Automatic Sound Classification of Radio Broadcast News

Theodoros Theodorou[1], Iosif Mporas[1,2] and Nikos Fakotakis[1]
[1]*Artificial Intelligence Group, Wire Communications Laboratory,*
Dept. of Electrical and Computer Engineering, University of Patras,
26500 Patras, Greece
[2]*Dept. of Informatics and Means of Mass Communication,*
Technological Educational Institute of Patras,
27100 Pyrgos, Greece
{theodorot, imporas, fakotaki}@upatras.gr

Abstract

Automatic extraction of the index of broadcast streams from radio and television has become a challenging research topic over the last years. The automatic classification of audio types, such as speech, music, noises/atypical events etc, has found numerous applications. In this paper we study the evaluation of different machine learning algorithms, which have successfully been used in other classification tasks, on the task of classification of audio broadcast news. The audio classification scheme consists of pre-processing, audio parameterization with established audio features, and classification to acoustic events. The experimental evaluation was carried out using the Voice of America broadcast recordings database for the Greek language. The experimental results indicated that the best performance, approximately 92% of accuracy, was achieved by the classification scheme using the boosting technique with decision trees.

Keywords: *audio classification, sound recognition, audio broadcast news.*

1. Introduction

The development of information and communication technologies (ICT) over the last years has lead to a rapid increase of the amount of audiovisual data broadcasted over the radio, the television and the Internet [1, 2, 3 ,4]. Due to this explosive growth of available multimedia content (e.g. video-on-demand web services), there exist need for technologies that can automatically process this audiovisual data and extract the corresponding content. In the area of audio processing, the challenge is the automatic indexing of audio streams, which typically consist of combinations of overlapping acoustic events, such as speech, music, typical and atypical noise sounds etc. Thus audio classification architectures are needed for the automatic segmentation of audio data to acoustic categories of interest [5, 6, 7], in order the corresponding audio segments to be post-processed by appropriate systems, such as speech recognizers, speaker recognizers, singer recognizers, song recognizers, music processing systems for automatic copyrights control etc.

Audio categorization architectures can briefly be analyzed in four parts: the feature extraction stage, the segmentation stage, the classification stage, and the post-processing stage [8, 9]. When the incoming audio signal is introduced to the system the feature extraction stage pre-processes the audio signal and frame blocks it. Using digital signal processing algorithms, for each frame of audio samples a parametric vector is computed (*feature vector*). The feature vector describes the temporal characteristics of the signal. Typically, the frame blocking of the audio signal is performed using frames of constant length, overlapping to their adjacent

ones by a constant number of audio samples. The output of this first stage is a sequence of feature vectors, which are fed to the segmentation stage for further processing.

The estimated sequence of feature vectors is processed by the segmentation stage. However, in some cases the segmentation and classification are performed together in one audio processing step. The goal of the segmentation step is to separate the input audio signal into segments, the content of which will correspond to unique acoustic types, e.g. speech, silence, music etc. The segmentation is performed under the assumption that audio frames of the same acoustic type will have similar acoustic characteristics, and thus similar feature frames. Within the segmentation stage a comparison of the adjacent frames is performed, usually frame by frame, and results to the division of the original audio signal to a number of consequent segments, assuming that only one acoustic type exists within each segment. The estimated sequence of segments will be forwarded to the classification stage. It is expected that adjacent segments correspond to different acoustic types [10]. There are a number of architectures and algorithms proposed in the literature. Most of them calculate a distance between adjacent feature vectors and compare this distance against an experimentally determined threshold. Such distance metrics are the Euclidean distance, the Bayesian information criterion (*BIC*), the Kullback Leibler (*KL2*) distance, the Hotelling's T2 statictic (*T2*) and the generalized likelihood ratio (*GLR*) [11, 12, 13]. Other approaches use probabilistic models to estimate the log-likelihood of the candidate positions of acoustic type transitions. In those approaches one probability density function (*pdf*) is used to describe each acoustic type and the evaluation of them to the incoming audio stream indicates the most probable among the modeled acoustic types [14]. In order to enhance the segmentation stage, some approaches include a silence detection module [14, 15]. This module pre-segments the initial audio signal to silence and non-silence parts and feeds the segmentation algorithm only with the non-silence parts.

