ΤΕΧΝΟΛΟΓΙΑ ΗΧΟΥ ΚΑΙ ΕΙΚΟΝΑΣ

ΕΡΓΑΣΙΑ 1

Η ΟΜΑΔΑ ΜΑΣ

Αποστόλης

Φρανκ

Χριστίνα

HOW IT'S DONE

SPEECH / MUSIC CLASSIFIER

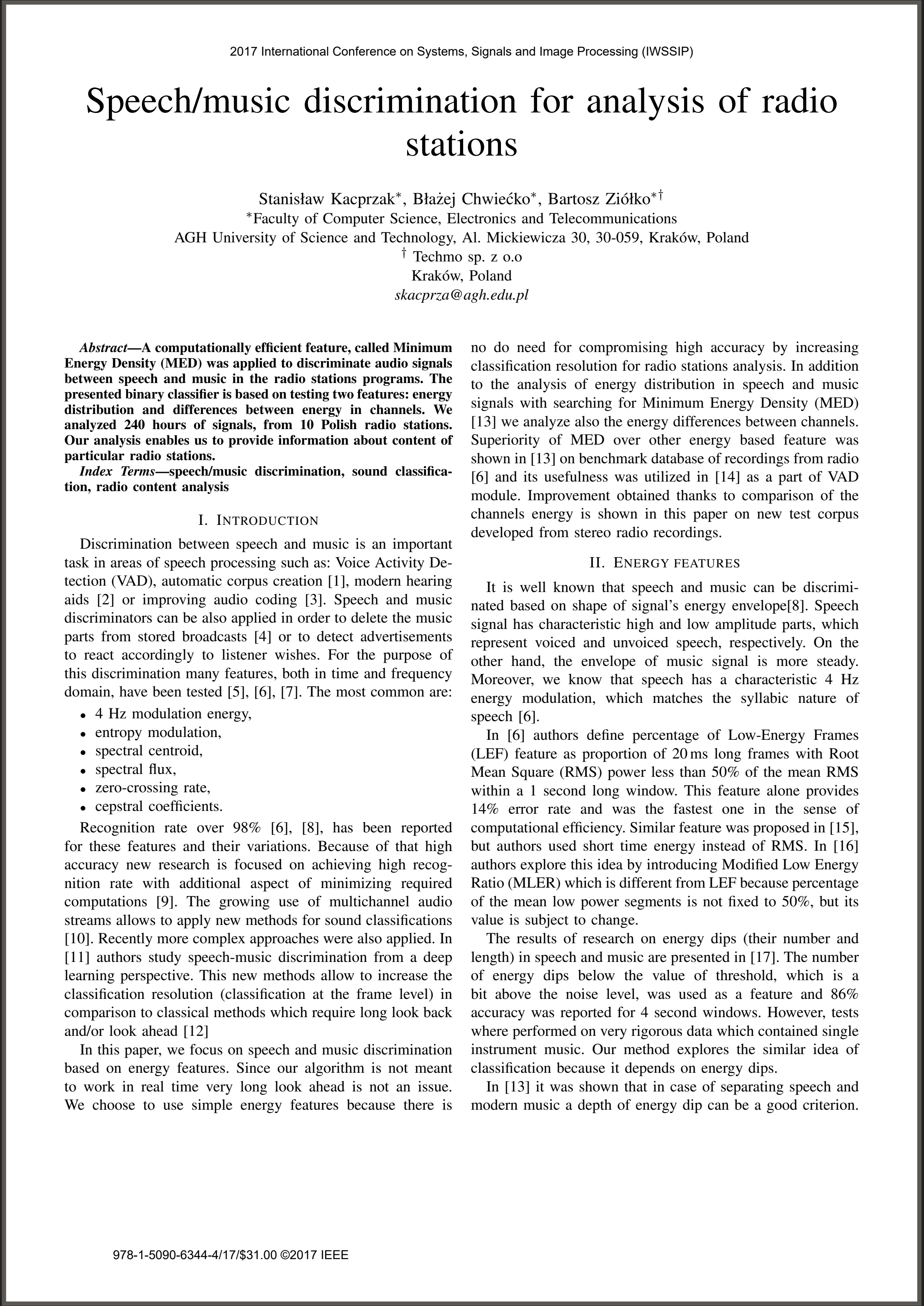
HOW IT'S DONE

SPEECH/MUSIC DISCRIMINATION FOR ANALYSIS OF RADIO STATIONS

STANISŁAW KACPRZAK, BŁAZEJ CHWIECKO, BARTOSZ ZIÓŁKO

- Εφαρμογή μη πραγματικού χρόνου
- Ενεργιακά features
  - Minimum Energy Density
  - Διαφορά ενέργειας μεταξύ καναλιών
- 100% επιτυχία*

*με εξαίρεση τις διαφημίσεις...



2017 International Conference on Systems, Signals and Image Processing (IWSSIP)

Speech/music discrimination for analysis of radio stations

Stanisław Kacprzak*, Błażej Chwiećko*, Bartosz Ziółko*†
*Faculty of Computer Science, Electronics and Telecommunications
AGH University of Science and Technology, Al. Mickiewicza 30, 30-059, Kraków, Poland
† Techmo sp. z o.o
Kraków, Poland
skacprza@agh.edu.pl

***Abstract*—A computationally efficient feature, called Minimum Energy Density (MED) was applied to discriminate audio signals between speech and music in the radio stations programs. The presented binary classifier is based on testing two features: energy distribution and differences between energy in channels. We analyzed 240 hours of signals, from 10 Polish radio stations. Our analysis enables us to provide information about content of particular radio stations.**

***Index Terms*—speech/music discrimination, sound classification, radio content analysis**

I. Introduction

Discrimination between speech and music is an important task in areas of speech processing such as: Voice Activity Detection (VAD), automatic corpus creation [1], modern hearing aids [2] or improving audio coding [3]. Speech and music discriminators can be also applied in order to delete the music parts from stored broadcasts [4] or to detect advertisements to react accordingly to listener wishes. For the purpose of this discrimination many features, both in time and frequency domain, have been tested [5], [6], [7]. The most common are:

- 4 Hz modulation energy,
- entropy modulation,
- spectral centroid,
- spectral flux,
- zero-crossing rate,
- cepstral coefficients.

