I.J. Information Technology and Computer Science, 2014, 11, 1-9
Published Online October 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2014.11.01

An Overview of Automatic Audio Segmentation

Theodoros Theodorou[1], Iosif Mporas[1,2], Nikos Fakotakis[1]
[1]Artificial Intelligence Group, Wire Communications Laboratory, Department of Electrical and Computer Engineering, University of Patras, Patras 26500, Greece
[2]Computer Engineering and Informatics Department, Technological Educational Institute of Western Greece, Antirion 30300, Greece
Email: {theodorou, imporas, fakotaki}@upatras.gr

***Abstract*—** In this report we present an overview of the approaches and techniques that are used in the task of automatic audio segmentation. Audio segmentation aims to find changing points in the audio content of an audio stream. Initially, we present the basic steps in an automatic audio segmentation procedure. Afterwards, the basic categories of segmentation algorithms, and more specific the unsupervised, the data-driven and the mixed algorithms, are presented. For each of the categorizations the segmentation analysis is followed by details about proposed architectural parameters, such us the audio descriptor set, the mathematical functions in unsupervised algorithms and the machine learning algorithms of data-driven modules. Finally a review of proposed architectures in the automatic audio segmentation literature appears, along with details about the experimenting audio environment (heading of database and list of audio events of interest), the basic modules of the procedure (categorization of the algorithm, audio descriptor set, architectural parameters and potential optional modules) along with the maximum achieved accuracy.

***Index Terms*—** Audio Segmentation; Sound Classification; Machine Learning; Mathematical Functions; Hybrid Architecture of Unsupervised and Data-Driven Algorithms

I. Introduction

Automatic audio segmentation aims to divide a digital audio signal into segments, each of which contains audio information from a specific acoustic type, such as speech, music, non-verbal human activity sounds, animal vocalizations, environmental sounds, noises, etc. The degree of detail in audio class analysis depends on the application. For example in radio broadcast signals segmentation the interest falls in the detection of the audio parts that contain speech, music, silence and noises.

In information processing frameworks dealing with audio data the role of the automatic segmentation sub-system is to divide the audio signal to the acoustic categories of interest in order to be further processed by the corresponding systems. Such post-processing systems can be speech recognizers, speaker recognizers, language recognizers, singer recognizers, song recognizers, sound event recognizers etc. The overall concept of such a framework and the role of audio segmentation within it are illustrated in Fig. 1. As can be seen, the initial audio stream is driven into an audio segmentation architecture, which is an open type of architecture in terms of its type (as it is described in the rest of the paper) can varies. The output stream, that holds an adjunct data series of segment level labels, is forward into a routing switch in which each audio type is driven into a suitable type of post-processing. For instance, in broadcast transmissions, speech parts can be driven into automatic speech recognizer for linguistic or speaker role processing while music parts can be driven into a sound effect collection library.

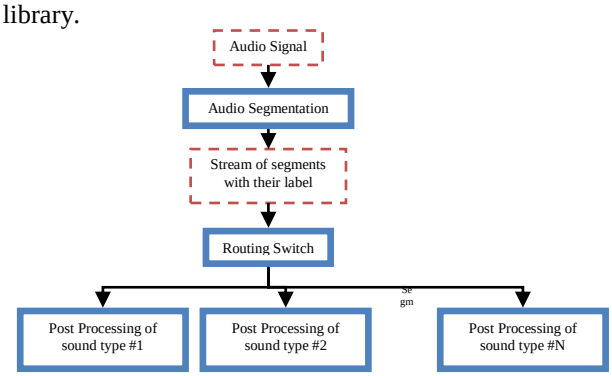

Fig. 1. Automatic audio segmentation and post-processing.

II. General Architecture of Audio Segmentation

The general architecture of the automatic audio segmentation scheme consists of four basic steps, namely (i) the feature extraction, (ii) the initial detection (optional stage), the (iii) the segmentation and (iv) the segment post-processing or smoothing (optional stage). In the first step the audio input is initially cut into overlapping frames of audio samples and for each frame a parametric feature vector is extracted. The computed sequence of feature vectors is forwarded to an initial detection module for a "garbage collection". This second step is optional and depending on the specifics of the application it is interpolated to the main structure for two reasons. The first reason is to remove the silence parts before the segmentation stage, instead of using a "silence" class. The second reason is to discard the parts of the signal that are out of interest (for example in the speaker segmentation task only the speech parts are needed for the segmentation stage). Typically the detection of silence and breathing noise is made from a simple energy-based detector (VAD) and the detection of music and noise is achieved using Gaussian mixture models (GMMs) [1;2;3;4;5;6]. The feature vector sequence is afterwards driven to the segmentation stage, where it is segmented to subsequences with common acoustics characteristic. For the segmentation stage two main approaches are followed, the distance-based techniques