In the literature there is a wide variety of acoustic features used [16, 17]. Depending on the structure of the architectures, there are two major groups of commonly used features. The frame-based processing architectures use spectral characteristics [15], while the segment-based processing architectures mainly use the linear predictive coefficients (*LPCs*), the mel-frequency cepstral coefficients (*MFCCs*) and the zero crossing rate (*ZCR*) [10, 11].

These features are based on extracting coefficients that describe a part of the signal identity. In the literature there is another point of view for the extraction procedure. Instead of producing a coefficient, the audio segment is transformed into a key signal. These signals are unlike to each other depending on the nature of the audio signal and the classification stage process them as they were symbols on a dictionary [18, 19].

After the feature processing or the segmentation stage the feature vectors that describe the segments are introduced to the classification stage. There is a wide variety of classifiers in the literature for sound recognition of audio broadcast news. Some of the most commonly used classification algorithms are the support vector machines (*SVMs*), the Gaussian mixture models (*GMMs*), the hidden Markov models (*HMMs*), the artificial neural networks (*ANNs*) and fuzzy logic (*FL*) techniques [15, 20, 21, 22, 23]. The selection of the appropriate classification algorithm is crucial for the overall performance of the audio classification system, both for architectures where segmentation and classification are performed separately or simultaneously within the same stage. The selection of the appropriate classification algorithm can be based on the type of existing acoustic events in the broadcasts of interest as well as on parameters like the time efficiency and the necessary ability for multitasking [15, 24] or the ability to process real time data without knowing the nature of upcoming sound sources [21].

The smoothing stage is optionally interpolated after the classification stage. In some cases the classifier creates regions of data of one class with too small duration and different categorization from the adjacent regions, e.g. the previous and next region, which could be of the same or different acoustic type. Such detected occurrences of acoustic types of very short duration, e.g. of one second, may correspond to false detections. For this reason smoothing decision rules are typically applied in order to refine the estimated acoustic classification [15, 25].

In this article we study the classification performance of a number of powerful machine learning algorithms, which are widely used in other classification applications. The evaluated scheme is performing classification of the oncoming broadcast audio stream in frame level. The structure of the article is organized as follows. In section 2 we offer a detailed description of the architecture of the scheme which was evaluated. Section 3 describes the experimental protocol that was followed, the broadcast news database and the classification algorithms that were used. In section 4 we present the experimental results of the evaluation. Finally, in section 5 the conclusions of this work are discussed.

2. Scheme Architecture Description

This section provides detailed description of the architectural scheme utilized in the present evaluation. The block diagram of the scheme used for audio classification of broadcast news is shown in Figure 1. The audio classification scheme can briefly be divided in two stages. In the first stage the incoming broadcast audio stream is pre-processed and parameterized, while in the second stage the parametric vectors are processed by an acoustic type classification module. The classified to acoustic categories audio frame fragments are forwarded for post-processing to other modules/systems.

In the first stage, the samples of the audio broadcast stream are frame blocked to overlapping frames of constant length, using constant time shifting. The corresponding frames will afterwards be processed by short-time frequency-domain and/or time-domain algorithms in order one feature vector for each frame of audio samples to be constructed. This procedure, which is typical in speech and audio processing, is essential in order to reduce the amount of information that the frames carry in the next stage. The overlapping of the adjacent frames allows the smooth capturing of potential short-time acoustic events. Each of the parametric techniques which are used by the parameterization module will estimate one feature sub-vector for every audio frame and the concatenation of all sub-vectors will result to one total feature vector for each of the processed audio frames. Thus, the output of the first stage will constitute of a sequence of feature vectors (output of the parameterization component in Figure 1) representing the time-varying acoustic characteristics of the input audio broadcast signal.

In the second stage of the present scheme, the sequence of feature vectors (i.e. the unknown or test audio data) will be processed in order to compute the corresponding audio classification estimations. In particular, each feature vector describing the acoustic characteristics of the corresponding audio frame will be feed to the classification algorithm utilized by the audio classification component. The classification algorithm will decide the sound category (e.g. speech, silence, music, etc) in which every frame belongs to, utilizing a number of pre-trained sound models, as illustrated in Figure 1. For the construction of the sound models, one for each audio category of interest, a sufficient amount of training audio data for each of these categories is used. The training data are used to estimate the parameters of the classification algorithm and build the corresponding sound models (e.g. speech model, silence model, music model, etc). During the operational phase, for each feature vector of the

test audio data a confidence score to belong to each of the sound models will be produced and the corresponding feature vector will be labeled with the sound category the model of which offered the maximum confidence score.

