interspeech 2008 incorporating SST 08

A Comparative Study in Automatic Recognition of Broadcast Audio

Stavros Ntalampiras, *Nikos Fakotakis*

Department of Electrical and Computer Engineering, University of Patras, Greece

sntalampiras@upatras.gr, fakotaki@wcl.ee.upatras.gr

Abstract

This paper provides a thorough description of a methodology which leads to high accuracy as regards automatic analysis of broadcast audio. The main objective is to find a feature set for efficient speech/music discrimination while keeping the number of its dimensions as small as possible. Three groups of parameters based on Mel-scale filterbank, MPEG-7 standard and wavelet decomposition are examined in detail. We annotated on-line radio recordings characterized by great diversity, for building probabilistic models and testing four frameworks. The proposed approach utilizes wavelets and MPEG-7 ASP descriptor for modeling speech and music respectively, and results to 98.5 % average recognition rate.

Index Terms: content-based audio recognition, speech/music discrimination, mfcc, mpeg-7, wavelets

1. Introduction

Humans experience a great deal of audio data in their everyday lives due to the massive growth of audio sources such as radio, telephone, television and the global network. Thus the need for automatic detection and classification of generated sounds in order to lighten the burden of manual annotation is increased. A robust methodology will have a great impact on numerous applications. Our work focuses on this problem from the scope of easing unsupervised news transcription and car-radio channel changing by automatic speech/music classification. The proposed methodology is applicable onto three tasks: (i) large audio database organization which is based on unattended annotation as speech or music, (ii) automatic news transcription systems can exploit such a method to elaborate on speech data *alone,* which results to improved performance and (iii) automatic radio channel changing. Think as a paradigm the situation where one is driving his car while listening to music on the radio. This device can be programmed in the following way: *if* a speech segment is detected *then* change the station *until* a music segment is found. An implementation with such capabilities would offer great convenience in our everyday routine.

The problem of speech/music classification has been addressed by a great deal of researchers. Pinquier et al [1] proposed a combination of two speech/music classification approaches. Both spectral and temporal features are taken into account for training Gaussian mixture models (GMM). Two stages are presented (speech/non-speech and music/non-music) which provide 94% accuracy for speech detection and 90% for music detection. The authors of [2] investigate the issue in order to assist an automatic speech recognition system. New features are utilized to analyze non-stationary signals and compute different energy types in each band obtained from the wavelet domain. Two classifiers are employed based on hidden Markov models (HMM) and experiments where carried out on three different databases. In [3] a system for radio broadcast indexing which is based on GMMs, normalization of features and fusion of classifiers is examined while a corpus of 12 hours was employed for evaluation. A different classification method was chosen in [4]. Spectral and temporal characteristics along with features derived from the wavelet transform are employed for training a multi-layer perceptron neural network while high classification rates are achieved.

This work is contributing to the field of automatic acoustic analysis for the purpose of "understanding" the surrounding environment by exploiting *only* the perceived auditory information similar to the way humans exhibit quite effortless. We use three different feature sets separately while their simultaneous usage is shown to produce the best recognition accuracies regarding each classification task. The first set is based on an alternate methodology for the derivation of the well known Mel frequency cepstral coefficients (MFCC). The second descriptor is Audio Spectrum Projection (ASP) which is a part of the MPEG-7 standard. Lastly we use a group of sound parameters that was presented in our previous work, and depends on multiresolution analysis.

The organization of this work is the following: in the next section we give a full description of all the involved feature sets. In section 3 the experimentations along with the classification method are reported. The last paragraph includes our conclusions and future work.

2. Feature sets

In this section we comment on the groups of descriptors that were employed in order to train probabilistic models which represent the a priori knowledge we have about the classes (speech and music). Mel-scale filterbank was selected because of its ability to sustain the most important information as regards human perception. The MPEG-7 standard is currently the state of the art methodology for automatic content-based sound recognition while the third set is based on the lower frequencies of wavelet analysis which include basic and distinctive information.

The parameters that were used during the extraction process of the first two groups were identical while a larger and non overlapped frame size was found to be sufficient for the third group. The final step of ASP descriptor reduces the dimensionality of Audio Spectrum Envelope (ASE) with respect to Audio Spectrum Basis (ASB). This method is data depended, thus a comparison between ASP and MFCC wouldn't be fair. Hence, we alternated the algorithm as regards the MFCC extraction and replaced the discrete cosine transform (DCT) stage with principal component analysis (PCA), inspired by ASB. Although the third set is consisted of a relatively small number of coefficients, PCA was exploited for its derivation. The purpose was to redefine a new orthogonal basis for better separation of the data. The above described approach is equivalent to identifying the best *set* of sound parameters for the specific task, instead of selecting the best individual parameters and combining them.