and the model-based techniques. Optionally, after the segmentation stage the detected segments are post-processed in order to refine/smooth the automatic segmentation results. This stage corrects the errors related to detected segments with duration smaller than empirically defined thresholds. In Fig. 2 we illustrate the general architecture for automatic audio segmentation.

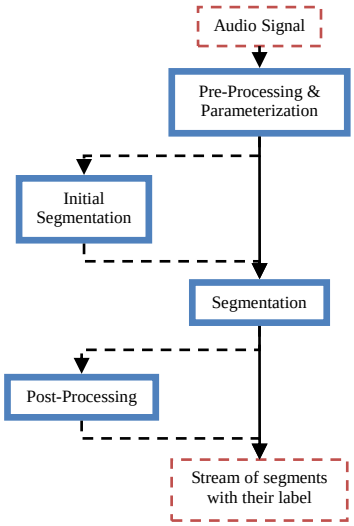

Fig. 2. General architecture of the automatic audio segmentation scheme.

III. Distance-Based Audio Segmentation

Distance-based audio segmentation algorithms estimate segments in the audio waveform, which correspond to specific acoustic categories, without labeling the segments with acoustic classes. In particular, the audio waveform is frame blocked and parameterized, and a distance metric is applied to adjacent feature vectors estimating the so called distance curve. The peaks of the distance curve correspond to frame boundaries where the distance is maximized, i.e. are positions with high acoustic change, and thus are considered as candidate audio segment boundaries. Post-processing over the candidate boundaries is applied in order to select which of the peaks on the distance curve will be considered as audio segment boundaries. The resulting audio sequence of segments will not be classified to a specific audio sound category. The categorization could be performed by forwarded the sequence into a classification scheme.

In order to improve the segmentation accuracy the distance is usually estimated over a time-shifting window of N frames, instead of measuring the distance between two adjacent frames. The length of the window typically varies from 1 sec to 5 sec with time-shift of approximately the 10% of the window length. The window length and the time-shift depend on the nature of the distance metric function as well as on the a priori knowledge of the minimum length of the corresponding audio segments of interest [7;8;9]. The use of frame windows results to a smooth estimation of the distance curve, which is less sensitive to potential local anomalies of the audio waveform. Except the use of overlapping windows for computing audio segments, divide and conquer strategy can be followed to segment the audio signal. In this approach the waveform is initially segmented into two non-overlapping parts, with respect to the maximization of their distance. In the same way, after this initial binary division, each of the segments is iteratively divided into two segments. The binary segmentation process continues until stopping criterions are met, e.g. the remaining pieces are too small. The segmentation points are decided according to the distance function computations [10].

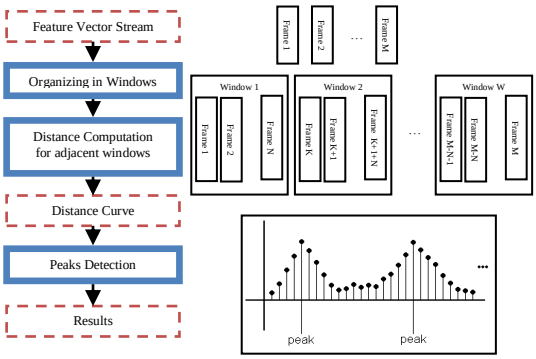

Fig. 3. Computation of the distance curve

A. Features for Distance-Based Segmentation

Several well known and extensively used audio descriptors have been used in the task of distance-based audio segmentation, for constructing the feature vectors. The most commonly used audio parameters are the Mel frequency cepstral coefficients (MFCC), and the zero crossing rate (ZCR) [7;9;8]. Other audio descriptors that have been used in the literature, especially when music information exists, are the dynamic, the timbre, the rhythm, the pitch and the tonality coefficients [11]. The linear prediction coefficients (LPC) and the linear spectral pairs (LSP) have also been used [12].

