City Research Online

City St George's, University of London

Citation: Kotti, M, Benetos, E., Kotropoulos, C. & Pitas, I (2007). A neural network approach to audio-assisted movie dialogue detection. Neurocomputing, 71(1-3), pp. 157-166. doi: 10.1016/j.neucom.2007.08.006

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/2046/

Link to published version: https://doi.org/10.1016/j.neucom.2007.08.006

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

NOTICE: this is the author's version of a work that was accepted for publication in Neurocomputing. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Neurocomputing, 71(1-3):157-166, Dec. 2007. doi: 10.1016/j.neucom.2007.08.006

A neural network approach to audio-assisted movie dialogue detection

Margarita Kotti, Emmanouil Benetos, Constantine Kotropoulos *, Ioannis Pitas

Artificial Intelligence and Information Analysis Lab, Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 54124, Greece

Abstract

A novel framework for audio-assisted dialogue detection based on indicator functions and neural networks is investigated. An indicator function defines that an actor is present at a particular time instant. The cross-correlation function of a pair of indicator functions and the magnitude of the corresponding cross-power spectral density are fed as input to neural networks for dialogue detection. Several types of artificial neural networks, including multilayer perceptrons, voted perceptrons, radial basis function networks, support vector machines, and particle swarm optimization-based multilayer perceptrons are tested. Experiments are carried out to validate the feasibility of the aforementioned approach by using ground-truth indicator functions determined by human observers on 6 different movies. A total of 41 dialogue instances and another 20 non-dialogue instances is employed. The average detection accuracy achieved is high, ranging between 84.78%±5.499% and 91.43%±4.239%.

Key words: Dialogue detection, Indicator functions, Cross-correlation, Cross-power spectral density.

1. Introduction

Nowadays digital archives is a commonplace. Movies alone constitute a large portion of the entertainment industry, as over 9.000 hours of video are released every year [1]. As the available bandwidth per user increases, online movie stores, the equivalent of digital music stores that prevail in music industry, emerge. There are various reasons explaining this phenomenon, such as the wide prevalence of personal computers, the decreasing cost of mass storage devices, and the advances in compression.

For efficient handling of digital movie archives, multimedia data management should be employed. As a consequence, research on movie content analysis has been very active. Browsing and retrieval are required to annotate and appropriately organize the data. To manage specific classes of movie content, the detection of dialogue scenes is essential. For example, a rough idea about the movie genre (i.e. drama, comedy, action) can be provided by a quantitative comparison between the duration of dialogue scenes and that of non-dialogue scenes. A second application derives from the fact that dialogue scenes follow some specific patterns that facilitate their detection in a video sequence. Hence, dialogue scene analysis is a key concept for movie content management.

The need for content-based audiovisual analysis has been noted by the MPEG committee, leading to the creation of the MPEG-7 standard (formerly known as Multimedia Content Description Interface) [27]. Current approaches to automatic movie analysis and annotation focus mainly on visual information, while audio information receives little or no attention. However, as it is proposed in this paper, significant information content exists in the audio channel, as well. It should be noted that, combined audio and video information yield a better analysis of the semantic movie content than processing separately the audio and video information channels. For example, dialogue detection experiments have been performed using low-level audio and visual features with a maximum classification accuracy of 96% [1]. Alternatively, emotional stages are proposed as a means for segmenting video in [24]. Detecting monologues based on audio-visual information is discussed in [11], where a maximum recall of 0.880 is reported. Related topics to dialogue detection are face detection and tracking, speaker turn detection [13], and speaker tracking [17].

Numerous definitions of a dialogue appear in the literature. According to Alatan, a dialogue scene can be defined as a set of consecutive shots, which contain conversations of people [2]. However, in the extreme case, the shots in a dialogue scene may not contain any conversation or even any person. According to Chen, the elements of a dialogue scene are: the people, the conversation, and the location where the dialogue is taking place [6]. 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

* Corresponding author.

Email addresses: mkotti@aiia.csd.auth.gr (Margarita Kotti), empeneto@aiia.csd.auth.gr (Emmanouil Benetos), costas@aiia.csd.auth.gr (Constantine Kotropoulos), pitas@aiia.csd.auth.gr (Ioannis Pitas).

similar classification of audio recordings can be made with respect to speakers' utterances, as well. Recognizable dialogue acts, according to semantic content, are [15]: (i) Statements, (ii) Questions, (iii) Backchannels, (iv) Incomplete utterances, (v) Agreements, (vi) Appreciations. Movie dialogue detection follows specific rules, since movie making is a kind of art and it has its own grammar [3]. Lehane states that dialogue detection is feasible, since there is usually an A-B-A-B structure of camera angles in a 2-person dialogue [16]. However, this is not the only case, since the person who speaks at any given time is not always the one displayed. For example, shots of other participants' reactions are frequently inserted. In addition, the shot of the speaker may not include his face, but the back view of his head. Various shots may be inserted in the dialogue scene, such as other persons or objects. Evidently, these shots add to the complexity of the dialogue detection problem, due to their nondeterministic nature.