The output of the acoustic type classification, i.e. the estimated sound identity of each audio frame, can further be post-processed. Depending on the type of the recognized sound, the corresponding audio frames can be processed by other modules/systems for different applications, i.e. speech parts can be processed by speech/speaker recognizers, music parts can be processed by music category recognizers etc.

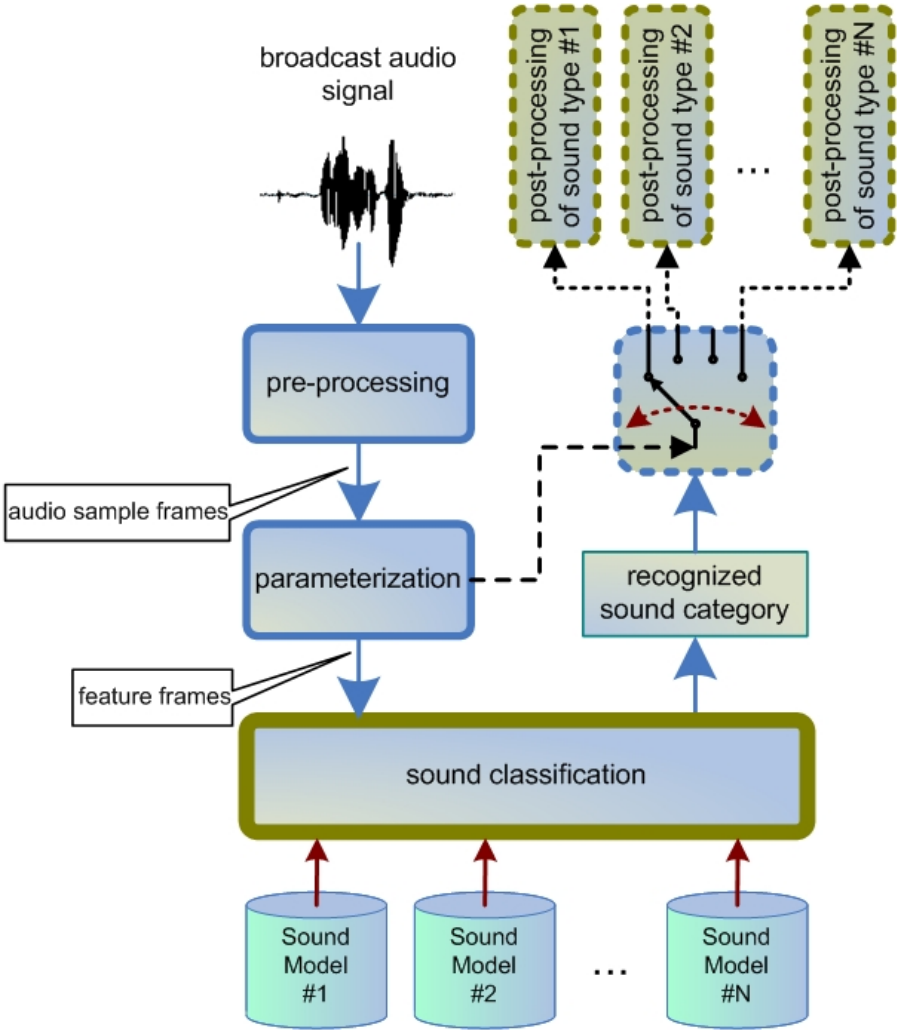

Figure 1: Block diagram of the scheme for audio classification of broadcast news.

The modular architecture of the above described evaluated scheme offers independence between the feature extraction process and the classification process, i.e. the length and time shift of the audio frames as well as the parametric techniques that are utilized are configurable and the classification algorithm can process them using sound models trained with similarly configured data. On the other hand different classification algorithms can be utilized regardless of the feature vector. The utilization of the appropriate audio parameters and the appropriate machine learning classification algorithm are of major importance for the

accuracy of automatic indexing of audio broadcast streams. In the present work we utilize a set of well known and widely used audio parameters, which are described in section 3, and focus on the examination of the audio classification performance for different classification algorithms.