The size of the frame length and the frame step varies, depending on the nature and the specifics of the audio data. Typical values of the frame length are from 10msec to 25msec, with frame step between successive overlapping frames approximately 50% of the frame length [7;9;8].

B. Distance Metrics

The distance metrics are distance-based algorithms that perform an analysis over a stream of data to find that point which gives the optimum characteristic event. In case of segmentation the distance metric is a mathematical function that rolls over the audio stream comparing shifted windows, creating a curve that expresses the difference between sequential parts of the stream and finally finding the maximum points of this curve. Many functions have been proposed in the audio segmentation literature, mainly because they can be blind to the audio stream characteristics i.e. type of audio (recording conditions, number of acoustic sources, etc) or type of the upcoming audio classes (speech, music, etc). The most commonly used are:

- The Euclidean distance [7];

- The Bayesian information criterion, BIC [7;13;8;10;4;14;15;12;5;6];
- The Kullback Leibler KL2 distance [8;12;5];
- The Generalized Likelihood Ratio, GLR [7;9;4;15];
- The Hotelling T2 statistic [7;8];

The simplest distance metric for comparing two windows of feature vectors is the Euclidean distance. For two windows of audio data described as Gaussian models G1(μ1,Σ1) and G2(μ2,Σ2), the Euclidean distance metric is given by:

$$Eucl\ \ Dist = (\mu_1 - \mu_2)^{\mathrm{T}}(\mu_1 - \mu_2). \quad (1)$$

The Bayesian information criterion aims to find the best models that describe a set of data. From the two given windows of audio stream the algorithm computes three models representing the windows separately and jointly. From each model the formula extracts the likelihood and a complexity term that expresses the number of the model parameters. For two windows of audio data described as Gaussian models G1(μ1,Σ1) and G2(μ2,Σ2) and with their combined windows described as G(μ,Σ), the ΔBIC distance metric is given by:

$$\Delta BIC = BIC\{G_1\} + BIC\{G_2\} - BIC\{G\}. \quad (2)$$

$$BIC\{G\} = -\frac{N\log|\Sigma|}{2}... - \frac{\lambda(d + \frac{d(d+1)\log N}{2}}{2}.... - \frac{dN\log 2\pi}{2} - \frac{\mathrm{N}}{2} \quad (3)$$

$$\Delta BIC = \frac{N\log|\Sigma|}{2} - \frac{N_1\log|\Sigma_1|}{2}... - \frac{N_2\log|\Sigma_2|}{2} - \frac{\lambda d}{2}... - \frac{\lambda}{4}d(d+1)... (\log N_1 + \log N_2 - \log N) \quad (4)$$

where N,N1,N2 are the number of frames in the corresponding streams, d is the number of features of the feature vectors and λ is an experimentally factor.

The KL2 is a popular tool in the domain of statistical analysis for comparing probabilistic distributions. For two windows of audio data described as Gaussian models G1(μ1,Σ1) and G2(μ2,Σ2), the KL2 distance metric is given by:

$$KL2 = \frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}}(inv(\Sigma_1) + inv(\Sigma_2))... (\mu_1 - \mu_2) + \frac{1}{2}tr(inv(\Sigma_1)\Sigma_2 inv(\Sigma_2)\Sigma_1 - 2\mathrm{I}) \quad (5)$$

The Generalized Likelihood Ratio is a modification created by simplifying the Bayesian Information Criterion. Like BIC, it finds the difference between two windows of audio stream using the three Gaussian Models that describe these windows separately and jointly. For two windows of audio data described as Gaussian models G1(μ1,Σ1) and G2(μ2,Σ2), the GLR distance is given by:

$$GLR = w(2\log|\Sigma| - \log|\Sigma_1| - \log|\Sigma_2|). \quad (6)$$

where w is the window size.

Hotelling T2 statistic is another popular tool for comparing distributions. The main difference with KL2 is the assumption that the two comparing windows of audio stream have no difference on their covariances. For two windows of audio data described as Gaussian models G1(μ1,Σ1) and G2(μ2,Σ2), the Hotelling T2 distance metric is given by:

$$\mathrm{T2} = \frac{N_1 N_2}{N_1 + N_2}(\mu_1 - \mu_2)^{\mathrm{T}} inv(\Sigma)(\mu_1 - \mu_2). \quad (7)$$

where Σ equals Σ1 and Σ2 and N1,N2 are the number of frames in the corresponding streams.