In this paper, a novel framework for *audio-assisted* dialogue detection based on *indicator functions* is proposed. In practice, indicator functions can be obtained by speaker turn detection followed by speaker clustering. However, in this work we are interested in setting up the detection framework in the *ideal* situation, where the indicator functions are *error-free*. To achieve this, human observers extracted the ground-truth indicator functions. The cross-correlation values of a pair of indicator functions and the magnitude of the corresponding cross-power spectral density are fed as input to neural networks for dialogue detection. Experiments are carried out using the audio scenes extracted from 6 different movies enlisted in Table 1. In total, 27 dialogue and 12 non-dialogue scenes are employed where dialogues last from 20 sec to 90 sec. The aforementioned dialogue scenes are used to model the empirical distribution of actor utterance duration. After examining several distributions, it is found that the Inverse Gaussian fits best the empirical distribution of actor utterance duration. The expected value of the actor utterance duration is found to be 5 sec. A time window of 25 sec is employed, to ensure that 4 actor changes are taken into account on average. A total of 41 dialogue instances and another 20 non-dialogue instances are extracted. Several types of neural networks are tested for dialogue detection, namely multilayer perceptrons, voted perceptrons, radial basis function networks, and support vector machines. Meta-classifiers like AdaBoost and MultiBoost are also employed, in order to enhance the performance of the aforementioned artificial neural networks. It is demonstrated that, a high dialogue detection accuracy is achieved, ranging between 84.78%±5.499% and 91.43%±4.239% with a mean F_1 measure of 0.906.

The remainder of the paper is as follows. In Section 2, the notion of indicator function is introduced in the framework of dialogue detection. In addition, the cross-correlation and the cross-power spectral density are described as features for dialogue detection. The dataset created, as well as its modeling is discussed in Section 3. Figures of merit are defined in Section 4. In Section 5, experimental results are described. Finally, conclusions are drawn in Section 6.

2. Audio Dialogue Detection

2.1. *Indicator functions*

Indicator functions are frequently employed in statistical signal processing. They are closely related to zero-one random variables, used in the computation of expected values, in order to derive the probabilities of events [19]. In maximum entropy probability estimation, indicator functions are used to insert constraints. The aforementioned constraints, quantify facts stemming from training data that constitute the knowledge about the experiment. An example of indicator function usage, is language modeling [12]. The analysis of the DNA sequences utilizes indicator functions as well [5].

Let us suppose that we know exactly when a particular actor (i.e. speaker) appears in an audio recording of N samples, where N is the product of the audio recording duration multiplied by the sampling frequency. Such information can be quantified by the indicator function of say actor A, $I_A(n)$, defined as:

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

At least two actors should be present in a dialogue. We shall confine ourselves to 2-person dialogues, without loss of generality. If the first actor is denoted by A and the second by B, their corresponding indicator functions are $I_A(n)$ and $I_B(n)$, respectively. For a dialogue scene, a characteristic plot of indicator functions can be seen in Figure 1.

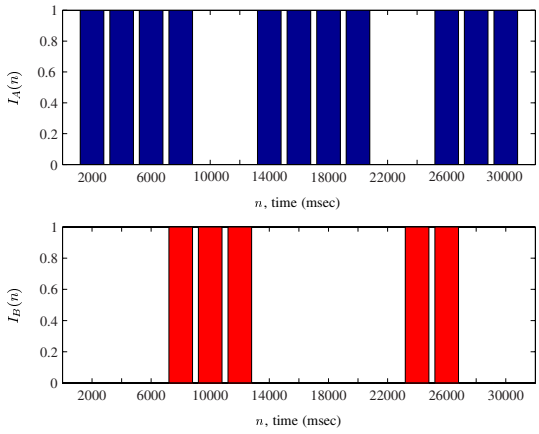

Fig. 1. Indicator functions of two actors in a dialogue scene.

There are several alternatives to describe a dialogue scene. In every-day 2-actor dialogues, the first actor rarely stops at n and the second actor starts at $n+1$. There might be audio frames corresponding to both actors. In addition, short periods of silence should be tolerated. Frequently, the audio background might contain music or environmental noise that should not prevent dialogue detection. In this paper, optimal, error-free (i.e. ground-truth) indicator functions are employed. In Figure 2, a typical example of a non-dialogue (i.e. a monologue)

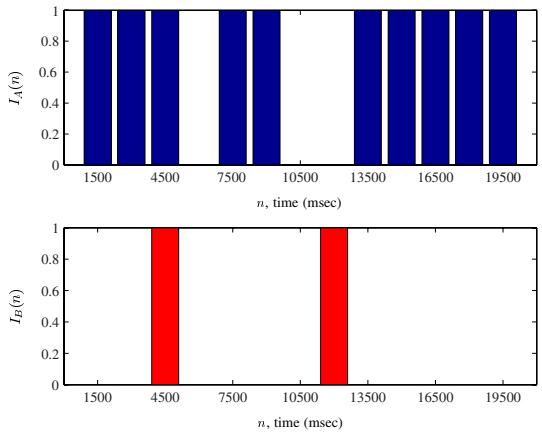

Fig. 2. Indicator functions of two actors in a non-dialogue scene (i.e. monologue).

is depicted, where $I_B(n)$ corresponds to short exclamations of the second actor.

