A Framework for Dialogue Detection in Movies

Margarita Kotti, Constantine Kotropoulos, Bartosz Ziólko, Ioannis Pitas, and Vassiliki Moschou*

Department of Informatics, Aristotle University of Thessaloniki
Box 452, Thessaloniki 54124, Greece
{mkotti, costas, pitas, vmoshou}@aiia.csd.auth.gr

Abstract. In this paper, we investigate a novel framework for dialogue detection that is based on indicator functions. An indicator function defines that a particular actor is present at each time instant. Two dialogue detection rules are developed and assessed. The first rule relies on the value of the cross-correlation function at zero time lag that is compared to a threshold. The second rule is based on the cross-power in a particular frequency band that is also compared to a threshold. Experiments are carried out in order to validate the feasibility of the aforementioned dialogue detection rules by using ground-truth indicator functions determined by human observers from six different movies. A total of 25 dialogue scenes and another 8 non-dialogue scenes are employed. The probabilities of false alarm and detection are estimated by cross-validation, where 70% of the available scenes are used to learn the thresholds employed in the dialogue detection rules and the remaining 30% of the scenes are used for testing. An almost perfect dialogue detection is reported for every distinct threshold.

1 Introduction

Digital movie archives have become a commonplace nowadays. Research on movie content analysis has been very active. A dialogue scene can be defined as a set of consecutive shots which contain conversations of people [1]. However, there is a possibility of having shots in a dialogue scene that do not contain any conversation or even any person. The elements of a dialogue scene are: the people, the conversation and the location is taking place in [2]. The basic shots in a dialogue scene are: (i) Type A shot: Shot of actor A's face; (ii) Type B shot: Shot of actor B's face; (iii) Type C shot: Shot with both faces visible. A set of recognizable dialogue acts according to semantic content is proposed in [3]: (i) Statements; (ii) Questions; (iii) Backchannels; (iv) Incomplete utterance; (v) Agreements; Appreciations. Dialogue detection in movies follows specific rules since movie making is a kind of art [5]. Lehane states that in a 2-person dialogue there is usually a A-B-A-B structure of camera angles, thus making dialogue detection feasible [4]. However, the person who speaks at any given time is not always the one displayed. Shots of other participants' reactions are frequently inserted. In addition, the shot of the speaker may not include his face, i.e. the rear view of his head might be

* This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation and LEarning" (FP6-507752).

B. Gunsel et al. (Eds.): MRCS 2006, LNCS 4105, pp. 371–378, 2006.

depicted. Furthermore, shots of other persons or objects might be inserted in the dialogue scene. Evidently, these shots add to the complexity of the dialogue detection problem, due to their nondeterministic nature. Numerous methods for dialogue detection have been proposed, because such a preprocessing step is useful for video analysis, indexing, browsing, searching, and summarization. Both video and audio information channels could be exploited for efficient dialogue detection. For example, automatically extracted low-level and mid-level visual features are used to detect different types of scenes, focusing on dialogue sequences [4]. Emotional stages as a means for segmenting video are proposed in [6]. The detection of monologues based on audio-visual information is discussed in [7] where a noticeably high average decision performance is reported. Related topics to dialogue detection are face detection and tracking [8], speaker turn detection [9], and speaker tracking [10]. The aforementioned research is compliant with the MPEG-7 standard.

In this paper, we propose a novel framework for dialogue detection that is based on *indicator functions*. In practice, indicator functions can be obtained by speaker turn detection followed by speaker clustering or by face detection followed by a similar clustering procedure. However, in this paper we are interested in setting up the detection framework in the *ideal* situation where the indicator functions are *error free*. Towards this goal ground truth indicator functions are employed. Two dialogue detection rules are developed. The first rule employs the value of the cross-correlation function at zero time-lag and the second one is based on the cross-power in specific frequency band. Both quantities are compared to corresponding thresholds. Experiments are carried out using the audio streams extracted from six different movies while the ground-truth indicator functions are defined by human observers. To validate the feasibility of the dialogue detection rules, the cross-validation approach is utilized, where 70% of the audio streams is used to define the two thresholds, and the remaining 30% is used for testing. Experimental results indicate that an almost perfect dialogue detection is achievable.