IV. Model-Based Segmentation

In contrast to the distance-based segmentation approaches, where only segment boundaries are detected, in the model-based segmentation algorithms each audio frame is separately classified to a specific sound class, i.e. speech, music, noises, etc. In particular, each sound class of interest is represented by a model. Training data have been used to train one model for each sound class of interest, while universal models could also been used. During the operational phase, the unknown sequence of frames is compared against each of the models in order to provide decision (sound labeling) on frame-level. Post-processing algorithms can be applied to refine the frame labeling and after that adjacent audio frames labeled with the same sound class are merged to construct the detected segments. In the model-based approaches the segmentation process is performed together with the classification of the frames to a set of sound categories.

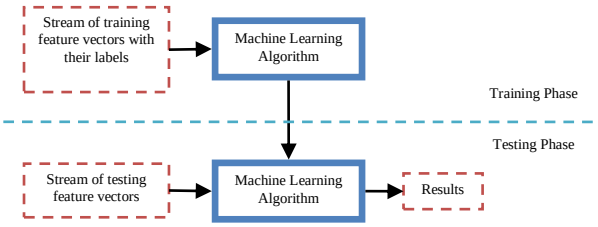

Fig. 4. Model-based audio segmentation

A. Features for Model-Based Segmentation

In the task of model-based audio segmentation, an extensively variety of well known audio descriptors have been used for constructing the feature vectors. The most commonly used audio parameters are the Mel frequency

cepstral coefficients (MFCC), the zero crossing rate (ZCR), spectrum based features such as energy, pitch, bandwidth, flux, centroid, roll-off, spread, flatness, projection, MPEG-7 features, etc [1;16;2], and other audio descriptors, especially when music information exists [11], such as the dynamic, the timbre, the rhythm, the pitch and the tonality coefficients.

The size of the frame length and the frame step varies, depending on the nature and the specifics of the audio data. Typical values of the frame length are from 10msec to 25msec, with frame step between successive overlapping frames approximately 50% of the frame length [12].

B. Machine Learning Algorithms

The machine learning algorithms are model-based algorithms that perform their analysis over a stream of data to classify each frame. In case of segmentation the machine learning algorithms are applied both into training and test phase. During training phase, the machine learning algorithms adjust their mathematical function to describe the a priori set of training data. For each one of the sounds of interest this probabilistic multidimensional mathematical function peaks expressing that each sound of interest corresponds to an (as possible as could be) unique combination of feature parameters. During the test phase, the sequence of feature vectors of the unknown audio signal are driven into this mathematical function, and each of the frame is examined for finding from the a priori close set of sound which is the most possible to be. Many machine learning algorithms have been proposed in audio segmentation literature mainly because the experimental conditions and the classes of interest are usually driven from the post processing systems. The most commonly used machine learning algorithms in audio segmentation are the GMM/HMM [16;17;6;1] and the SVM [1;2;14;11]. Other algorithms that have been used are the artificial neural networks [6], the boosting technology [11], the k-nearest neighbor [2], the decision trees [2;6] and the fuzzy logic [18].

V. Hybrid Architecture

While distance-based and model-based algorithms differ on handling vector streams they both tend to include all their procedures into a single stage. Hybrid architecture is a type of mixture of distance-based and model-based algorithms in a multistage architecture. The major roles of stages are either extracting from the feature stream results about the current acoustic phenomena or exploring sequential results for readjustment. All possible fine tuning on hybrid architecture has been proposed, examining the existence or the absence of classifier, the advantages or disadvantages for the upcoming stages while operating in frame or segment level as well as developing details about the sequence, the number and the repetition of stages. Hybrid technique is usually preferred in cases of a complex segmentation task, in tasks with post processing systems target into events of interest and in tasks in which segmentation is an essential component of the entire system. Analyzing the distance-based and the model-based algorithms in the domain of potential hybrid segmentation stages, a secondary operation for both distance-based and model-based algorithms creates another type of algorithm called resegmentation algorithm. Resegmentation algorithms differ not on the mathematical framework but on the usage of it.