Recognition rate over 98% [6], [8], has been reported for these features and their variations. Because of that high accuracy new research is focused on achieving high recognition rate with additional aspect of minimizing required computations [9]. The growing use of multichannel audio streams allows to apply new methods for sound classifications [10]. Recently more complex approaches were also applied. In [11] authors study speech-music discrimination from a deep learning perspective. This new methods allow to increase the classification resolution (classification at the frame level) in comparison to classical methods which require long look back and/or look ahead [12]

In this paper, we focus on speech and music discrimination based on energy features. Since our algorithm is not meant to work in real time very long look ahead is not an issue. We choose to use simple energy features because there is no do need for compromising high accuracy by increasing classification resolution for radio stations analysis. In addition to the analysis of energy distribution in speech and music signals with searching for Minimum Energy Density (MED) [13] we analyze also the energy differences between channels. Superiority of MED over other energy based feature was shown in [13] on benchmark database of recordings from radio [6] and its usefulness was utilized in [14] as a part of VAD module. Improvement obtained thanks to comparison of the channels energy is shown in this paper on new test corpus developed from stereo radio recordings.

II. Energy features

It is well known that speech and music can be discriminated based on shape of signal's energy envelope[8]. Speech signal has characteristic high and low amplitude parts, which represent voiced and unvoiced speech, respectively. On the other hand, the envelope of music signal is more steady. Moreover, we know that speech has a characteristic 4 Hz energy modulation, which matches the syllabic nature of speech [6].

In [6] authors define percentage of Low-Energy Frames (LEF) feature as proportion of 20 ms long frames with Root Mean Square (RMS) power less than 50% of the mean RMS within a 1 second long window. This feature alone provides 14% error rate and was the fastest one in the sense of computational efficiency. Similar feature was proposed in [15], but authors used short time energy instead of RMS. In [16] authors explore this idea by introducing Modified Low Energy Ratio (MLER) which is different from LEF because percentage of the mean low power segments is not fixed to 50%, but its value is subject to change.

The results of research on energy dips (their number and length) in speech and music are presented in [17]. The number of energy dips below the value of threshold, which is a bit above the noise level, was used as a feature and 86% accuracy was reported for 4 second windows. However, tests where performed on very rigorous data which contained single instrument music. Our method explores the similar idea of classification because it depends on energy dips.

In [13] it was shown that in case of separating speech and modern music a depth of energy dip can be a good criterion.