The outline of the paper is as follows. The proposed dialogue detection rules are discussed in Section 2. In Section 3, the dialogue scenes used for the experimental evaluation of the proposed method and the training procedure are described. In Section 4, performance evaluation is presented and finally conclusions are drawn in Section 5.

2 Dialogue Detection

2.1 Indicator Functions

Indicator functions are frequently used in statistical signal processing. They are closely related to zero-one random variables used in the derivation of the probabilities of events through expected values [11]. In maximum entropy probability estimation, indicator functions are used to insert constrains quantifying facts stemming from the training data that constitute our knowledge about the random experiment. An example is language modeling [12]. Indicator functions have also been used in the analysis of the DNA sequences [13].

Let us suppose that we have an audio recording of N samples, where N is the product of duration of the audio recording multiplied by the sampling frequency and we

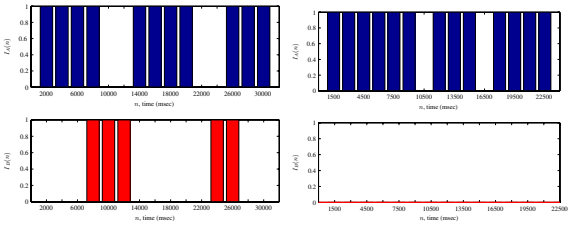
know exactly when a particular actor (i.e. speaker) appears. Such information can be quantified by the indicator function of say actor A, $I_A(n)$, defined as:

$$I_A(n) = \begin{cases} 1 & \text{when actor } A \text{ is present at sample } n \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For a dialogue, at least two actors should be present. Let us call them A and B with corresponding indicator functions $I_A(n)$ and $I_B(n)$, respectively. Besides their presence, the actors should be active, that is their indicator functions should not be zero during the entire scene duration. To avoid such irregularities, we can measure a proper norm of the indicator function, e.g. the L_1 norm or the L_2 norm, etc. Since the indicator functions admit non-negative values, their L_1 norm is simply the sum of the indicator function values:

$$S_A = \sum_{n=1}^{N} I_A(n). \tag{2}$$

Two characteristic indicator functions for a dialogue scene are plotted in Figure 1(a). There are several possibilities for a dialogue scene. For example, there might be audio



(a) (b)

Fig. 1. (a) Indicator functions of two actors in a dialogue scene. (b) Indicator functions of two actors in a non-dialogue scene (i.e. monologue).

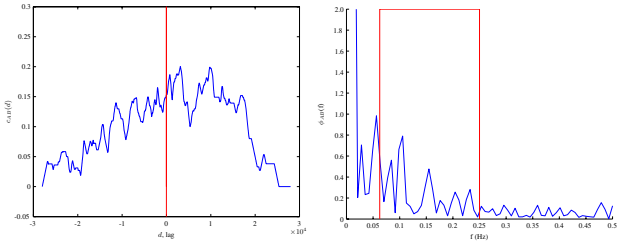
frames where both actors speak. Audio frames corresponding to short silences should be tolerated. In addition, the audio background in dialogue scenes might contain music or environment noise that should not prevent dialogue detection. For the time-being, since optimal (i.e. ground-truth) indicator functions are employed, such cases are not dealt with explicitly. An example of a scene where there is no dialogue is shown in Figure 1(b). It is seen that $I_B(n)$ is zero for all n. This is the case of an inactive actor for whom $S_B = 0$.

2.2 Cross-Correlation

The cross-correlation is a measure of similarity between two signals. It is defined as:

$$c_{AB}(d) = \begin{cases} \frac{1}{N} \sum_{n=1}^{N-d} I_A(n+d) I_B(n) & \text{if} \quad 0 \le d \le N-1 \\ c_{BA}(-d) & \text{if} \quad -(N-1) \le d \le 0 \end{cases} \tag{3}$$

where N is the total number of samples in the audio stream and d is the time-lag. For $d = 0$, the cross-correlation is equal to the product of the two indicator functions $I_A(n)$ and $I_B(n)$. Practically, this means that the greater the value of $c_{AB}(0)$ is, the longer time the two actors speak simultaneously. The cross-correlation for the dialogue shown in Figure 1(a) is depicted in Figure 2(a). It can be seen that $c_{AB}(0) > 0$.