2.2. *Cross-correlation and cross-power spectral density*

A common measure of similarity between two signals is their cross-correlation [22]. The cross-correlation is commonly used to find a linear relationship between the two signals. The cross-correlation of a pair of indicator functions is defined by:

$$c_{AB}(d) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{n=1}^{N-d} I_A(n+d)I_B(n), \\ \qquad \text{when } 0 \le d \le N-1 \\ c_{BA}(-d), \\ \qquad \text{when } -(N-1) \le d \le 0 \end{cases} \tag{2}$$

where d is the time-lag. Significantly large values of the cross-correlation function, indicate the presence of a dialogue. For a dialogue scene, a typical cross-correlation function is depicted in Figure 3.

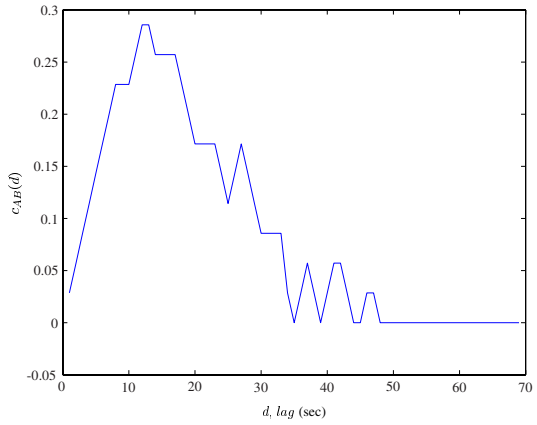

Fig. 3. Cross-correlation of the indicator functions of two actors in a dialogue.

Another useful notion to be exploited for dialogue detection is the discrete-time Fourier transform of the cross-correlation,

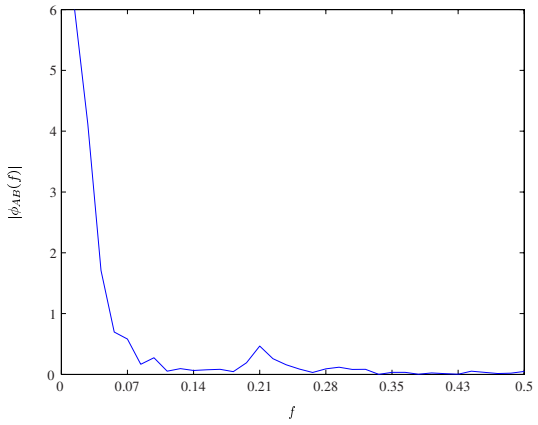

Fig. 4. Magnitude of the cross-power spectral density for two actors in a dialogue.

i.e. the cross-power spectral density [22]. The cross-power spectral density is defined as:

$$\phi_{AB}(f) = \sum_{d=-(N-1)}^{N-1} c_{AB}(d)\ \exp\left(-j2\pi\, f\, d\right) \tag{3}$$

where $f \in [-0.5, 0.5]$ is the frequency in cycles per sampling interval. For negative frequencies, $\phi_{AB}(-f) = \phi^*_{AB}(f)$, where * denotes complex conjugation. In audio processing experiments, the magnitude of the cross-power spectral density is commonly employed. When there is a dialogue, the area under $|\phi_{AB}(f)|$ is considerably large, whereas it admits a rather small value for a non-dialogue. Figure 4 shows the magnitude of the cross-power spectral density derived from the same audio stream, whose cross-correlation is depicted in Figure 3.

In preliminary experiments on dialogue detection, two values were only used, namely the value admitted by cross-correlation at zero lag $c_{AB}(0)$ and the cross-spectrum energy in the frequency band [0.065, 0.25] [14]. Both values were compared against properly set thresholds, derived by training, in order to detect dialogues. The interpretation of $c_{AB}(0)$ is straightforward, since it is the product of the two indicator functions. The greater the value of $c_{AB}(0)$ is, the longer time the two actors speak simultaneously. In this paper, we avoid dealing with scalar values, derived from the cross-correlation and the corresponding cross-power spectral density. A more generic approach is adopted, that considers the cross-correlation sequence evaluated on properly chosen time-windows, as is described in Section 3.2, as well as the magnitude of its Discrete Fourier Transform, i.e. the uniform frequency sampling of the cross-power spectral density.

3. Dataset

3.1. *Data description*

A dataset is created by extracting audio scenes from 6 movies, as indicated in Table 1. There are multiple reasons explaining justifying the choice of these movies. First of all, they are considered to be quite popular and accordingly, they are

easily accessible. Secondly, they cover a wide area of movie genres. For example, Analyze That is a comedy, Platoon is an action, and Cold Mountain is a drama. Finally, they have already been widely used in movie analysis experiments. In total, 39 scenes are extracted from the aforementioned movies. To the best of the authors' knowledge, this is the largest movie set used in audio-assisted dialogue detection research. As can be seen in Table 1, 27 out of the 39 scenes correspond to dialogue scenes, while the remaining 12 do not contain any dialogue.