3. Experimental Setup

In this section the experimental protocol followed in the evaluation of the audio classification scheme is described. In particular, the evaluated audio data, the audio parameterization techniques and the machine learning algorithms that were used are presented.

3.1. Broadcast News Database

The evaluation of the performance of the audio categorization scheme for different classification algorithms was carried out using recordings of the Voice of America (*VOA*) radio broadcast news [26] for the Greek language. The *Greek VOA* newscasts data consists of files with broadcast audio streaming recordings of twelve minutes duration. The recordings include speech in the Greek language from different male and female speakers (broadcasters, reporters, interviewees, etc), different kinds of music, human and non-human noises as well as simultaneous presence (overlapping) of speech and music. The broadcasts are in broadband quality, while some telephone quality speech parts from reporters-correspondents exist in the audio recordings. All recordings are stored in mu-law compressed single-channel audio files with sampling frequency 8 kHz and resolution analysis 8-bits.

For the evaluation of the audio classification scheme part of the Greek data, which were distributed by NIST during the 2009 Language Recognition Evaluation were selected [27]. In detail, we used 100 audio recording files of the *Greek VOA*, which were manually annotated by engineer expert in audio processing. The annotation was performed with the utilization of Praat [28] software tool. The manually produced annotations consist of time-aligned marking of the sound events that appear in each recording file. The sound events that were adopted in the present study are: (i) *speech*, (ii) *music*, (iii) *silence*, (iv) *noises*, (v) *speech and music*. As considers the speech sound type, it consists of the voices of native Greek speakers (both single and multiple speakers at each time). The music category included mainly instrumental songs of different music kinds. Parts of the recordings without the presence of speech or music were labeled as silence, while interfering noises constituted another sound category. Finally, a sound category with the presence of both speech and music was labeled.

3.2. Audio Parameterization

The audio recordings of the *Greek VOA* audio data were parameterized using both time-domain and frequency-domain features. In particular, each audio file was blocked to overlapping frames of 20 milliseconds length and 10 milliseconds time shift. For each frame of samples a number of parameterization algorithms were applied in order to extract the corresponding audio features. The features used in the present evaluation were (i) the 12 first Mel frequency cepstral coefficients (*MFCCs*) [29], (ii) the *energy*, (iii) the zero-crossing rate (*ZCR*), (iv) the *voicing probability*, (v) the fundamental frequency (*F0*) and (vi) the *F0 contour envelope*. These features [26] have been used in the literature for the speech recognition [31], speech/music discrimination and audio segmentation tasks [32]. The computed for each frame audio parameters were concatenated in a single feature vector,

1×18, in order to be used by the audio classification module. The audio parameterization procedure is illustrated in Figure 2.

