20 speakers, each speaking 10 utterances each, we estimated the mutual information between audio MFCCs and each pixel in the video using local Gaussian distributions. Figure 3 shows the estimated mutual information plots as an image, using (increased) pixel brightness to reflect (increased) mutual information with the audio stream. In all 200 cases, we could clearly localize the face region from the rest of the video. The parts of the face that had high mutual information with the speech are fairly person dependent, but it is promising to see that only the face region has high mutual information with the speech stream. To get a synchrony score from such a mutual informa-



Fig. 3. Mutual Information faces

tion image, we compute the ratio between the mutual information of the face region and the average mutual information across the entire image. Intuitively, the higher this score the greater the synchrony between audio and video. Our experiments indicate that the face location information from the face detector is error-prone and varies considerably across the video. Hence, rather than rely on the face location estimates from the face detector, we compute this ratio between the best $m \times m$ pixel region in the mutual information plot and the background, where m is chosen empirically from the validation set. The search for the best region begins at the top left pixel and proceeds through the entire image in raster order. To speed up the search, we only consider regions whose center pixel is atleast 80% of the maximum mutual information value in the plot.

3. MONOLOGUE DETECTION USING FACE, SPEECH AND SYNCHRONY

We investigated two simple fusion approaches between the three scores. The first approach is a linear weighting of normalized speech, face and synchrony scores. In the second approach, we combine scores using weighted products. The two combination rules are detailed in Equation 3 below. We do a grid search for weights in the range (0, 1) and we variance normalize the scores prior to the grid search for weights. The learned variance normalization parameters and weights are applied to unseen test data prior to testing.

$$\begin{aligned} M_+(i) &= w_1 * F(i) + w_2 * Sp(i) + w_3 * Sy(i) \\ M_x(i) &= F(i)^{w_1} * Sp(i)^{w_2} * Sy(i)^{w_3} \end{aligned} \quad (3)$$

where $M_+(i)$, $M_x(i)$, $F(i)$, $Sp(i)$ and $Sy(i)$ are the monologue score using weighted sum, monologue score using weighted product, face, speech and synchrony scores for shot i , respectively. We note here that these approaches can be thought of as Naive Bayes classifiers because of the implicit independence assumption.

4. RESULTS

The number of monologue shots are extremely limited in both the validation and test corpora. Based on NIST ground truth, of the 2295 validation shots, only 33 (1.4%) shots are classified as monologues. Similarly, of the 1849 shots in the test corpus, only 38 (2.1%) shots are monologues. In contrast, more than 75% of the shots contain speech and more than 27% of the shots contain faces. Monologue is an extremely rare class in this dataset and therefore makes it a very interesting detection problem.

In Figure 4, we present the average precision performance for the monologue detector, both on the validation set and on the test set. The first two columns represent the performance of the two different fusion rules on the validation set. We see that the weighted sum approach is slightly better than product weighting and chose this for our VT02 submission. The next two columns show the average precision numbers released by NIST for the average (across 18 different monologue detector submissions), and our submitted detector (the best in VT02).

To quantify the effect of synchrony on monologue detection, in Table 1 we compare the performance of using all 3 scores with using only face and speech information alone. The first column indicates the approach used. The second column shows the AP and the third column shows the Recall at 1000 documents. It is clear

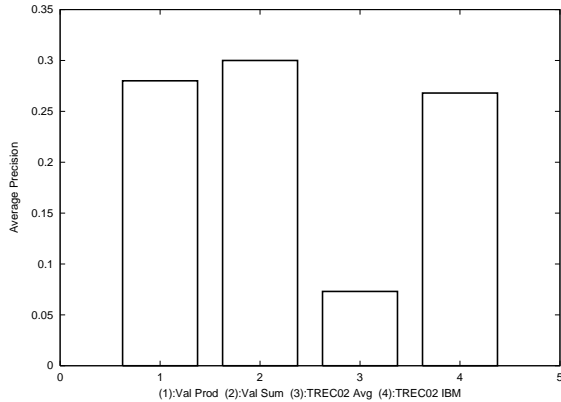


Fig. 4. Average precision plots for the monologue detector for the VT02 corpus. The first two columns represent the performance of the two fusion schemes on the validation set. The next column illustrates the average performance across all submitted detectors and the last column represents our submission to VT02.

that using synchrony in addition to face and speech information adds significant performance improvement.