(a) (b)

Fig. 2. (a) Cross-correlation of the indicator functions for two actors participating in a dialogue. (b) Cross-power spectrum for two actors participating in a dialogue.

For the scene corresponding to the two indicator functions plotted in Figure 1(b), the cross-correlation is zero throughout its domain. From the aforementioned observations, a plausible dialogue detection rule is:

$$c_{AB}(0) \geq \vartheta_1 \tag{4}$$

where ϑ_1 is an appropriately chosen threshold.

2.3 Cross-Power Spectrum

Another useful notion to be exploited for dialogue detection is the cross-power spectrum, i.e., the discrete-time Fourier transform of the cross-correlation:

$$\phi_{AB}(f) = \sum_{d=-(N-1)}^{N-1} c_{AB}(d)\ \exp\left(-j2\pi\ f\ d\right) \tag{5}$$

where $f \in [-0.5, 0.5]$ is the frequency in cycles per sampling interval. In order to robustify the dialogue detection, we propose to examine the cross-power p in the frequency band $[0.065, 0.25]$ that has been determined by analyzing the measured cross-power spectra

$$p = \int_{0.065}^{0.25} |\phi_{AB}(f)|^2\ df. \tag{6}$$

When there is a dialogue, p admits a value that depends on the area under the cross-power spectrum $\phi_{AB}(f)$. Figure 2(b) shows the cross-power spectrum density over the frequencies $[0, 0.5]$. For negative frequencies, $\phi_{AB}(-f) = \phi^*_{AB}(f)$. On the other

hand, in the non-dialogue scene corresponding to the two indicator functions plotted in Figure 1(b), the cross-power spectrum is identically zero. Accordingly, the second dialogue detection rule proposed is:

$$p \geq \vartheta_2 \tag{7}$$

where ϑ_2 is clearly an appropriately chosen threshold.

3 Data Set and Training Procedure

In total, 33 recordings were extracted from the following six movies: "Analyze That", "Cold Mountain", "Jackie Brown", "Lord of the Rings I", "Platoon", and "Secret Window". The total duration of the 33 recordings is 31 min and 7 sec. The audio track was digitized in PCM at a sampling rate of 48 kHz and the quantized sample length was 16 bit two-channel. 25 out of the 33 recordings correspond to dialogue scenes, while the remaining 8 do not contain any dialogue. For each recording, the ground-truth indicator function of the actors appearing in the scene is determined and for each pair of indicator functions their cross-correlation sequence is calculated.

In order to check the efficiency of the proposed detection rules (4) and (7), we need to estimate for each rule the probability of detection and the probability of false alarm. The aforementioned probabilities stem from the binary hypothesis detection problem where the null hypothesis is H_0: the scene is not a dialogue and the alternative hypothesis H_1: the scene is a dialogue. Then, the probability of detection is for rule (4):

$$P_d^{(1)} = \text{Prob}(\text{rule (4) decides the scene is dialogue}|H_1) \tag{8}$$

and the probability of false alarm is given by:

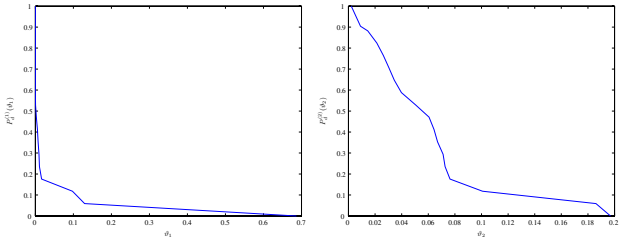
$$P_f^{(1)} = \text{Prob}(\text{rule (4) decides the scene is dialogue}|H_0). \tag{9}$$

$P_d^{(2)}$ and $P_f^{(2)}$ are defined similarly for rule (7). To estimate $P_d^{(i)}$ and $P_f^{(i)}$, $i = 1, 2$, cross-validation is employed.