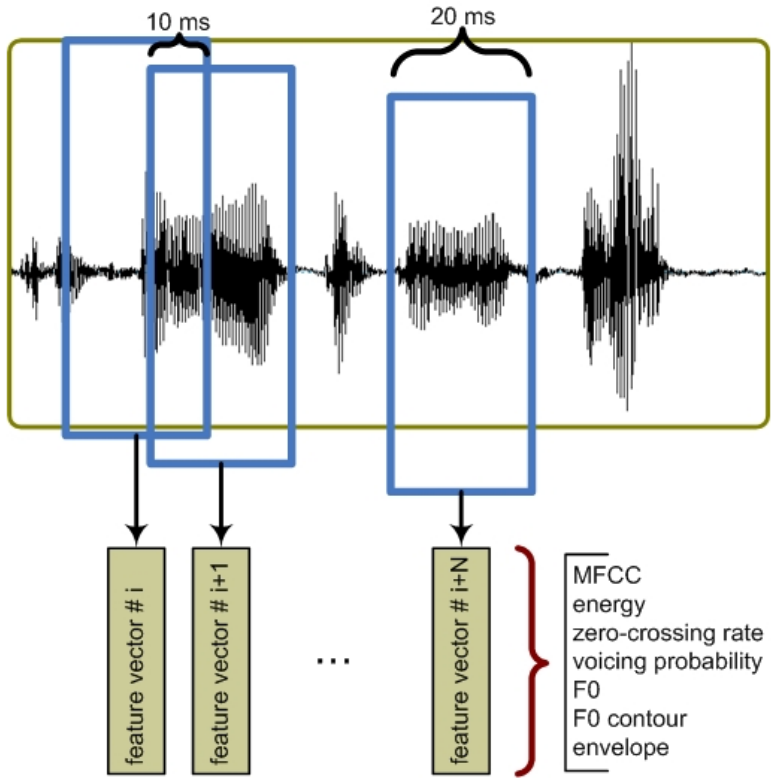

Figure 2: Construction of feature vectors from the audio broadcast signal.

3.3. Classification Algorithms

The audio classification scheme, presented in section 2, was evaluated using different machine learning algorithms. In detail, we used a two-layered back-propagation multilayer perceptron neural network (*MLP*) [33], a *naïve Bayes* classifier [34], a support vector classifier with radial basis function kernel utilizing the sequential minimal optimization algorithm (*SVM*) [35], a k-nearest neighbour classifier (*IBk*) [36] and a C4.5 decision tree learner (*J48*) [37]. Except these, we employed one bagging algorithm, using fast tree learner with reduced error pruning (*REPTree*) [38]. Finally, a boosting algorithm combined with J48 decision trees (*Adaboost.M1.J48*) [39] was used. For the evaluation of these classification algorithms we relied on their implementation in the WEKA machine learning toolkit [40].

In order to ensure the reliability of our experimental results we performed ten-fold cross validation experimentations, using ninety percent of the data for training and ten percent for testing during each fold. Thus, there was no overlapping between the training and test data for any of the performed experiments.

4. Experimental Results

The audio classification scheme presented in section 2 was evaluated under the experimental protocol described in section 3. In this section we present the experimental results of the evaluation. All classification algorithms were evaluated utilizing the audio features described in section 3 on the same evaluation data. The performance of each classification algorithm on the task of audio categorization of radio broadcasts was evaluated by estimating the percentage of the audio frames which were correctly recognized. As ground truth for the sound identity of each frame we adopted the manual annotations of the VOA recordings. The audio categorization performance for each of the evaluated classification algorithms is shown in Table 1.

Table 1. Audio categorization performance results, in percentages, for each of the seven evaluated classification algorithms.

Method	**Accuracy (%)**
Naïve Bayes	70.38
SMO-SVM	72.49
MLP	83.63
IBk	86.80
J48	88.07
Bagging (REP Trees)	88.80
AdaBoost.M1 (J48)	91.82

As can be seen in Table 1, the best performance, 91.82%, was achieved by the boosting algorithm combined with decision trees (AdaBoost.M1-J48). The best performing AdaBoost.M1 (J48) was followed by the bagging algorithm, using fast tree learner with reduced error pruning (Bagging REP Trees), which achieved overall audio classification accuracy of 88.80%. The two meta-classifiers were followed in terms of accuracy ranking by the decision tree J48 with 88.07%, the IBk algorithm with accuracy 86.80%, the MLP neural networks with 83.63%, the SMO support vector machines by 72.49% and the Naïve Bayes algorithm with 70.38%. The superior performance of the meta-classifiers is mainly owed to the ability of them to create more stable and accurate learners, using the existing classifiers by random redistribution of the training data (bagging) or combination of weak learners to build strong ones (boosting) [40]. Particularly, the combination of meta-classification techniques with a well-performing on the specific task algorithm, such as the J48 decision tree, increases significantly the performance (91.82%). On the other hand, support vector machines did not show competitive performance probably, since they do not suffer from the curse of dimensionality and thus are advantageous when processing large length feature vectors. However, the present feature vectors' size is limited to 1×18, since we utilized established audio descriptors [30, 31, 32]. It is worth mentioning that the above reported performance depends also on the audio data characteristics that were evaluated, as well as on the five sound categories that have been adopted.