Technique	AP	Recall@1000
face+speech	.19	.84
with Sync	.30	.88

Table 1. Comparison between synchrony and face+speech for monologue detection

5. SUMMARY AND DISCUSSION

In this paper we proposed the use of synchrony between audio and video features as a strong cue to detect monologues in video archives. This is used in concert with speech and face detection to build a model for monologues. We applied this monologue detector to the VT02 corpus as part of the benchmarking activity. Based on these results, it appears that the synchrony based approach is an extremely promising approach for detecting monologues in video sequences. We note that, as currently formulated, our scheme for monologue detection cannot differentiate between monologues and dialogue scenes where the faces of both parties are visible to the camera. However, this approach in combination with a speaker change detection scheme (cf. [11]) is a possibility for such instances.

We did not make use of the face location information provided by the face detector. This is mainly because of the erroneous estimates of the scale, orientation, and position of the face. One possibility is to use the synchrony as a mechanism to guide the face detection to get better localization of the face in an image. This approach is promising for talking head detection under varying lighting and visual noise conditions. With better face localization, rather than searching the entire image for synchrony, only the localized region could be searched, thereby enabling an efficient mechanism for synchrony detection. An iterative approach where

face localization and synchrony detection can be refined in steps is an interesting pursuit to enable fast localization and detection of talking heads in applications such as meeting transcription, smart environments, biometric authentication, and pervasive computing.

6. ACKNOWLEDGMENTS

We would like to thank G. Saon of IBM Research for the Viterbi decoder, and J. R. Smith of IBM Research for providing the video key frames, shot descriptions and the retrieval infrastructure.

7. REFERENCES

- [1] “Text retrieval conference (trec) video track,” <http://trec.nist.gov>.
- [2] Ross Cutler and Larry Davis, “Look Who’s Talking: Speaker Detection using Video and Audio Correlation,” in *Proc. ICME*, 2000.
- [3] Harriet J. Nock, Giridharan Iyengar, and Chalapathy Neti, “Assessing face and speech consistency for monologue detection in video,” in *Proc. ACM Multimedia*, 2002.
- [4] Benoit Maison, Chalapathy Neti, and Andrew Senior, “Audio-visual speaker recognition for video broadcast news: some fusion techniques,” in *IEEE Multimedia Signal Processing (MMSP99)*, Denmark, September 1999.
- [5] John Hershey and Javier Movellan, “Using audio-visual synchrony to locate sounds,” in *Proc. NIPS*, 1999.
- [6] JW Fisher III, T Darrell, WT Freeman, and P Viola, “Learning Joint Statistical Models for Audio-Visual Fusion and Segregation,” in *Proc. NIPS*, 2001.
- [7] John W Fisher III and Trevor Darrell, “Informative subspaces for audiovisual processing: High-level function from low-level fusion,” in *Proc. ICASSP*, 2002.
- [8] Malcolm Slaney and Michele Covell, “Facesync: a linear operator for measuring synchronization of video facial images and audio tracks,” in *Proc. NIPS*, 2001.
- [9] C. Neti, G. Potamianos, J. Leuttin, I. Matthews, H. Glotin, D. Vergyri, J. Sisson, A. Mashari, and J. Zhou, “Audio-visual speech recognition,” CLSP Summer Workshop Tech. Rep. WS00AVSR, Johns-Hopkins University, Baltimore, MD, 2000.
- [10] Brian Kingsbury, George Saon, Lidia Mangu, Mukund Padmanabhan, and Ruhi Sarikaya, “Robust Speech Recognition in Noisy Environments: The IBM Spine-2 Evaluation System,” in *Proc. ICASSP*, 2002.
- [11] 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.