The available cross-correlation sequences and their cross-power spectrum densities are divided into two disjoint subsets. The first subset is used for training and the second subset is used for testing. 70% of the available data are used for training and the remaining 30% for testing. This means that the 23 randomly selected cross-correlation sequences and their corresponding cross-power spectrum densities are used for training and the remaining 9 are used for testing. When selecting the 23 training sequences we simultaneously preserved the ratio between dialogue and non dialogue scenes, i.e. 18 cross-correlation sequences corresponding to dialogue scenes and another 6 corresponding to non dialogue scenes. Similarly, the testing cross-correlation sequences were formed by 7 audio streams corresponding to dialogue scenes and another 2 corresponding to non dialogue scenes.

Because of the relatively small amount of the training sequences we applied the leave-one-out method to estimate the probability of detection. That is 22 out of the 23 training sequences are used to estimate the probability of detection and the estimation

is repeated by leaving a different training sequence out for all training sequences (i.e. 23 times). Let $P_d^{(i;r)}(\vartheta_i^r)$ be the probability of detection for the ith rule that employs the threshold ϑ_i^r when the rth training sequence is left out. Figure 3(a) shows the average $P_d^{(1)}(\vartheta_1)$ versus ϑ_1. The curve is estimated by averaging the probabilities measured in the 23 repetitions. The corresponding plot of the average $P_d^{(2)}(\vartheta_2)$ versus ϑ_2 is depicted in Figure 3(b).



(a) (b)

Fig. 3. (a) The average $P_d^{(1)}(\vartheta_1)$ versus ϑ_1 for the first rule. (b) The average $P_d^{(2)}(\vartheta_2)$ versus ϑ_2 for the second rule.

Let ϑ_i be chosen as the minimum threshold value such that $P_d^{(i;r)}(\vartheta_i^r)$=1. Table 1a summarizes the thresholds determined for each training sequence left out. By applying the minimum threshold value and using the entries of Table 1a, we find that $\vartheta_1 = 3.52 \times 10^{-18}$ and $\vartheta_2 = 0.004$, respectively.

4 Performance Evaluation During Testing

For the 9 audio streams left out for testing, their corresponding cross-correlations and cross-power spectrum densities are computed and the values of $c_{AB}(0)$ and p are collected in Table 1b. The first seven rows in Table 1b correspond to dialogue scenes and the last two correspond to non-dialogues. From the inspection of Table 1b, it is seen that only the 6th cross-correlation sequence is not detected as corresponding to a dialogue scene by applying the detection rule (4), although it is. It is also seen that there are no false alarms. The second detection rule (7) can rectify the just described miss-detection. A simple OR rule, i.e.

$$c_{AB}(0) \geq \vartheta_1 \quad OR \quad p \geq \vartheta_2. \tag{10}$$

can yield a perfect dialogue detection.

To compensate the lack of real indicator functions, a number of synthetic indicator functions admitting real values within $[0, 1]$ have been created and including in the test phase. The nature of syntectic indicator functions created and the performance of rule (10) is summarized in Table 2.

Table 1. (a) The 23 pairs of ϑ_1 and ϑ_2 during the training procedure. (b) The 9 pairs of cross-correlation value at zero lag and cross-power in the frequency band $f \in [0.065, 0.25]$ for the test recordings.

sequence left out, r	ϑ_1^r	ϑ_2^r
1	3.52×10^{-18}	0.010
2	3.52×10^{-18}	0.010
3	3.53×10^{-18}	0.010
4	3.52×10^{-18}	0.0082
5	3.53×10^{-18}	0.010
6	3.53×10^{-18}	0.010
7	3.52×10^{-18}	0.0082
8	3.52×10^{-18}	0.0082
9	3.52×10^{-18}	0.0082
10	3.52×10^{-18}	0.0082
11	3.52×10^{-18}	0.0082
12	3.52×10^{-18}	0.0082
13	3.53×10^{-18}	0.010
14	3.52×10^{-18}	0.0082
15	3.52×10^{-18}	0.0082
16	3.53×10^{-18}	0.010
17	3.52×10^{-18}	0.0082
18	3.53×10^{-18}	0.010
19	3.52×10^{-18}	0.0082
20	3.53×10^{-18}	0.0082
21	3.52×10^{-18}	0.004
22	3.52×10^{-18}	0.004
23	3.52×10^{-18}	0.004

test audio stream index	$c_{AB}(0)$	p
1	1.61×10^{-5}	0.0254
2	0.0176	0.0859
3	0.0854	0.0854
4	1.42×10^{-17}	0.0307
5	0.0018	0.0529
6	$\mathbf{1.73} \times 10^{-18}$	0.0999
7	0.0043	0.0859
8	0	0
9	0	0