Accepted after peer review of full paper

2.1. Mel filterbank-based features

For the derivation of the first feature set, 23 Mel filter bank log-energies are utilized. The extraction method is the following: Firstly the short time Fourier transform (STFT) is computed for every frame while its outcome is filtered using triangular Mel scale filterbank. This process is proven to emphasize components which play an important role to human perception. Consecutively we obtain the logarithm to adequately space the data. At this point we explore the usage of an orthogonal decomposition technique instead of DCT. PCA is employed to reduce the dimensionality of the data while projecting them on axes derived from the data themselves. The basic kernel, which is composed of all the eigenvectors, is computed from the feature values coming from the whole training set. Subsequently only the first twelve vectors with the highest eigenvalues are kept. With this procedure the data are transformed to a new coordination system based on the relationships between them. The above explained process is shown in Fig. 1. Testing feature vectors are transformed to a lower dimension space based on the kernel derived from the training data. With this procedure 23 log-Mel filterbank coefficients are reduced to 12. It should be noted that PCA is a *data-driven* procedure unlike DCT which compacts data's energy with a standard weighting procedure. Moreover, each sound is cut into frames of 30 ms with 10 ms overlap after MPEG-7 standard recommendation for enabling robustness to possible misalignments. The sampled data are hamming-windowed to smooth any discontinuities while the FFT size is 512.

2.2. Audio Spectrum Projection

The ASP descriptor constitutes a powerful audio signal representation technique, which was introduced as a part of MPEG-7 Audio standard. MPEG-7 provides standardized tools for automatic multimedia content description and offers a degree of "explanation" of the information meaning. It eases navigation of audio data by providing a general framework for efficient audio management. Furthermore, it includes a group of fundamental descriptors and description schemes for indexing and retrieval of audio data.

ASP is based on the projection of signal's spectrum onto a low-dimensional feature space using decorrelated basis functions. In Figure 2, we depict the stages which are involved in the derivation of the particular descriptor. Initially the Audio Spectrum Envelope (ASE) descriptor is computed via STFT. All the parameters during this procedure were identical to the ones used in the previous feature set. ASE belongs to the basic spectral descriptors and is derived for the generation of a reduced spectrogram of the original audio

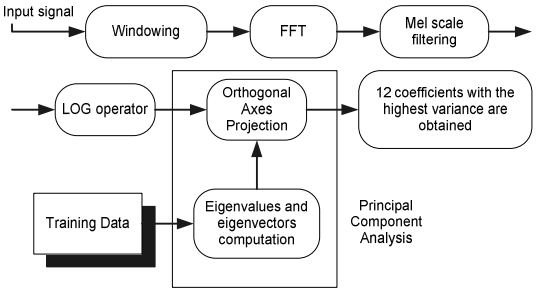

Figure 1: *Mel filterbank-based feature extraction process.*

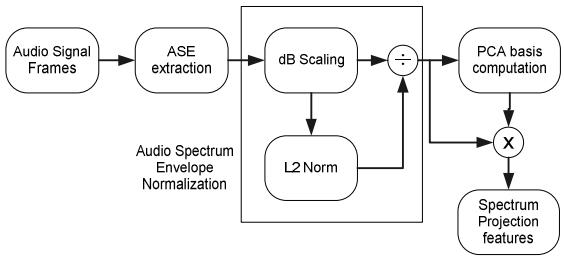

Figure 2: *Block diagram of ASP extraction.*

signal. It is a log-frequency power spectrum and calculated by summing the energy of the original power spectrum within a series of logarithmically distributed frequency bands utilizing a predefined resolution. Subsequently ASE is normalized and PCA basis is derived from the training data. Finally spectrum's projection is obtained by multiplying the normalized ASE (NASE) with the extracted basis functions [5]. During this process 27 normalized ASE vectors are reduced down to 12.

2.3. DWT-based sound parameters

In our previous work we examined audio processing by employing a well known multiresolution technique, discrete wavelet transform (DWT). The main advantage of the wavelet transform is that it can process time series which include non stationary power at many different frequencies (Daubechies 1990). This kind of analysis has been used in many different fields of research including denoising of signals and applications in geophysics (tropical convention, the dispersion of ocean waves etc) [6]. Wavelet comprises a dynamic windowing technique which can treat with different precision low and high frequency information content.