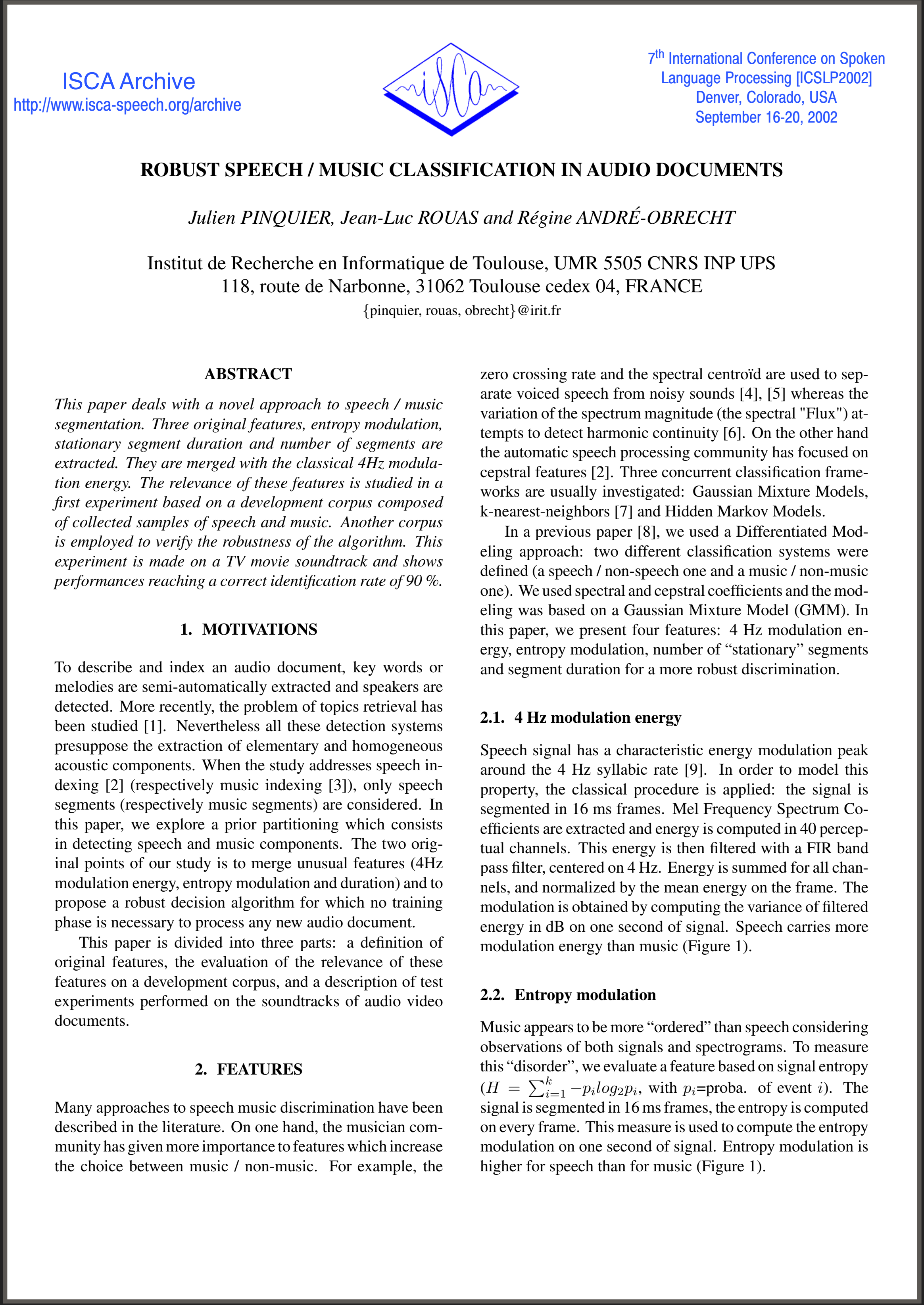
978-1-5090-6344-4/17/$31.00 ©2017 IEEE

HOW IT'S DONE

ROBUST SPEECH / MUSIC CLASSIFICATION IN AUDIO DOCUMENTS

JULIEN PINQUIER, JEAN-LUC ROUAS AND REGINE ANDRE-OBRECHT

- Features:
  - Διαμόρφωση ενέργειας στα 4Hz του σήματος
  - Διαμόρφωση εντροπίας του σήματος
  - Αριθμό των στατικών τμημάτων
  - Διάρκεια των τμημάτων
- Με ιεραρχικό αλγόριθμο
  - επιτυχία 90.1%



ISCA Archive
http://www.isca-speech.org/archive

7th International Conference on Spoken
Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

ROBUST SPEECH / MUSIC CLASSIFICATION IN AUDIO DOCUMENTS

Julien PINQUIER, Jean-Luc ROUAS and Régine ANDRÉ-OBRECHT

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INP UPS
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
{pinquier, rouas, obrecht}@irit.fr

ABSTRACT

This paper deals with a novel approach to speech / music segmentation. Three original features, entropy modulation, stationary segment duration and number of segments are extracted. They are merged with the classical 4Hz modulation energy. The relevance of these features is studied in a first experiment based on a development corpus composed of collected samples of speech and music. Another corpus is employed to verify the robustness of the algorithm. This experiment is made on a TV movie soundtrack and shows performances reaching a correct identification rate of 90 %.

1. MOTIVATIONS

To describe and index an audio document, key words or melodies are semi-automatically extracted and speakers are detected. More recently, the problem of topics retrieval has been studied [1]. Nevertheless all these detection systems presuppose the extraction of elementary and homogeneous acoustic components. When the study addresses speech indexing [2] (respectively music indexing [3]), only speech segments (respectively music segments) are considered. In this paper, we explore a prior partitioning which consists in detecting speech and music components. The two original points of our study is to merge unusual features (4Hz modulation energy, entropy modulation and duration) and to propose a robust decision algorithm for which no training phase is necessary to process any new audio document.

This paper is divided into three parts: a definition of original features, the evaluation of the relevance of these features on a development corpus, and a description of test experiments performed on the soundtracks of audio video documents.

2. FEATURES

Many approaches to speech music discrimination have been described in the literature. On one hand, the musician community has given more importance to features which increase the choice between music / non-music. For example, the zero crossing rate and the spectral centroïd are used to separate voiced speech from noisy sounds [4], [5] whereas the variation of the spectrum magnitude (the spectral "Flux") attempts to detect harmonic continuity [6]. On the other hand the automatic speech processing community has focused on cepstral features [2]. Three concurrent classification frameworks are usually investigated: Gaussian Mixture Models, k-nearest-neighbors [7] and Hidden Markov Models.

In a previous paper [8], we used a Differentiated Modeling approach: two different classification systems were defined (a speech / non-speech one and a music / non-music one). We used spectral and cepstral coefficients and the modeling was based on a Gaussian Mixture Model (GMM). In this paper, we present four features: 4 Hz modulation energy, entropy modulation, number of "stationary" segments and segment duration for a more robust discrimination.

2.1. 4 Hz modulation energy

Speech signal has a characteristic energy modulation peak around the 4 Hz syllabic rate [9]. In order to model this property, the classical procedure is applied: the signal is segmented in 16 ms frames. Mel Frequency Spectrum Coefficients are extracted and energy is computed in 40 perceptual channels. This energy is then filtered with a FIR band pass filter, centered on 4 Hz. Energy is summed for all channels, and normalized by the mean energy on the frame. The modulation is obtained by computing the variance of filtered energy in dB on one second of signal. Speech carries more modulation energy than music (Figure 1).

2.2. Entropy modulation

Music appears to be more "ordered" than speech considering observations of both signals and spectrograms. To measure this "disorder", we evaluate a feature based on signal entropy ($H = \sum_{i=1}^{k} -p_i log_2 p_i$, with p_i=proba. of event i). The signal is segmented in 16 ms frames, the entropy is computed on every frame. This measure is used to compute the entropy modulation on one second of signal. Entropy modulation is higher for speech than for music (Figure 1).