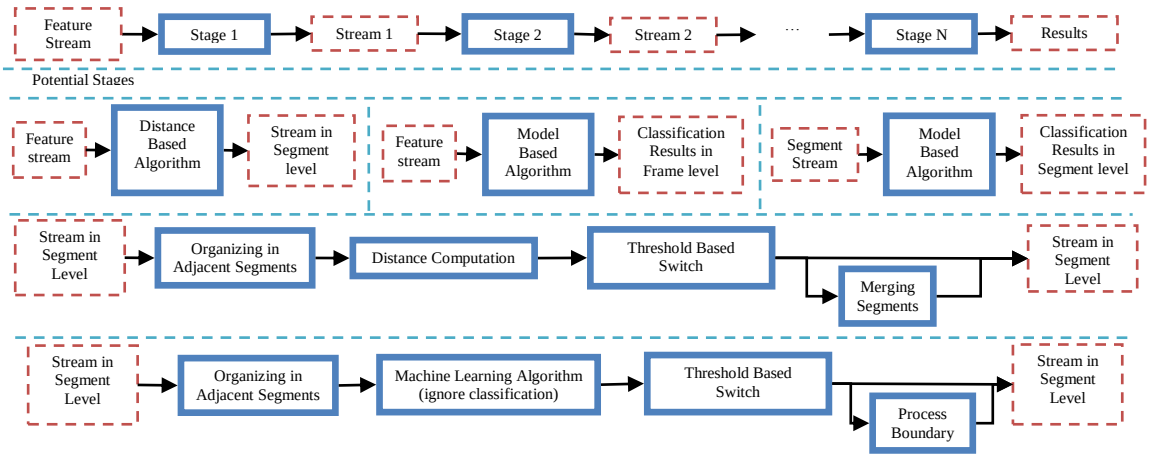

Fig. 5. Hybrid Architecture.

While distance-based and model-based algorithms tend to create results based on the incoming data, resegmentation algorithms tend to process the results from the precedent stage for error detection or for better tuning. Therefore, the most common components in hybrid segmentation are:

- Distance based algorithm for initial division of the audio stream into segments of the same acoustic content;
- Model based algorithm for initial classification of the audio stream in frame level;
- Model based algorithm for classification of the audio stream in segment level (the segments have been produced from a previous stage);

- Distance based resegmentation algorithm for error detection made by the precedent stage;
- Model based resegmentation algorithm for error detection made by the precedent stage without the usage of classification in either frame or segment level;

A. Features in Hybrid Architecture

In the task of hybrid segmentation, all the well known and previously described as extensively used audio descriptors in distance-based or model-based segmentation task have been used. Also the domain on the usage of the features has been examined. In particular, the extracted features could be used as an universal set for all the stages or the extracted features could be divided into sets, one set for each stage, and each feature could be assigned exclusively into one set or be assigned into more than one sets.

B. Distance Metric and Machine Learning Algorithms for the Resegmentation Stage

Considering the hybrid architecture as a potential combination between the described distance-based and model based algorithms, the mathematical framework is constructed based on the extensively variety that distance based and model based algorithm carry with them. Nevertheless, the two most common algorithms from the domains of distance metrics and machine learning algorithms for the resegmentation stage are the the BIC agglomerative clustering and the Viterbi resegmentation using GMM [4;12;5] correspondingly. The BIC agglomerative clustering computes the distance between two adjacent segments for potential merging. The Viterbi resegmentation using GMM creates models for two adjacent segments and reposes their common boundary.

VI. Post-Processing (Smoothing)

Post-processing (Smoothing) affects on the computed sequence of segmentation results, based on empirical rules existing on the database. A typical example relies on reclassifying segments when their length or their adjacency is inconsistent with the length and adjacency statistics of the database. A common post-processing (smoothing) algorithm [1][6] relies on reclassifying short duration segments adjacent to long duration segments when their acoustic content do not refer into sound of transactions from one content to another that normally should exist between them. Typical examples of such sounds are the environmental sounds of low energy and almost stable signal level [19] such as silence and background noise.

VII. Results

In the following table we present several proposed segmentation architectures as where presented in the literature. For each of the following architectures we present the database used in the experiments, the classes that described the events of interest and the accuracy results. The procedure followed in each architecture is organized based in a schematic form with the optional stages appear in their position in the schematic block. The architectures are organized based on the main type of algorithm: distance-based, model-based, hybrid. The distance metrics, the machine learning algorithms and the features used in these architectures are also appeared.