Table 1
The 6 movies used to create the dataset.

Movie name	**Dialogue scenes**	**Non-dialogue scenes**	**Total scenes**
Analyze That	4	2	6
Cold Mountain	5	1	6
Jackie Brown	3	2	5
Lord of the Rings I	5	2	7
Platoon	4	0	4
Secret Window	6	5	11
Total	27	12	39

The audio track of these scenes is digitized in PCM, at a sampling rate of 48 kHz and each sample is quantized in 16 bit two-channel. The total duration of the 39 scenes is 32 min and 10 sec. The dialogue scenes have a total duration of 25 min and 9 sec, while the total duration of non-dialogues is 7 min and 1 sec. Examples of non-dialogue scenes include monologues, music soundtrack, songs, street noise, or instances where the first actor is talking and the second one is just making exclamations.

To fix the number of inputs in the neural networks under study, a running time-window of 25 sec duration is applied to each audio scene. The particular choice of time window duration is justified in Section 3.2. After applying the 25 sec window to the 39 audio scenes, 61 instances are extracted. 41 out of the 61 instances correspond to dialogue instances and the remaining 20 to non-dialogue ones. For a 25 sec window and a sampling frequency of 1 Hz, 2·25-1=49 samples of $c_{AB}(d)$ and another 49 samples of $|\phi_{AB}(f)|$ are computed. The aforementioned 98 samples, plus the label, stating whether the instance is a dialogue or not, are fed as input to the artificial neural networks described in Section 5.

3.2. *Modeling the dataset*

Using the 27 dialogue scenes described in Section 3.1, an effort is made to model the actor utterance duration. The just mentioned duration can be computed as the difference between two successive actor change points. Such a modeling is advantageous, because it enables the use of an informative time-window in the analysis of audio scenes, which contains an adequate number of speaker turn points.

In this context, several distributions are tested for modeling the duration of actor utterances, namely Birnbaum-Saunders, Exponential, Extreme value, Gamma, Inverse Gaussian, Log-logistic, Logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t-location scale, and Weibull. All the aforementioned distributions are parameterized by location and scale parameters μ and σ. To estimate these parameters, maximum likelihood estimation (MLE) is employed. It is worth mentioning that both the log-likelihood and the Kolmogorov-Smirnov criteria yield the Inverse Gaussian as the best fit.

An illustration of the best fit for the aforementioned distributions to the distribution of actor utterances can be depicted in the form of a probability-probability plot (P-P plot). P-P plots are graphical methods to evaluate the goodness-of-fit of the empirical data x to a theoretical distribution. Let $q_i = \frac{i}{n+1}$ be the uniform quantiles, $x_{(i)}$ be the order statistics of the empirical data, and $F()$ be the cumulative density function. The P-P plot is given by the pairs $(q_i, F(\frac{x_{(i)}-\mu}{\sigma}))$. A strong deviation of the P-P plot from the main diagonal in the unit square indicates that the considered model is incorrect [21]. In Figures 5-6, one can see the P-P plots for the aforementioned distributions. The P-P plots for the best and worst model assumptions are shown in Figure 7.

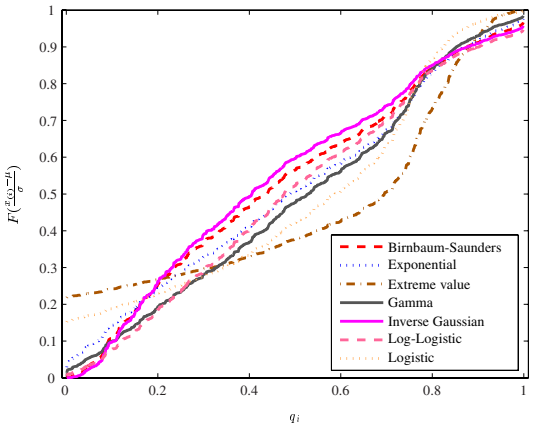

Fig. 5. The P-P plots for distributions Birnbaum-Saunders, Exponential, Extreme value, Gamma, Inverse Gaussian, Log-logistic, and Logistic.

The expected value of the utterance duration has been found equal to 5 sec. This means that actor changes are expected to occur, on average, every 5 sec. We consider that 4 actor changes should occur within the time-window employed in our analysis. Accordingly, an A-B-A-B-A structure is assumed. Similar assumptions were also invoked in [16,26]. As a result, an appropriate time-window should have a duration of $5 \times (4 + 1)$ = 25 sec.