HOW IT'S DONE

SPEECH CLASSIFICATION BASED ON CUCKOO ALGORITHM AND SUPPORT VECTOR MACHINES

WENLEI SHI, XINHAI FAN

- Μόνο Mel frequency cepstrum coefficient (MFCC) ως feature
- Συνδιασμός SVM με τον Cuckoo algorithm
- Κατά μέσο όρο επιτυχία
  - BP 89.08%
  - SVM 90.12%
  - CS-SVM 92.75%

Speech classification based on cuckoo algorithm and support vector machines

Wenlei Shi
Department of Mechanical Engineering
Academy of Armored Forces Engineering
Beijing, China
e-mail: leishen03@126.com

Xinhai Fan
Department of Mechanical Engineering
Academy of Armored Forces Engineering
Beijing, China
e-mail: zgyfxh1210@sina.com

Abstract—**Speech classification is an important part of speech signal processing. It is significant to classify speech accurately and quickly in speech coding and speech synthesis. Because of the diversity and uncertainty of the speech signals, the traditional classification method is slow and not so accurate in the large-scale application of real speech classification. In order to improve the accuracy and precision of speech classification, a speech classification method based on support vector machine optimized by cuckoo algorithm(CS-SVM) is proposed. Firstly, choose four types of music: folk songs, Guzheng, rock and pop. And adopt the cepstral coefficient to extract speech feature, then use the support vector machine optimized by the cuckoo algorithm to train the characteristic signals, and establish the optimal classifier model, and finally classify the tested speech. The results of the simulation experiment show that the support vector machine based on the cuckoo algorithm(CS-SVM) is better than the traditional SVM and BP neural network in speech recognition.**

Keywords-support vector machine;cuckoo algorithm; speech classification; classification model; BP neural network

I. INTRODUCTION

Speech classification is a subject that divides speech into voiced sounds and unvoiced sounds, and the unvoiced sounds are divided into vowels and voiced consonants according to the glottal excitation form. It is a vital part of speech signal processing. The traditional way is to extract some feature parameters, then judge them by linear processing and predefined threshold value, and the threshold value is usually determined by personal experience. Although it is simple and easy to achieve, there is no guarantee that the results are accurate and reliable. With the rapid development of artificial intelligence and machine learning in recent years, it has provided a great foundation for the research of independent learning ability and automatic audio classification [1-3]. Support vector machine (SVM) is a new machine learning method based on statistical learning theory, which can better solve the practical problems such as small samples, nonlinear etc. It has become a hotspot in the research of intelligent technology. Besides, at present, it has been widely used in condition assessment, fault diagnosis, pattern recognition, chemical modeling, and many other fields.

II. SUPPORT VECTOR MACHINE

The original SVM algorithm [4] was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. In 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik suggested a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplanes [5]. The current standard SVM was proposed by Corinna Cortes and Vapnik in 1993 and published in 1995 [6].

SVM originated from linear separable optimal classification hyperplane, which essence is to find out the support vector used to construct the optimal classification hyperplane in the training sample. And it can be attributed to the solution of a quadratic optimization problem in mathematics. For the nonlinear classification, the general idea of SVM is to use a nonlinear transformation to map the input data into a high dimensional vector space firstly, then construct the optimal classification hyperplane in the feature space to develop the linear classification, finally after mapping it back to the original space, it becomes a nonlinear classification in the impute space.

Set linear separable sample $(x_i，y_i)$,$i = 1,2,3 \dots, l$,$x_i \in R_n$,$y_i \in [-l，l]$, l is the total number of training samples，n is the dimension of sample space, $\mathrm{y_i}$ is the category of sample. As shown in the following figure, circles and squares represent two different types of samples respectively, H represents the hyperplane that separates the two kinds of samples correctly, its direction is represented by the the normal vector of the hyperplane. $\mathrm{H_1}$、$\mathrm{H_2}$ represent the plane that is parallel to hyperplane and closest to hyperplane Hin two types of samples, and the distance is called classification interval. The optimal classification hyperplane refers to the hyperplane which not only separates the two types of samples correctly, makes the error of the model training zero, but also makes the classification interval of the two types arrive to the maximum. The linear discriminant function in the dimension space d is $g(x) = w * x + b$. The hyperplane equation is： $w * x + b = 0$. In the formula, w is the parameter variable, $\mathrm{w} \in \mathrm{R_n}$， namely the normal line of the hyperplane, $\mathrm{b} \in \mathrm{R}$ is the threshold value of the classification，$\mathrm{w} * \mathrm{x}$ is the inner product operation of the vector.

When $|g(x)| \geq 1$ ， the closest sample to the classification plan is$|g(x)| = 1$， the classification interval is $2/\|w\|$， if the interval is required to be maximum, that is, require$\|w\|$ or $\|w\|^2$ to be minimum and if the classification plane is accurate for all sample classification, it means to meet the requirement:

$$y_i[(w * x_i + b)] \geq 1，\quad i = 1,2,3, \dots, l \quad (1)$$

The optimal classification hyperplane problem can be transformed into the following optimization problem with constrained conditions:

978-1-5386-2030-4/17/$31.00 ©2017 IEEE

HOW IT'S DONE

SPEECH/MUSIC DISCRIMINATION USING HYBRID-BASED FEATURE EXTRACTION FOR AUDIO DATA INDEXING

KUN-CHING, WANG, MEMBER, IEEE, YUNG-MING, YANG AND YING-RU, YANG

- Features:
  - MFCC ~90%
  - - ZCR (zero crossing rate) ~90%
  - - SC (Spectral Centroid) ~70%
  - - SR (Spectral Rolloff) ~83%
  - - SF (Specral Flux) ~90%

- Με χρήση SVM
  - επιτυχία 95.68%
- Αλλά είναι ανισόρροπη
  - Speech: 98.25%
  - Music: 93.1%