In particular, the list of the articles is organized in a term in which all articles that referring to distance base audio segmentation to precede and those that referring to hybrid architectures to be placed at the final rows of the table. The basic characteristics of each article are presented in four columns in which the database, the procedure, the detailed report of the audio classes and the details for the maximum accuracy are presented. The procedure column is organized in a form in which the type of segmentation algorithm is followed by details about the stage steps (in order that are placed in the architecture), the features used in those steps, the optional stages (where and when there are) and the distance metrics of the machine learning algorithms used in the current article.

Table 1. Proposed architectures in literature

Article	Database	Procedure	Classes	Maximum Accuracy
7	CCTV news corpus	Distance Based: Euclidean distance, BIC, T2, GLR Feature: 26 dimensions MFCC	-	for mEdist-BIC false alarm rate (FAR) 9.8% - missed detection rate (MDR) 10.7 %
8	DARPA	Distance Based: BIC, KL2, T2 Several features	Speech/non speech in various conditions (structures of audio, recording equipment, background noise)	96.9% total accuracy for speech/non speech classification; 97.8% frame accuracy NGSW Data classification (1960s)
9	2002 Rich Transcription Evaluation Database	Distance Based: GLR Feature: MFCC	-	FAR = 16.94 % MDR =19.47 %
10	MATBN Mandarin Chinese Broadcast News	Distance Based: BIC Divide and Conquer Feature: MFCC	Speaker change detection	17% EER
13	BBC news corpus 16KHz	Distance Based: BIC Feature: 13 MFCC	Speaker change detection	Before clustering FAR=9.8% MDR=11%

Article	Database	Procedure	Classes	Maximum Accuracy
20	MuscleFish	Silence detection Distance Based: Similarity Map Feature: MFCC, total spectrum power, bandwidth, brightness and pitch	16 classes including music, speech, animal sound and everyday sounds	91.9% for 2-phase Euclidean
1	TRECVID 2003 ABC World News Tonight, CNN headline news, Internet, music CDs	Silence Detection Model Based: SVM, HMM Feature: MPEG-7 features Post Processing: 3sec smoothing based on rules	Speech over environmental sound, speech over music, environmental sound, music, pure speech, silence	91.9% with speech – non speech classification; 89% pure speech – mixed speech classification; 85.2% speech over environmental sound – speech over music classification; 95.6% environmental sound – music classification
2	VRT Flemish Radio and Television Network extracted 16bit/48KHz converted 16bit/32KHz	Silence Detection based on rms Model Based: Bayes Network, k-NN, Decision Trees, SVM Feature: ZCR, RMS, pitch, spectral flux, low frequency RMS, high order zero crossing, sub-band correlation	Speech in various conditions, music (with or without singing), silence, various types of noise	above 96%
3	TDT-3 news broadcast	Silence Detection Model Based: MDL Feature: several features	Speech, music, speech and music, speech and noise, noise	88%
16	Broadcast material 16bit, 16KHz	Model Based: GMM Feature: MFCC, ZCR, Percentage of Low Energy Frames, spectral roll-off, spectral centroid, spectral flux	Speech (both male female speakers under studio or telephone quality) vs. music	98,3% for MFCCV mix as features
4	Different French Broadcast 16KHz	GMM classifier based on MFCC for speech discrimination Hybrid (A) Distance Based: BIC/GLR (B) Distance Based / Model Based without classification: BIC segmentation, Viterbi resegmentation Anchor duration can't be less than 4sec	Speaker detection	89.2% the F-measure
5	French TV stations in Quebec	Silence Detection using an energy based voice activity detection and GMM removes noise, music, music & speech Hybrid (A) Distance Based: KL2 (B) Model Based without classification: Viterbi resegmentation (C) Distance Based: BIC agglomerative clustering (D) Model Based without classification: Viterbi resegmentation (E) Distance Based: BIC agglomerative clustering (F) Model Based without classification: Gender Labeling (G) Model Based without classification: Agglomerative SID clustering (H) Model Based without classification: Viterbi resegmentation (I) Model Based with classification: GMM Feature: MFCC	Speaker Diarization	The best DER is 14.5% on the test set