Table 2. Synthetic indicator functions, their corresponding $c_{AB}(0)$ and p values, and final decision

Nature of the indicator function	$c_{AB}(0)$	p	Dialogue detection
Adding Gaussian noise $\sim \mathcal{N}(0.2, 0.05)$ independently to both indicator functions and hard limiting to $[0, 1]$.	0.1899	0.1132	correct
Adding a considerable amount of silence between speaker turn points (here 33.3% of the average dialogue duration is silence).	1.3817×10^{-18}	0.0191	correct
Adding a considerable amount of overlap between speaker activities (the overlap amounts to 33.3% of the average dialogue duration).	0.0761	0.3371	correct
Modeling between-speaker silence as a Gaussian random variable $\sim \mathcal{N}(0.5, 0.05)$	0.1053	3.6405×10^{-17}	correct
Modeling between-speaker silence as a uniform random variable	0.3654	2.6820×10^{-17}	correct
Modeling between-speaker silence/music/noise as constant value of 0.2.	3.8892×10^{-5}	0.0239	correct

5 Conclusions

In this paper, we have proposed a novel framework for dialogue detection in movies based on indicator functions. Experiments are carried out using indicator function ground truth extracted from real movies. Cross-validation was used to estimate the

probabilities of detection and false alarm. The experimental results demonstrate the feasibility of the proposed detection rules in 33 movie segments. In the future, we plan to extend our movie database. Moreover, the ground truth indicator functions will be replaced by actual ones derived by either speaker turn detection followed by speaker tracking or face detection followed by face tracking by their combination.

References

1. A. A. Alatan and A. N. Akansu, "Multi-modal dialog scene detection using hidden-markov models for content-based multimedia indexing," *J. Multimedia Tools and Applications*, vol. 14, pp. 137-151, 2001.
2. L. Chen and M. T. Özsu, "Rule-based extraction from video," in Proc. *2002 IEEE Int. Conf. Image Processing*, vol. II, pp. 737-740, 2002.
3. P. Král, C. Cerisara, and J. Kleckova, "Combination of classifiers for automatic recognition of dialogue acts," in Proc. *9th European Conf. Speech Communication and Technology*, pp. 825-828, 2005.
4. B. Lehane, N. O'Connor, and N. Murphy, "Dialogue scene detection in movies using low and mid-level visual features," in Proc. *Int. Conf. Image and Video Retrieval*, pp. 286-296, 2005.
5. D. Arijon, *Grammar of the Film Language*. Silman-James Press, 1991.
6. A. Vassiliou, A. Salway, and D.Pitt, "Formalising stories: sequences of events and state changes", in Proc. *2004 IEEE Int. Conf. Multimedia and Expo*, vol. I, pp. 587-590, Hong-Kong, Taiwan 2004.
7. G. Iyengal, H. J. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in Proc. *2003 IEEE lnt. Conf. Acoustics, Speech, and Signal Processing*, vol. I, pp. 329-332, April 2003, Hong Kong.
8. K. Sobottka and I.Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *Image Communication and Signal Processing*, vol. 12, no. 3, pp. 263-281, June 1998.
9. M. Kotti, E. Benetos, and C. Kotropoulos, "Automatic speaker change detection with the bayesian information criterion using MPEG-7 features and a fusion scheme," in Proc. *2006 IEEE Int. Symp. Circuits and Systems*, May 2006, Kos, Greece.
10. L. Lu and H. Zhang, "Speaker change detection and tracking in real-time news broadcast analysis," in Proc. *2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. I, pp. 741-744, June 2004.
11. A. Papoulis and S. V. Pillai, *Probabilities, Random Variables, and Stochastic Processes*, 4/e. N.Y.: McGraw-Hill, 2002.
12. F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, Massachusetts: The MIT Press, 1997.
13. R. J. Boys and D. A. Henderson, "A Bayesian approach to DNA sequence segmetation", in Proc. *2004 Biometrics*, vol. 60, no 3, pp. 573, September 2004.