2017 International Conference on System Science and Engineering (ICSSE)

Speech/Music Discrimination using Hybrid-Based Feature Extraction for Audio Data Indexing

Kun-Ching, Wang, *Member, IEEE,* Yung-Ming, Yang and Ying-Ru, Yang

***Abstract*—In this paper, we present a speech/music discrimination (SMD) using hybrid manner of feature extraction to discriminate the noisy audio signal into speech and music. The hybrid-based SMD performs the combination of 1D signal processing and 2D image processing to extract multiple features. In general, the noisy audio segment can be regarded as music, speech or noise (silence). The proposed hybrid-based SMD approach has been successfully applied into audio data indexing to classify the noisy audio signal into speech, music and noise. The approach includes three main stages: pre-processing/voice activity detection (VAD), speech/music discrimination (SMD) and rule-based post-processing. Both of pre-processing and VAD are regarded as the first stage for discriminating audio recording stream into noise-only segments and noisy audio segments. Next, the hybrid-based SMD is regarded as the second stage to classify noisy audio segments into speech segments and music segments. In third stage, a rule-based post-filtering method will be applied in order to improve the discrimination accuracy and to reflect the continuity of audio data in time. Experimental results will show that the proposed hybrid-based SMD approach can successfully apply into the audio data indexing. The overall system accuracy will be evaluated on radio recordings from various sources. Performance results can provide significant classification for the envisaged tasks compared to existing methods is given.**

***Keywords*— spectrogram image, speech/music classification, wavelet packet, support vector machine, hybrid-based feature extraction**

I. Introduction

Speech/Music discrimination (SMD) is an important task in multimedia indexing. It is usually the basic step for further processing on audio data. In previous work, feature analysis in SMD has been utilized widely to represent audio data. These feature analyses can be categorized into three major categories. One is based on temporal features, such as zero-crossing rate (ZCR) and short time energy [1-2]. Another category of approaches adopts frequency-domain features (such as mel frequency cepstral coefficients (MFCCs) popularly used in automatic speech recognition [3] and spectral features [4]). Spectral analysis based features were used to represent audio data likes Spectral Centroid (SC), Spectral Rolloff (SR) and Specral Flux (SF) [5-6]. The third category of approaches are based on mixed feature exploiting both time-domain and frequency-domain features (such as 4 Hz modulation energy [6], low frequency modulation amplitude features [7], percentage of low energy frames [5]).

The hybrid-based SMD performs the combination of 1D signal processing and 2D image processing to extract multiple features. In general, the noisy audio segment can be regarded as music, speech or noise (silence). The proposed hybrid-based SMD approach has been successfully applied into audio data indexing to classify the noisy audio signal into speech, music and noise. The approach includes three main stages: pre-processing, speech/music discrimination (SMD) and rule-based post-processing. Both of pre-processing and voice activity detection are regarded as the first stage for discriminating audio recording stream into noise-only segments and noisy audio segments. Next, the hybrid-based SMD is regarded as the second stage to classify noisy audio segments into speech segments and music segments. In third stage, a rule-based post-filtering method will be applied in order to improve the discrimination accuracy and to reflect the continuity of audio data in time.

For noisy segmented audio input, the features are extracted from the 1D perceptual wavelet packet transform (PWPT) and zoning spectrogram image, respectively. The hybrid features include 1D subband energy information, 2D texture information and spectral peak tracking information. For the feature extraction of 1D subband energy information, we use 1D-PWPT to get the 24 critical subbands. Through the useful subband selection [10], the correct energy information is used to discriminate the difference between speech and music. In the feature extraction of 2D texture information, gray-scale spectrogram is first generated. Zoning the range from 0 kHz to 4 kHz, the local information is enough to character speech and music, respectively [11]. Using Laws' mask through 2D-PWPT, we can get the 2D texture information [12]. In addition, the spectral peak tracking information will be computed after binary processing for gray-scale spectrogram [13]. Next, the three hybrid-features are inputted into a classifier of support vector machine (SVM).

Figure 1 shows a diagram of the proposed audio analysis approach to index the noisy audio signal as speech, music and noise. The approach includes three main stages: pre-processing, speech/music discrimination (SMD) and rule-based post-processing. Both of pre-processing and voice activity detection are regarded as the first stage for discriminating audio recording stream into noise-only segments and noisy audio segments. Next, the hybrid-based SMD, which performs the combination of 1D signal processing and 2D image processing to extract multiple features including subband energy information, textural information, spectral peak tracking, is regarded as the second stage to classify noisy audio segments into speech segments and music segments. In third stage, a rule-based post-filtering method will be applied in order to improve the discrimination accuracy and to reflect the continuity of audio data in time [14].

The Authors are with the Department of Information Technology & Communication, Shih Chien University, 200 University Road, Neimen, Kaohsiung 84550, Taiwan, R.O.C. (corresponding author to provide phone: + 886 – 076678888-5723; fax: + 886 – 076678888-4332; e-mail: kunching@mail.kh.usc.edu.tw, a0434344@g2.usc.edu.tw and Oscarking1332@gmail.com).