It is interesting to evaluate the audio categorization performance in terms of pairs of sound categories, in order to examine the degree of misclassification between pairs of sound types. In Table 2, we present the confusion matrix of the best, in average, performing boosting algorithm combined with decision trees (AdaBoost.M1-J48). The confusion matrix shows for each of the adopted sound categories (rows in Table 2) the percentage of the audio data of radio broadcast streams that were classified to each sound category (columns in Table 2). The

cells that correspond to the correctly classified audio frames are highlighted in grey. The cells highlighted in bold correspond to the sound type that introduced the maximum classification error to each sound category.

As can be seen in Table 2, the sound category that was most difficult to be classified was the simultaneous presence of music and speech, which was classified correctly 83.50% of the times. On the other hand, silence sound category was classified correctly 98.67% of the times it was found. The maximum classification errors of the audio frames are distributed as follows. Music sound category was recognized as music and speech by 2.19%. Music and speech sound type was classified as music by 2.66% and as speech by 12.28%. Speech was classified as speech and music by 9.38%. Finally, noise category was classified as speech by 3.21% and as music and speech by 1.02%.

Table 2. Confusion matrix of the audio categorization performance results, in percentages, for each of the adopted sound categories for the best performing AdaBoost.M1 (J48) algorithm.

Input \ Recognized as	**Silence (%)**	**Music (%)**	**Music & Speech (%)**	**Speech (%)**	**Noise (%)**
Silence	98.67	**00.70**	00.00	00.00	00.63
Music	00.70	96.64	**02.19**	00.08	00.39
Music & Speech	00.00	02.66	83.50	**12.28**	01.56
Speech	00.00	00.16	**09.38**	85.93	04.53
Noise	00.86	00.55	01.02	**03.21**	94.37

The above results clearly indicate the difficulty in the task of speech-music discrimination from audio data, which is in agreement with other related studies [15]. In the environment of radio broadcast audio streaming the presence of speech from broadcasters and music at the same time is not rare, thus making the specific task more challenging. To resolve this difficulty, speech and music can be considered as one sound class and further be processed by classifiers dedicated to speech/music discrimination. Moreover, the presence of non constant number of speakers speaking simultaneously makes the task even more complicated. In the VOA radio broadcast audio data, the boosting algorithm combined with decision trees (AdaBoost.M1-J48) offered a classification accuracy of more than 90% for the three classes: music, speech, music and speech. The performance of the audio categorization stage will directly affect the performance of the post-processing modules, i.e. the speech recognition engine, which should process purely speech data since introduction of non-speech data would result to reduction of the word recognition rate.

5. Conclusions

Automatic audio classification of radio broadcast streaming has become a research area of great interest over the last years. The availability of enormous amounts of multimedia data in combination with the need for automatic processing of them for tasks, such as speech transcription extraction, speaker change detection, scene change detection etc, widows the development of accurate architectures/algorithms for automatic audio categorization very important.

In this article we presented a scheme for automatic categorization of radio broadcast data of the Voice of America database for the Greek language. The scheme utilizes a common set of pre-processing and parameterization modules, utilizing well known and widely used acoustic parameters. The audio classification module was evaluated using a number of classification algorithms, which have successfully been used in other machine learning applications. The experimental results showed that the meta-classification algorithms that were evaluated here achieved higher classification scores than all the other classification techniques. In particular, the boosting algorithm combined with decision trees (AdaBoost.M1-J48) achieved approximately 92% accuracy, when categorizing the radio broadcast data among the classes: speech, silence, music, speech and music, noise. The analysis of the experimental results per sound category showed that the simultaneous existence of speech and music significantly drops the audio classification performance. The utilization of the appropriate, in terms of audio classification accuracy, machine learning algorithm will affect the overall performance of the broadcast audio processing and it's sub-products, i.e. speech transcribers, speaker diarization, sound event detection etc.

6. References

[1] H. D. Tran and H. Li, "Sound Event Recognition with Probabilistic Distance SVMs", IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 6, August 2011.

[2] Y. Itoh, S. Sakaki, K. Kojima and M. Ishigame, "Highlight Scene Extraction of Sports Broadcasts Using Sports News Programs", In Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing, 2008, pp. 646-649.

[3] M.-L. Shyu, G. Ravitz, and S.-C. Chen, "Unit Detection in American Football TV Broadcasts Using Average Energy of Audio Track", In Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering, MSE, 2004, Miami FL USA, pp. 193-200.