4. Figures of Merit

The most commonly used figures of merit for dialogue detection are described in this Section, in order to enable a comparable performance assessment with other similar works. Let us call the correctly classified dialogue instances $hits_d$ and the correctly classified non-dialogue instances $hits_{nd}$. Then, $misses$ are the dialogue instances that are not classified correctly and $false\ alarms$ are the non-dialogue instances classi-

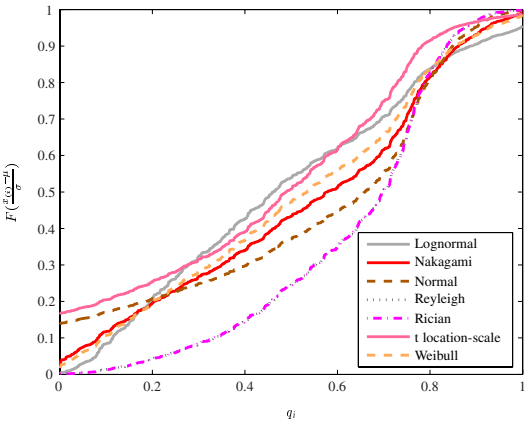

Fig. 6. The P-P plots for distributions Lognormal, Nakagami, Normal, Rayleigh, Rician, t-location scale, and Weibull.

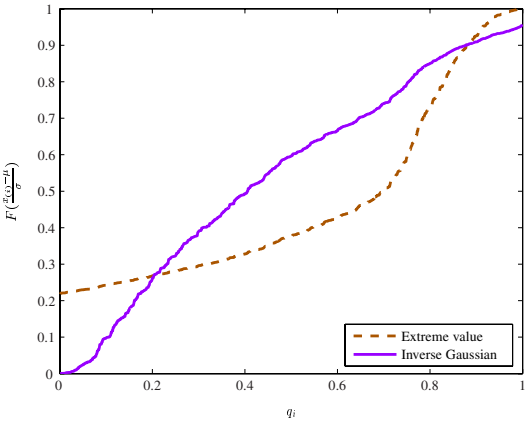

Fig. 7. The P-P plots for the Inverse Gaussian (i.e. the best fit) and the Extreme value (i.e. the worst fit).

fied as dialogue ones. Obviously, the total number of dialogue instances is equal to the sum of $hits_d$ plus $misses$.

Two triplets of figures of merit are employed. The first includes the percentage of correctly classified instances, the percentage of the incorrectly classified instances, and its respective root. The percentage of correctly classified instances (CCI) is defined as:

$$CCI = \frac{hits_d + hits_{nd}}{hits_d + hits_{nd} + misses + false\ alarms} \cdot 100\%. \tag{4}$$

The percentage of incorrectly found instances (ICI) is given by:

$$ICI = \frac{misses + false\ alarms}{hits_d + hits_{nd} + misses + false\ alarms} \cdot 100\%. \tag{5}$$

The root mean squared error ($RMSE$) has also been utilized. For the 2-class classification problem, it is defined as:

$$RMSE = \sqrt{ICI}. \tag{6}$$

Precision (PRC), recall (RCL), and F_1 measure is another commonly used triplet. For the dialogue instances, they are defined as:

$$PRC = \frac{hits_d}{hits_d + false\ alarms} \tag{7}$$

$$RCL = \frac{hits_d}{hits_d + misses}. \tag{8}$$

F_1 measure admits a value between 0 and 1. It is defined as:

$$F_1 = \frac{2\ PRC \cdot RCL}{PRC + RCL}. \tag{9}$$

The higher its value is, the better performance is obtained.

5. Experimental Results using Artificial Neural Networks

Several types of artificial neural networks (ANNs) have been employed for audio-assisted movie dialogue detection. Their performance is evaluated in Sections 5.1-5.7. The experiments are conducted in two distinct stages. In the first stage, simple ANNs have been trained, using as input the cross-correlation and the magnitude of the cross-power spectral density of indicator function pairs. For each feature vector used in training, the class label (dialogue or non-dialogue) is also supplied. The following ANNs are tested: multilayer perceptrons, voted perceptrons, radial-basis function networks, and support vector machines. In the second stage, meta-classifiers, such as the AdaBoost and the MultiBoost algorithms are used, aiming at improving the performance of the aforementioned simple classifiers. The experiments are performed using 7-fold cross validation. For comparative reasons, two commonly used splits between training and test sets are utilized: the 70%/30% and 50%/50% ratios.

5.1. *Perceptrons*

Two variants of the perceptron networks are discussed. The first variant is the multilayer perceptron (MLP) and the second one is the voted perceptron (VP). The latter is a perceptron operating in a higher dimensional space using kernel functions.

MLPs are feed-forward networks, consisting of multiple layers of computational units. In this particular case, there are three layers: the input layer consisting of 98 input nodes (i.e. 49 for $c_{AB}(d)$ and another 49 for $|\phi_{AB}(f)|$), the hidden layer, and the output layer. The learning technique is the back-propagation algorithm. The sigmoid function is utilized as an activation function, the learning rate is equal to 0.3, and the momentum equals 0.2. In general, MLPs tend to overfit the training data, especially when limited training samples are available. In addition, there are problems with computation speed and convergence. The optimization problem using the back-propagation algorithm can be solved with reduced computational cost, by utilizing the fast artificial neural network library (FANN) [18]. 7-fold dialogue detection results using MLPs with 70%/30% and 50%/50% training/test set splits are enlisted in Table 2.