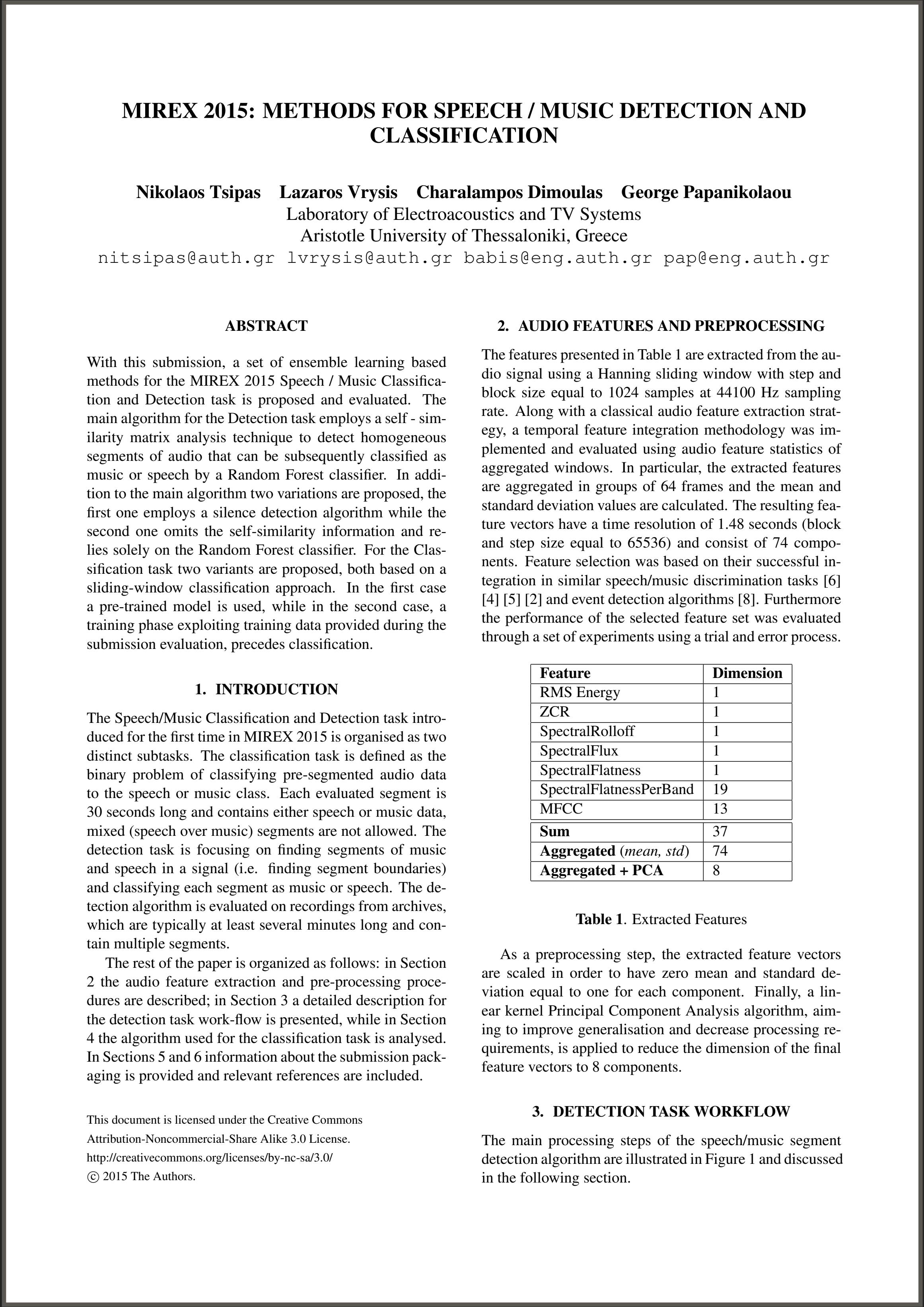
978-1-5386-3422-6/17/$31.00 ©2017 IEEE

HOW IT'S DONE

MIREX 2015: METHODS FOR SPEECH / MUSIC DETECTION AND CLASSIFICATION

NIKOLAOS TSIPAS LAZAROS VRYSIS CHARALAMPOS DIMOULAS GEORGE PAPANIKOLAOU

- Εντοπισμό δειγμάτων με τον αλγόριθμο Random Forest
- Features:
  - RMS ενέργεια
  - ZCR
  - Spectral rolloff
  - Spectral flux
  - Spectral flatness
  - Spectral flatness per Band
  - MFCCs
- Βελτιστοποίηση διανυσμάτων κριτηρίων



MIREX 2015: METHODS FOR SPEECH / MUSIC DETECTION AND CLASSIFICATION

Nikolaos Tsipas Lazaros Vrysis Charalampos Dimoulas George Papanikolaou
Laboratory of Electroacoustics and TV Systems
Aristotle University of Thessaloniki, Greece
nitsipas@auth.gr lvrysis@auth.gr babis@eng.auth.gr pap@eng.auth.gr

ABSTRACT

With this submission, a set of ensemble learning based methods for the MIREX 2015 Speech / Music Classification and Detection task is proposed and evaluated. The main algorithm for the Detection task employs a self - similarity matrix analysis technique to detect homogeneous segments of audio that can be subsequently classified as music or speech by a Random Forest classifier. In addition to the main algorithm two variations are proposed, the first one employs a silence detection algorithm while the second one omits the self-similarity information and relies solely on the Random Forest classifier. For the Classification task two variants are proposed, both based on a sliding-window classification approach. In the first case a pre-trained model is used, while in the second case, a training phase exploiting training data provided during the submission evaluation, precedes classification.

1. INTRODUCTION

The Speech/Music Classification and Detection task introduced for the first time in MIREX 2015 is organised as two distinct subtasks. The classification task is defined as the binary problem of classifying pre-segmented audio data to the speech or music class. Each evaluated segment is 30 seconds long and contains either speech or music data, mixed (speech over music) segments are not allowed. The detection task is focusing on finding segments of music and speech in a signal (i.e. finding segment boundaries) and classifying each segment as music or speech. The detection algorithm is evaluated on recordings from archives, which are typically at least several minutes long and contain multiple segments.