[4] S.-C. Chen, M.-L. Shyu, W. Liao, and C. Zhang, "Scene Change Detection by Audio and Video Clues", In Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2002, 2002, Lausanne Switzerland, pp 365-368.

[5] J. Zhang, B. Jiang, L. Lu, and Q. Zhao, "Audio Segmentation System for Sport Games", In Proceedings of the 2010 International Conference on Electrical and Control Engineering, 2010, pp. 505-508.

[6] M. Kos, M. Grasic, D. Vlaj, and Z. Kacic, "On-line Speech/Music Segmentation for Broadcast News Domain", In Proceedings of the 16th International Conference on Systems, Signals and Image Processing, IWSSIP 2009, 2009.

[7] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast News Subtitling System in Portuguese", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, 2008, pp. 1561-1564.

[8] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker Segmentation and Clustering", Signal Processing, 2008, vol. 88, pp. 1091-1124.

[9] G. Richard, M. Ramona and S. Essid, "Combines Supervised and Unsupervised Approaches for Automatic Segmentationa of Radiophonic Audio Streams", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, 2007, vol. 2, pp. II-461 - II-464.

[10] J. Huang, Y. Dong, J. Liu, C. Dong and H. Wang, "Sports audio segmentation and classification", In Proceedings of the IEEE International Conference on Network Infrastructure and Digital Content, 2009, IC-NIDC 2009, 2009, pp. 379-383.

[11] R. Huang and J.H.L. Hansen, "Advances in unsupervised audio segmentation for the broadcast news and NGSW corpora", In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, 2004, vol. 1, pp. I-741-4.

[12] B. Zhou and J.H.L. Hansen, "Efficient Audio Stream Segmentation via the Combined T2 Statistic and Bayesian Information Criterion", IEEE Transactions on Speech and Audio Processing, July 2005, vol. 13, no. 4, pp. 467-474.

[13] D. Wang, R. Vogt, M. Mason, and S. Sridharan, "Automatic Audio Segmentation Using the Generalized Likelihood Ratio", In Proceedings of the 2nd International Conference on Signal Processing and Communication Systems, ICSPCS 2008, 2008.

[14] C.-H. Wu and C.-H. Hsieh, “Multiple Change-Point Audio Segmentation and Classification Using an MDL-Based Gaussian Model”, IEEE Transactions on Audio, Speech, and Language Processing, March 2006, vol. 14, no. 2, pp. 647-657.

[15] E. Dogan, M. Sert and A. Yazici, “Content-Based Classification and Segmentation of Mixed-Type Audio by Using MPEG-7 Features”, In Proceedings of the 2009 First International Conference on Advances in Multimedia, 2009, pp. 152-157.

[16] V. Gupta, G. Boulianne, P. Kenny, P. Ouellet and P. Dumouchel, “Speaker Diarization of French Broadcast News”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, 2008, pp. 4365-4368.

[17] Y. Patsis and W. Verhelst, “A Speech/Music/Silence/Garbage Classifier for Searching and Indexing Broadcast News Material”, In Proceedings of the 19th International Workshop on Database and Expert Systems Application, DEXA ’08, 2008, pp. 585-589.

[18] T. H. Dat and L. Haizhou, “Jump function komogorov and its application for audio stream segmentation and classification”, IEEE Transactions on Signal Processing, August 2009, vol. 57, no. 8, pp. 2908-2918.

[19] S. Chu, S. Narayanan and C.-C. Jay Kuo, “Environmental sound recognition using MP-based features”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, 2008.

[20] H. Meinedo and J. Neto, “Audio Segmentation, Classification and Clustering in a Broadcast News Task”, In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003, 2003, pp. II 5-8.

[21] M. Liu, C. Wan and L. Wang, “Content-based audio classification and retrieval using a fuzzy logic system: towards multimedia search engines”, Soft Comput., 2002, vol. 6, no. 5, pp. 357-364.

[22] H. Aronowitz, “Segmental modeling for audio segmentation”, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, 2007, pp. IV-393 - IV-396.

[23] Y.A. Alotaibi, and A. Hussain, “Comparative Analysis of Arabic Vowels using Formants and an Automatic Speech Recognition System”, International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 3, no. 2, 2010.