Article	Database	Procedure	Classes	Maximum Accuracy
11	MIRAX 2009 16bit 44.1KHz stereo music database and separation in 45 types of music	Hybrid (A) Distance Based: Self Similarity Matrix (B) Model Based with classification: SVM, AdaBoost Feature: (A) MFCC (B) 174 dimensional feature vector from MIRToolbox 1.1 [21]	45 types of music	0.9201 ±0.0003 tag accuracy
12	Eurosport TV program	Hybrid (A) Distance Based: KL2 (B) Distance Based: BIC agglomerative clustering (C) Model Based without classification: Viterbi resegmentation (D) Model Based with classification: GMM Feature: (A) LSP (B) LSP (C) LSP (D) 39 MFCC	Speech, non speech & speech, music, background	average accuracy 87.3%
14	ESTER campaign	Hybrid (A) Model based: SVM (B) Distance Based / Model Based without classification: BIC, One-class SVM, probabilistic distance Feature: 500 features	Speech, music, speech and music	91.7% the F-measure with 70 features selected by the IRMFSP algorithm
15	ESTER2; 3 classes: Anchorman, Journalist and Other	Hybrid (A) Distance based: BIC/GLR (B) Model based with classification: GMM, k-NN, SVM Feature: (B) 34 features based on temporal time measurements on signal energy and on pitch	Speaker Detection between Anchorman, Journalist and Other	78.66%
17	Audio data from Sports 16KHz	Hybrid (A) Distance Based: Bhattachayya Distance splitting and BIC merging (B) Model Based with classification: GMM Feature (A) MFCC (B) 24 features include plp, short time energy, spectrum flux, sub-band energy distribution, brightness, bandwidth	commentator's speech in the audio stream	93.56% the F value for the average
6	3/24 Catalan TV channel, recorded by TALP Research Cender from UPC, annotated by Verbio Technologies, 16bit 16KHz	Different types of architectures	Sounds (speech, music, noise, none) over various conditions (studio, telephone, outside, none), Speaker and Speaking mode, Speech Transcription and Acoustic Events	30.22 error rate for the first system
		Model Based: HMM Feature: MFCC, energy, spectral entropy, CHROMA		
		Hybrid (A) Distance Based: BIC (B) Model Based with classification: GMM Feature: MFCC and energy		
		Model Based: HMM Feature: MFCC Post Processing		
		Initial Silence Detection Model Based: HMM Feature: 16 frequency-filtered log-bank energies		
		Model Based: HMM Feature: PLP, local energy		
		Initial Silence and Music Detection Model Based: HMM, MLP Feature: MFCC, energy, eight perceptual coefficients (zero crossing rate, spectral centroid, spectral roll-off, etc)		
		Initial Silence Detection Hybrid (A) Distance Based: BIC Segmentation (B) Model Based with classification: GMM and Decision Trees Feature: MFCC		
		Model Based: GMM Feature: MFCC and energy Post Processing Smoothing		

REFERENCES

[1] E. Dogan, M. Sert, A. Yazici (2009). "Content-Based Classification and Segmentation of Mixed-Type Audio by Using MPEG-7 Features", 2009 First International Conference on Advances in Multimedia MMEDIA '09, on pages(s) 152-157

[2] Y. Patsis, W. Verhelst (2008). "A Speech/ Music/Silence /Garbage Classifier for Searching and Indexing Broadcast News Material", 2008 19th International Workshop on Database and Expert Systems Application DEXA '08, on page(s) 585-589

[3] C.-H. Wu, C.-H. Hsieh (2006). "Multiple Change Point Audio Segmentation and Classification using an MDL-based Gaussian Model", IEEE Transactions on Audio, Speech and Language Processing, Issue Date March 2006, volume 14, Issue 2, on page(s) 647-657

[4] C. Delphine 2010. "Model-Free Anchor Speaker Turn Detection for Automatic Chapter Generation in Broadcast News", 2010 IEEE International Conference on Acoustics Speech and Signal Processing ICASSP, on page(s) 4966-4969

[5] V. Gupta, G. Boulianne, P. Kenny, P. Ouellet, P. Dumouchel (2008). "Speaker Diarization of French Broadcast News", 2008 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008, on page(s) 4365-4368

[6] T. Butko, C. Nadeu (2011). "Audio Segmentation of Broadcast News in the Albayzin-2010 Evaluation: Overview, Results and Discussion", EURASIP Journal on Audio, Speech and Music Processing 2011, volume 2011 issue 1

[7] H. Xue, H. Li, C. Gao, Z. Shi (2010). "Computationally Efficient Audio Segmentation through a Multi-Stage BIC Approach", 2010 3rd International Congress on Image ad Signal Processing CISP, volume 8, on page(s) 3774-3777