The rest of the paper is organized as follows: in Section 2 the audio feature extraction and pre-processing procedures are described; in Section 3 a detailed description for the detection task work-flow is presented, while in Section 4 the algorithm used for the classification task is analysed. In Sections 5 and 6 information about the submission packaging is provided and relevant references are included.

2. AUDIO FEATURES AND PREPROCESSING

The features presented in Table 1 are extracted from the audio signal using a Hanning sliding window with step and block size equal to 1024 samples at 44100 Hz sampling rate. Along with a classical audio feature extraction strategy, a temporal feature integration methodology was implemented and evaluated using audio feature statistics of aggregated windows. In particular, the extracted features are aggregated in groups of 64 frames and the mean and standard deviation values are calculated. The resulting feature vectors have a time resolution of 1.48 seconds (block and step size equal to 65536) and consist of 74 components. Feature selection was based on their successful integration in similar speech/music discrimination tasks [6] [4] [5] [2] and event detection algorithms [8]. Furthermore the performance of the selected feature set was evaluated through a set of experiments using a trial and error process.

Feature	**Dimension**
RMS Energy	1
ZCR	1
SpectralRolloff	1
SpectralFlux	1
SpectralFlatness	1
SpectralFlatnessPerBand	19
MFCC	13
Sum	37
Aggregated (*mean, std*)	74
Aggregated + PCA	8

Table 1. Extracted Features

As a preprocessing step, the extracted feature vectors are scaled in order to have zero mean and standard deviation equal to one for each component. Finally, a linear kernel Principal Component Analysis algorithm, aiming to improve generalisation and decrease processing requirements, is applied to reduce the dimension of the final feature vectors to 8 components.

3. DETECTION TASK WORKFLOW

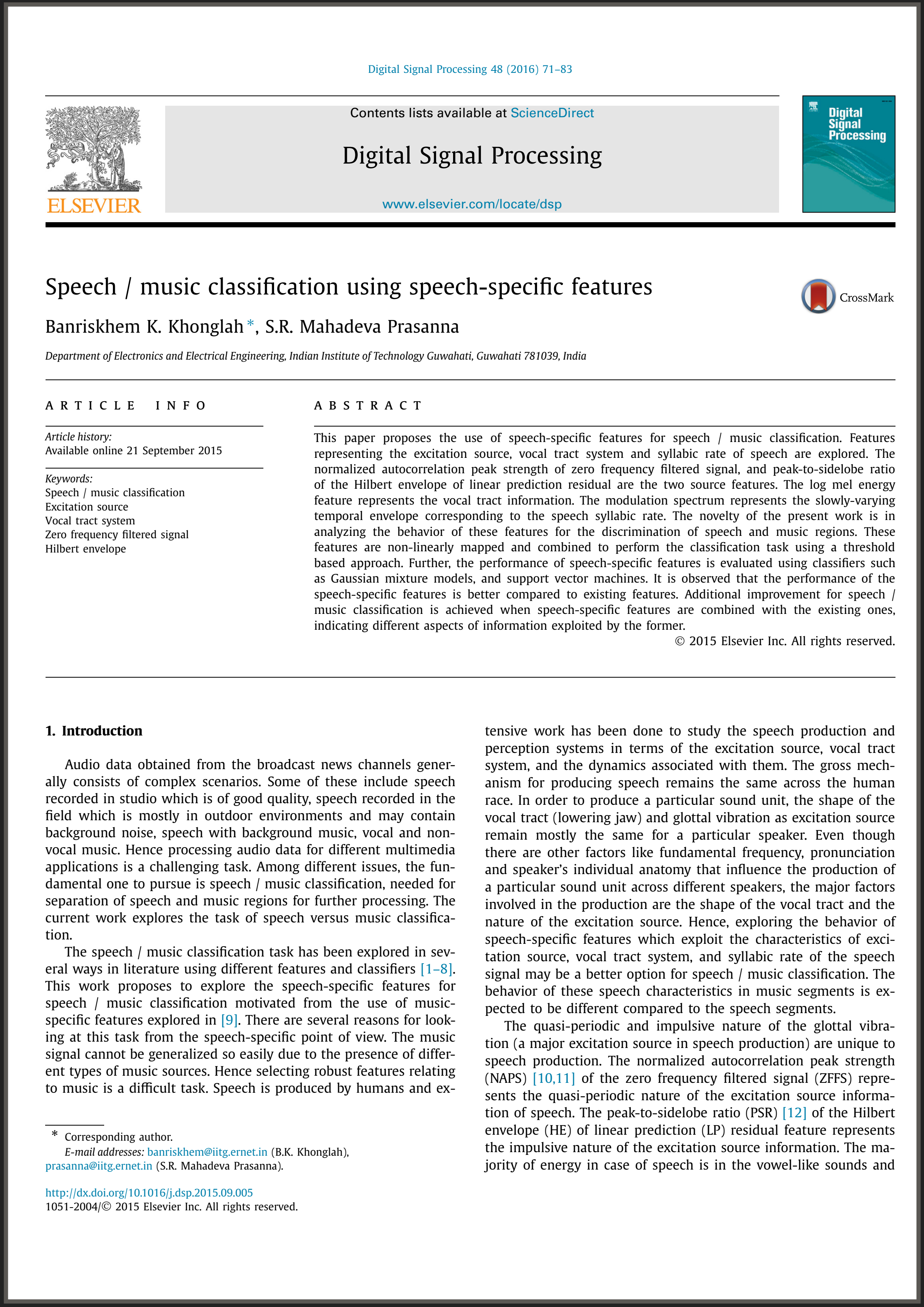
The main processing steps of the speech/music segment detection algorithm are illustrated in Figure 1 and discussed in the following section.