[24] L. Lu, S. Z. Li, and H.-J. Zhang, “Content-based audio segmentation using support vector machines”, In Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2001, 2001, pp. 749-752.

[25] L. Lu, H.-J. Zhang and H. Jiang, “Content Analysis for Audio Classification and Segmentation”, IEEE Transactions on Speech and Audio Processing, October 2002, vol. 10, no. 7, pp. 504-516.

[26] http://www.voanews.com/.

[27] http://www.itl.nist.gov/iad/mig/tests/lre/2009/.

[28] P. Boersma, and D. Weenink, “Praat: doing phonetics by computer (Version 5.1.05) [Computer program]”. Retrieved May 1, 2009, from http://www.praat.org/.

[29] M. Slaney, “Auditory Toolbox Version 2 Technical Report #1998-010 Interval Research Corporation”, 1998.

[30] F. Eyben, M. Wöllmer, and B. Schuller, “OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit”, In Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction, ACII 2009, 2009.

[31] A.N. Mishra, M. Chandra, A. Biswas, and S.N. Sharan, “Robust Features for Connected Hindi Digits Recognition”, International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 4, no. 2, 2011.

[32] H.-Y. Lo, J.-C. Wang, and H.-M. Wang, “Homogeneous Segmentation and Classifier Ensemble for Audio Tag Annotation and Retrieval”, In Proceedings of the IEEE International Conference on Multimedia & EXPO, ICME 2010, Singapore, July 2010.

[33] T.M. Mitchell, Machine Learning, McGraw-Hill International Editions, 1997.

[34] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[35] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy, “Improvements to Platt's SMO Algorithm for SVM Classifier Design”, Neural Computation, vol. 13, no. 3, 2001, pp. 637-649.

[36] D. Aha, and D. Kibler, “Instance-based learning algorithms”, Machine Learning, vol. 6, 1991, pp. 37-66.

[37] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[38] L. Breiman, “Bagging predictors”, Machine Learning, vol. 24, no. 2, 1996, pp. 123-140.

[39] Y. Freund, and R.E. Schapire, “Experiments with a new boosting algorithm”, In Proceedings of the 13th International Conference on Machine Learning, San Francisco, USA, 1996, pp. 148-156.

[40] I.H. Witten, and E. Frank, Data mining: practical machine learning tools and techniques (2nd ed.), Morgan-Kaufman Series of Data Management Systems, Elsevier, San Francisco, 2005.

Authors

Mr. Theodoros Theodorou was born in Athens, Greece, in 1986. He graduated in 2008 (Diploma) with excellent grade from the Department of Electrical and Computer Engineering of the University of Patras, Greece. Currently he is PhD candidate at the Department of Electrical and Computer Engineering of the University of Patras. His research interests include audio segmentation, audio and speech processing and sound source enumeration.

Dr. Iosif Mporas was born in Athens, Greece, in 1981. He graduated in 2004 (Diploma) from the Department of Electrical and Computer Engineering of the University of Patras, Greece. He received his PhD degree in July 2009. Currently he is post-doctoral researcher at the University of Patras and non-tenured Assistant Professor at the Technological Educational Institute of Patras. He is author and co-

author in more than 50 publications in scientific journals and international conferences. His research interests include speech and audio signal processing, pattern recognition, automatic speech recognition, automatic speech segmentation and spoken language/dialect identification.

Prof. Nikos Fakotakis received his B.Sc. degree from the University of London (UK) in Electronics in 1978, M.Sc. degree in Electronics from the University of Wales (UK), and his Ph.D. degree in Speech Processing from the University of Patras, (Greece), in 1986. From 1986 to 1992 he was a lecturer in the Electrical and Computer Engineering Dept. of the University of Patras, from 1992 to 1999 an Assistant Professor, from 2000 to 2003 an Associate Professor, and since 2003 he is a full Professor in the area of Speech and Natural Language Processing. Prof. Fakotakis is director of the Communication and Information Technology Division, director of the Wire Communications Laboratory (WCL), and Head of the Artificial Intelligence Group. He is author of over 300 publications in the area of Speech and Natural Language Engineering, and Artificial Intelligence. His current research interests include AI, Speech Recognition/Understanding, Speaker Recognition, User Modeling, Spoken Dialogue Processing, and Natural Language Processing.