In VP, the algorithm takes advantage of data that are linearly separable with large margins [8]. VP also utilizes the leave-one-out method. For the marginal case of one epoch, VP is equivalent to MLP. The main expectation underlying VP, is that data are more likely to be linearly separable into higher dimension spaces. VP is easy to implement and also saves computation time. Dialogue detection results using 7-fold cross-validation with the two splits, are enlisted in Table 3.

Table 2
7-fold averaged figures of merit for dialogue detection using MLPs.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	90.97%	86.17%
CCI (st. dev.)	3.976%	5.172%
$RMSE$	0.259	0.326
PRC	0.978	0.948
RCL	0.892	0.843
F_1	0.931	0.890

Table 3
7-fold averaged figures of merit for dialogue detection using VPs.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	88.72%	86.63%
CCI (st. dev.)	6.393%	4.337%
$RMSE$	0.305	0.360
PRC	0.864	0.849
RCL	0.998	0.979
F_1	0.920	0.908

5.2. *Radial basis functions*

Radial basis functions (RBFs) can replace the sigmoidal hidden layer transfer function in MLPs. In classification problems, the output layer is typically a sigmoid function of a linear combination of hidden layer values, representing the posterior probability. RBF networks do not suffer from local minima, in contrast to the MLPs. This is because the linear mapping from the hidden layer to the output layer is adjusted in the learning process. In classification problems, the fixed non-linearity introduced by the sigmoid output function, is most efficiently dealt with using iterated reweighted least squares.

A normalized Gaussian RBF network is used in this paper. The k-means clustering algorithm is used to provide the basis functions while the logistic regression model [10] is employed for learning. Symmetric multivariate Gaussians fit the data of each cluster. All features are standardized to zero mean and unit variance. Dialogue detection results using the RBF network are summarized in Table 4.

Table 4
7-fold averaged figures of merit for dialogue detection using RBF networks.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	87.21%	84.78%
CCI (st. dev.)	5.135%	5.499%
$RMSE$	0.318	0.357
PRC	0.908	0.923
RCL	0.913	0.855
F_1	0.906	0.885

5.3. *Support Vector Machines*

Support vector machines (SVMs) are supervised learning methods that can be applied either to classification or regression. SVMs and RBFs are closely connected. SVMs take a different approach to avoid overfitting by finding the maximum-margin hyperplane. In the dialogue detection experiments performed, the sequential minimal optimization algorithm is used for training the support vector classifier [20]. In this implementation, the polynomial kernel is employed, with exponent value equal to 1. Experimental results are detailed in Table 5.

Table 5
7-fold averaged figures of merit for dialogue detection using the SVM classifier.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	87.21%	87.55%
CCI (st. dev.)	5.966%	5.720%
$RMSE$	0.290	0.308
PRC	0.889	0.923
RCL	0.933	0.897
F_1	0.907	0.907

5.4. *AdaBoost*

AdaBoost is a meta-classifier for constructing a strong classifier as linear combination of simple weak classifiers [9]. It is adaptive, in the sense that subsequently built classifiers are tweaked in favor of those instances misclassified by previous classifiers. The biggest drawback of AdaBoost is its sensitivity to noisy data and outliers. Otherwise, it has a better generalization performance than most learning algorithms. In this paper, the AdaBoost algorithm is used to build a strong classifier based on 3-layered MLPs and VPs discussed in Section 5.1, the RBF network presented in Section 5.2, and the SVM classifier outlined in Section 5.3. Dialogue detection results using the AdaBoost algorithm for MLP, VP, RBF networks, and the SVM classifier are shown in Tables 6, 7, 8, and 9, respectively.

Table 6
7-fold averaged figures of merit for dialogue detection using the AdaBoost algorithm with MLPs.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	89.47%	86.62%
CCI (st. dev.)	5.263%	6.014%
$RMSE$	0.317	0.345
PRC	0.934	0.931
RCL	0.924	0.859
F_1	0.924	0.890

5.5. *MultiBoost*

The MultiBoost meta-classifier is an extension to AdaBoost. It can be viewed as combining AdaBoost with wagging [25]. Wagging is a base learning algorithm, that utilizes training cases with differing weights. Using wagging, the high bias of AdaBoost can be significantly diminished. In addition, MultiBoost is more efficient in error reduction than AdaBoost. It has also the advantage of a parallel execution. In this paper, the MultiBoost algorithm is used to boost the classification performance

Table 7
7-fold averaged figures of merit for dialogue detection using the AdaBoost algorithm with VPs.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	87.96%	87.09%
CCI (st. dev.)	5.003%	3.226%
$RMSE$	0.338	0.356
PRC	0.873	0.853
RCL	0.968	0.980
F_1	0.916	0.912