HOW IT'S DONE

SPEECH / MUSIC CLASSIFICATION USING SPEECH-SPECIFIC FEATURES

BANRISKHEM K. KHONGLAH, S.R. MAHADEVA PRASANNA

- Έμφαση σε features που εντοπίζουν ομιλία
  - NAPS of ZFFS (normalized autocorrelation peak strength of the zero frequency filtered signal
  - PSR (peak-to-sidelobe ratio)
  - HE of LP (hilbert envelope of linear prediction)
  - variance in the log mel spectrum energy
- Βελτίωση μεμονωμένης χρήσης και σε συνδυασμό



Digital Signal Processing 48 (2016) 71–83

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Speech / music classification using speech-specific features

Banriskhem K. Khonglah*, S.R. Mahadeva Prasanna

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

ARTICLE INFO

Article history:
Available online 21 September 2015

Keywords:
Speech / music classification
Excitation source
Vocal tract system
Zero frequency filtered signal
Hilbert envelope

ABSTRACT

This paper proposes the use of speech-specific features for speech / music classification. Features representing the excitation source, vocal tract system and syllabic rate of speech are explored. The normalized autocorrelation peak strength of zero frequency filtered signal, and peak-to-sidelobe ratio of the Hilbert envelope of linear prediction residual are the two source features. The log mel energy feature represents the vocal tract information. The modulation spectrum represents the slowly-varying temporal envelope corresponding to the speech syllabic rate. The novelty of the present work is in analyzing the behavior of these features for the discrimination of speech and music regions. These features are non-linearly mapped and combined to perform the classification task using a threshold based approach. Further, the performance of speech-specific features is evaluated using classifiers such as Gaussian mixture models, and support vector machines. It is observed that the performance of the speech-specific features is better compared to existing features. Additional improvement for speech / music classification is achieved when speech-specific features are combined with the existing ones, indicating different aspects of information exploited by the former.

1. Introduction

Audio data obtained from the broadcast news channels generally consists of complex scenarios. Some of these include speech recorded in studio which is of good quality, speech recorded in the field which is mostly in outdoor environments and may contain background noise, speech with background music, vocal and non-vocal music. Hence processing audio data for different multimedia applications is a challenging task. Among different issues, the fundamental one to pursue is speech / music classification, needed for separation of speech and music regions for further processing. The current work explores the task of speech versus music classification.

The speech / music classification task has been explored in several ways in literature using different features and classifiers [1–8]. This work proposes to explore the speech-specific features for speech / music classification motivated from the use of music-specific features explored in [9]. There are several reasons for looking at this task from the speech-specific point of view. The music signal cannot be generalized so easily due to the presence of different types of music sources. Hence selecting robust features relating to music is a difficult task. Speech is produced by humans and extensive work has been done to study the speech production and perception systems in terms of the excitation source, vocal tract system, and the dynamics associated with them. The gross mechanism for producing speech remains the same across the human race. In order to produce a particular sound unit, the shape of the vocal tract (lowering jaw) and glottal vibration as excitation source remain mostly the same for a particular speaker. Even though there are other factors like fundamental frequency, pronunciation and speaker's individual anatomy that influence the production of a particular sound unit across different speakers, the major factors involved in the production are the shape of the vocal tract and the nature of the excitation source. Hence, exploring the behavior of speech-specific features which exploit the characteristics of excitation source, vocal tract system, and syllabic rate of the speech signal may be a better option for speech / music classification. The behavior of these speech characteristics in music segments is expected to be different compared to the speech segments.

The quasi-periodic and impulsive nature of the glottal vibration (a major excitation source in speech production) are unique to speech production. The normalized autocorrelation peak strength (NAPS) [10,11] of the zero frequency filtered signal (ZFFS) represents the quasi-periodic nature of the excitation source information of speech. The peak-to-sidelobe ratio (PSR) [12] of the Hilbert envelope (HE) of linear prediction (LP) residual feature represents the impulsive nature of the excitation source information. The majority of energy in case of speech is in the vowel-like sounds and

* Corresponding author.
E-mail addresses: banriskhem@iitg.ernet.in (B.K. Khonglah), prasanna@iitg.ernet.in (S.R. Mahadeva Prasanna).

http://dx.doi.org/10.1016/j.dsp.2015.09.005
1051-2004/© 2015 Elsevier Inc. All rights reserved.

HOW WE ARE GOING TO DO IT

SPEECH / MUSIC CLASSIFIER

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

- MFCCs (Mel Frequency Cepstral Coefficients)
- Silence ratio
- ZCR (Zero-Crossing Rate)
- SC (Spectral Centroid)
- SR (Spectral Rolloff)
- SF (Specral Flux)
- 4Hz modulation
- Minimum Energy Density (MED)

ΠΑΡΑΘΥΡΟΠΟΙΗΣΗ

- Hamming παράθυρα με επικάλυψη 50%
- Μέγεθος 0.5 - 1 sec

ΜΟΝΤΕΛΟ ΤΑΞΙΝΟΜΗΣΗΣ

- Εξαρτάται…
  - NB, GMM, SVM, NN
- Όπως και οι αλγόριθμοι εκμάθησης εξαρτάται

HOW WE ARE GOING TO DO IT

STACK

- MATLAB
- R / Python
- WaveSurfer
- jMir / jAudio
- Julia / MusicProcessing.jl

ΕΥΧΑΡΙΣΤΟΥΜΕ

Αποστόλης, Φρανκ, Χριστίνα