[8] R. Huang, J. H.L. Hansen (2006). "Advances in Unsupervised Audio Classification and Segmentation for Broadcast News and NGSW Corpora", IEEE Transactions on Audio, Speech and Language Processing, issue Date May 2006, Volume 14, Issue 3, on page(s) 907-919

[9] D. Wang, R. Vogt, M. Mason, S. Sridharan (2008). "Automatic Audio Segmentation Using the Generalized Likelihood Ratio", 2008 2nd International Conference on Signal Processing and Communication Systems ICSPCS 2008, on page(s) 1-5

[10] S.-S. Cheng, H.-M. Wang, H.-C. Fu (2008). "BIC-Based Audio Segmentation by Divide and Conquer", 2008 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008, on page(s) 4841-4844

[11] H.-Y. Lo, J.-C. Wang, H.-M. Wang (2010). "Homogeneous Segmentation and Classifier Ensemble for Audio Tag Annotation and Retrieval", 2010 IEEE International Conference on Multimedia and Expo ICME, on page(s) 304-309

[12] J. Huang, Y. Dong, J. Liu, C. Dong, H. Wang (2009). "Sports Audio Segmentation and Classification", 2009 IEEE International Conference on Network Infrastructure and Digital Content IC-NIDC 2009, on page(s) 379-383

[13] S. Harsha Yella, V. Varma, K. Prahallad (2010). "Significance of Anchor Speaker Segments for Constructing Extractive Audio Summaries of Broadcast News", 2010 IEEE Spoken Language Technology Workshop SLT, on page(s) 13-18

[14] G. Richard, M. Ramona, S. Essid (2007). "Combined Supervised and Unsupervised Approaches for Automatic Segmentation of Radiophonic Audio Streams", 2007 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007, on page(s) II-461 – II-464

[15] B. Bigot, I. Ferrane, J. Pinquier (2010). "Exploiting Speaker Segmentations for Automatic Role Detection". An Application to Broadcast News Documents, 2010 International Workshop on Content-Based Multimedia Indexing CBMI, on page(s) 1-6

[16] M. Kos, M. Grasic, D. Vlaj, Z. Kacic (2009). "On-line Speech/Music Segmentation for Broadcast News Domain", 2009 16th International Conference on Systems, Signal and Image Processing IWSSIP 2009, on page(s) 1-4

[17] J. Zhang, B. Jiang, L. Lu, Q. Zhao (2010). "Audio Segmentation System for Sport Games", 2010 International Conference on Electrical and Control Engineering ICECE, on page(s) 505-508

[18] M. Liu, C. Wan, L. Wang (2002). "Content-Based Audio Classification and Retrieval using a Fuzzy Logic System: Towards Multimedia Search Engines", Soft Computing 6 (2002) 357-364

[19] T. Perperis, T. Giannakopoulos, A. Makris, D. I. Kosmopoulos, S. Tsekeridou, S. J. Perantonis, S. Theodoridis (2010). "Mutlimodal and Ontology-based Fusion Approaches of Audio and Visual Processing for Violence Detection in Movies", Expert Systems with Applications Volume 38, Issue 11, October 2011, Pages 14102–14116

[20] J. X. Zhang, J. Whalley, S. Brooks (2009). "A Two Phase Method for General Audio Segmentation", 2009 IEEE International Conference on Multimedia and Expo ICME 2009, on page(s) 626-629

[21] http://users.jyu.fi/~lartillo/mirtoolbox/

Authors' Profiles

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 University of Patras, Greece.

Afterwards, he was accepted as a PhD candidate at the Department of Electrical and Computer Engineering of the University of Patras. During his research activity he participated in researched programs and published articles in the domain of audio processing. His current research interests also include the domain of audio segmentation.

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 from the Department of Electrical and Computer Engineering of the University of Patras, Greece. Currently he is post-doctoral researcher at the University of Patras and non-tenured Assistant Professor at the Technological Educational Institute of Western Greece. 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.

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.

How to cite this paper: Theodoros Theodorou, Iosif Mporas, Nikos Fakotakis,"An Overview of Automatic Audio Segmentation", International Journal of Information Technology and Computer Science(IJITCS), vol.6, no.11, pp.1-9, 2014. DOI: 10.5815/ijitcs.2014.11.01