Table 8
7-fold averaged figures of merit for dialogue detection using the AdaBoost algorithm with RBF networks.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	86.46%	87.09%
CCI (st. dev.)	6.695%	5.265%
$RMSE$	0.327	0.347
PRC	0.899	0.932
RCL	0.912	0.878
F_1	0.900	0.902

Table 9
7-fold averaged figures of merit for dialogue detection using the AdaBoost algorithm with the SVM classifier.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	87.21%	89.39%
CCI (st. dev.)	5.966%	3.587%
$RMSE$	0.306	0.309
PRC	0.932	0.921
RCL	0.879	0.927
F_1	0.903	0.922

of the MLPs, the VPs, the RBF network, and the SVM classifier. In Tables 10, 11, 12, and 13, dialogue detection results for the aforementioned classifiers are depicted.

Table 10
7-fold averaged figures of merit for dialogue detection using the MultiBoost algorithm with MLPs.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	90.52%	87.09%
CCI (st. dev.)	5.766%	4.927%
$RMSE$	0.299	0.353
PRC	0.932	0.934
RCL	0.936	0.878
F_1	0.931	0.902

5.6. *Particle Swarm Optimization*

Particle swarm optimization (PSO) is an algorithm inspired by the social behavior of bird flocks and fish schools [7]. In PSO, each candidate solution of the optimization problem is called a particle, which has a current position and a velocity. Particles fly through the problem hyperspace by following the current optimum particles. PSO shares many similarities to genetic algorithms, but has no evolution operators, such as crossover and mutation. In ANNs, the PSO algorithm is utilized as a replacement of the back-propagation algorithm used in feed-forward networks, saving computation time and yielding better results. It should be noted that, PSO networks do not overfit the data and require less computational time, compared to MLP networks. In addition, PSO networks can approximate a nonlinear function better than an MLP network, thus exhibiting a better global convergence. In this paper, a 3-layered feed-forward network is employed, that uses the sigmoid activation function in the hidden layer. The Trelea type-II PSO is employed for learning [23]. In Table 14, dialogue detection results for the 3-layered PSO-trained MLP network are depicted.

Table 11
7-fold averaged figures of merit for dialogue detection using the MultiBoost algorithm with VPs.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	88.72%	87.09%
CCI (st. dev.)	6.393%	4.163%
$RMSE$	0.305	0.354
PRC	0.864	0.859
RCL	0.988	0.973
F_1	0.920	0.912

Table 12
7-fold averaged figures of merit for dialogue detection using the MultiBoost algorithm with RBF networks.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	86.46%	87.09%
CCI (st. dev.)	6.696%	5.265%
$RMSE$	0.327	0.347
PRC	0.899	0.932
RCL	0.912	0.878
F_1	0.900	0.902

Table 13
7-fold averaged figures of merit for dialogue detection using the MultiBoost algorithm with SVM classifier.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	85.71%	88.01%
CCI (st. dev.)	5.856%	3.067%
$RMSE$	0.328	0.324
PRC	0.896	0.928
RCL	0.897	0.898
F_1	0.893	0.911

5.7. *Performance comparison*

Regarding the classification performance of the aforementioned ANNs, the best results are obtained by the 3-layered PSO-based MLP, using the 50%/50% split. While the remaining classifiers are sensitive to initial weights, the performance of the 3-layered PSO-based MLP does not depend on initialization. The second best performance is achieved by the MLP for

Table 14
7-fold averaged figures of merit for dialogue detection using a 3-layered PSO-trained MLP feed-forward network.

Figures of merit	**70%/30%**	**50%/50%**
CCI (mean)	88.88%	91.43%
CCI (st. dev.)	4.535%	4.239%
$RMSE$	0.326	0.283
PRC	0.895	0.900
RCL	0.982	0.987
F_1	0.934	0.941

the 70%/30% split. The high accuracy achieved by the MLPs can be attributed to data overfitting. In the latter case, the network may not yield such a high dialogue detection accuracy, when it is fed by input patterns from a new dataset. This is manifested in the case of 50%/50% split, where the training data are not enough to efficiently train the classifier.

The worst performance is achieved by the RBFs, for the 50%/50% split. It is worth mentioning that RBFs are the ANNs with the lowest performance, even after applying the meta-classifiers of AdaBoost and MultiBoost. This implies that RBFs are not suitable for the classification problem under consideration.

As far as the boosting algorithms are concerned, the AdaBoost algorithm failed to improve the classification accuracy of the MLP and VP networks. However, the accuracy of the RBF network and the SVM classifier using the 50%/50% split is greatly improved. As expected, the MultiBoost algorithm yields a slightly improved performance. The RBF network and the SVM classifier are the most favored in contrast to the MLP and VP networks.

6. Conclusions

In this paper, a novel framework for audio dialogue detection was described, using the cross-correlation of a pair of indicator functions and the corresponding cross-power spectral density as features. Audio scenes, containing dialogues and non-dialogues, were extracted from six movies. To the best of the authors' knowledge, this is the largest set of movies used in works related to audio-assisted dialogue detection. Dialogue scenes were used to model the duration of actor utterances. A variety of artificial neural networks was employed for dialogue detection namely MLP, VP, RBF, SVM (with and without application of AdaBoost and MultiBoost), and 3-layered PSO-based MLP networks. The experimental results indicate that, the highest average dialogue detection accuracy is achieved by the 3-layered PSO-based MLP, equal to 91.43%. All results are apparently superior to state-of-the-art dialogue detection techniques [15]. However, there is a lack of direct comparison, due to the fact that there is no common database for performance evaluation. The present paper uses a multitude of commonly employed objective figures, in order to assess the performance of the ANNs studied for dialogue detection. The reported dialogue detection accuracy and F_1 measure can be treated as an upper bound of the actual figures of merit for dialogue detection in movies that could be obtained by employing speaker turn detection, speaker clustering, and speaker tracking.

Acknowledgement

This work has been supported by the "Pythagoras-II" Programme, funded in part by the European Union (75%) and in part by the Greek Ministry of National Education and Religious Affairs (25%). E. Benetos is a scholar of the "Alexander S. Onassis" Public Benefit Foundation.

References

[1] A. A. Alatan, A. N. Akansu, and W. Wolf, Comparative analysis of hidden Markov models for multi-modal dialogue scene indexing, in: Proc. 2000 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Vol. 4 (2000) 2401-2404.

[2] 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 (2001) 137-151.

[3] D. Arijon, Grammar of the Film Language, (Silman-James Press 1991).

[4] B. Birge, PSOt - A particle swarm optimization toolbox for Matlab, in: Proc. 2003 IEEE Swarm Intelligence Symp., (2003) 182-186.

[5] R. J. Boys and D. A. Henderson, A Bayesian approach to DNA sequence segmentation, in: Proc. 2004 Biometrics, Vol. 60 (3) (September 2004) 573-588.

[6] L. Chen and M. T. Özsu, Rule-based extraction from video, in: Proc. 2002 IEEE Int. Conf. Image Processing, Vol. II (2002) 737-740.

[7] R. Eberhart and J. Kennedy, A new optimizer using particle swarm theory, in: Proc. 6th Int. Symp. Micro Machine and Human Sceince, (October 1995) 39-43.

[8] Y. Freund and R. E. Schapire, Large margin classification using the perceptron algorithm, Machine Learning, Vol. 37 (3) (1999) 277-296.

[9] Y. Freund and R. E. Schapire, A short introduction to boosting, J. Japanese Society for Artificial Intelligence, Vol. 14 (5) (September 1999) 771-780.

[10] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, (N.Y.: Wiley, 2000).

[11] G. Iyengal, H. J. Nock, and C. Neti, Audio-visual synchrony for detection of monologues in video archives, in: Proc. 2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Vol. I (April 2003) 329-332.

[12] F. Jelinek, Statistical Methods for Speech Recognition, (Cambridge, Massachusetts: The MIT Press, 1997).

[13] 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) 1856-1859.

[14] M. Kotti, C. Kotropoulos, B. Ziólko, I. Pitas, and V. Moschou, A framework for dialogue detection in movies, in: Lecture Notes in Computer Science, Vol. 4105 (Springer, Istanbul, September 2006) 371-378.

[15] 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 (2005) 825-828.

[16] 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 (2005) 286-296.

[17] 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 (May 2004) 741-744.

[18] S. Nissen, Implementation of a fast artificial neural network library (FANN), Technical Report, Department Computer Science University of Copenhagen, Denmark (October 2003).

[19] A. Papoulis and S. V. Pillai, Probabilities, Random Variables, and Stochastic Processes (4/e. N.Y.: McGraw-Hill, 2002).

[20] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schoelkopf, C. Burges, and A. Smola, eds., Advances in Kernel Methods - Support Learning (MIT Press, 1999).

[21] R. D. Reiss and M. Thomas, Statistical Analysis of Extreme Values, (Basel, Switzerland: Birkhäuser Verlag, 1997).

[22] P. Stoica and R. L. Moses, Introduction to Spectral Analysis, (Upper Saddle River, NJ: Prentice Hall, 1997).

[23] I. C. Trelea, The particle swarm optimization algorithm: convergence analysis and parameter selection, Information Processing Letters, Vol. 85 (2003) 317-325.

[24] 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 (2004) 587-590.

[25] I. Webb, MultiBoosting: A technique for combining boosting and wagging, J. Machine Learning, Vol. 40 (2) (2000) 159-196.

[26] Y. Zhai, Z. Rasheed, and M. Shah, Semantic Classification of Movie Scenes Using Finite State Machines, IEE Proc. - Vision, Image, and Signal Processing, Vol. 152 (6) (December 2005) 896-901.

[27] MPEG-7 overview (version 9), *ISO/IEC JTC1/SC29/WG11 N5525* (March 2003).