The first step of the DWT is the choice of the original (or mother) wavelet and by utilizing this function, the transformation breaks up the signal into shifted and scaled versions of it. We chose the Haar wavelet function, although this decision doesn't really affect the system's recognition ability as derived from our former work's experiments. The *Approximation* coefficient is taken under consideration which contains the low frequency information of the input sound. The feature vector is consisted of six statistical measurements. At the primary stage the DWT coefficient is cut into equal chunks of data using a texture size of 480 samples. The size of the Approximation coefficient is always half the size of the original signal due to the downsampling which is the last stage of DWT (downsampling is applied here in order not to end up having the double amount of data, as Nyquist theorem requests). It should be noted that no preemphasis is applied

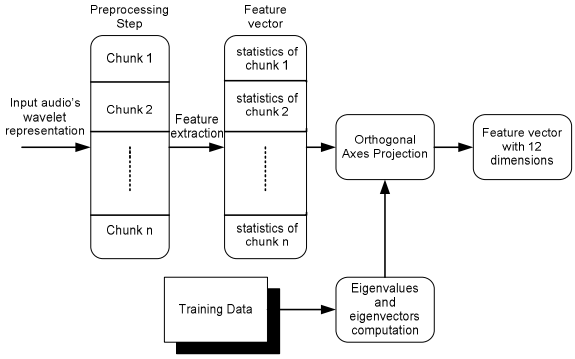

Figure 3: *DWT-based sound parameters derivation.*

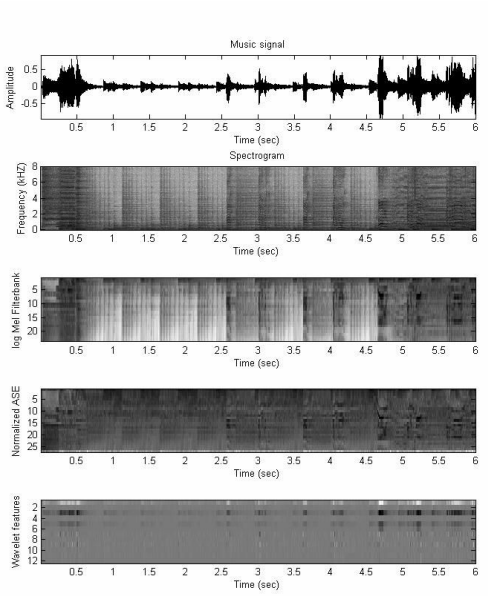

Figure 4: *Mel Filterbank, NASE and wavelet feature values against time as regards a music sample.*

and the analysis is performed onto non overlapping chunks. We sustained distinct information of the audio signal using only the following six statistical measurements taken over the texture size: (i) mean, (ii) variance, (iii) minimum value, (iv) maximum value, (v) standard deviation and (vi) median. The process is completed by juxtaposing their derivatives for obtaining a vector with 12 dimensions. Finally PCA is applied for projecting the data onto new orthogonal axes which provide better representation and distinction among the two classes.

3. Experimental set-up

3.1. Corpus

A descriptive analysis of the dataset that was used for conducting experiments is given in this paragraph. A total of 12 hours of BBC and CNN on-line radio recordings were manually annotated during this work phase. Silent segments were excluded using a statistical model based silence elimination algorithm [7] while average duration of a segment is 5.2 seconds (0.4secs<Seg_Dur<9.2secs). Approximately 3 hours of audio data were used for training while three different classification frameworks were tested onto the rest 9 hours (divided equally among the two classes). Care has been taken in order not to have portions of the same speech or music sample in both training and testing sets simultaneously. All the sounds were sampled at 16 KHz with 16 bit analysis (monophonic format). Speech and music sound samples are characterized by great diversity. Seven different speakers (male and female) are included while three of them exist only in the testing data. Speech category mainly contains news reports, weather forecasts, political arguments and interviews. The variability of music class is expressed in terms of how many different musical genres are contained. In this study the following six genres are included: pop, rock, jazz, classical, soul and country. Therefore, this corpus is considered to be more than adequate for building strong statistical models that describe each sound class.

3.2. Classification schema

Sound recognition is based on the fact that every sound source is distributing its energy across different frequencies in a different way. In order to model this property we employ a state of the art probability density function estimation technique, HMMs. The underlying assumption is that, sounds belonging to the same class follow a certain pattern across time. This pattern is tried to be approximated and identified during training and testing phases respectively.

The same schema is utilized for testing the recognition ability of the previously described feature sets. Ergodic (fully-connected) HMMs composed of 3 states are trained for each category using 50 iterations of the Baum-Welch algorithm. Each state is modeled by a Gaussian mixture consisting of 128 components. The parameters were selected after a series of experiments where the above combination resulted in a very good performance. While testing, feature streams enter each HMM and the outcome is a measure of the possibility that the particular model generated this succession of features. For the probability computation the well known Viterbi algorithm is used. System's final outcome is consisted of conventional maximum log likelihood estimation while decisions are made per frame. The implementation of HMMs is based on Torch machine learning library which is written in C++ [8]. It should be mentioned that the same training and testing data were employed for all the experiments and the results are tabulated in Table 1.

As can be seen, the Mel-filterbank based feature set provides the highest average recognition accuracy followed by ASP. This proves that this particular set of descriptors manage to sustain the most distinctive information between the two categories. However it doesn't provide the best results for neither of the two partial classification tasks. ASP was designed to better capture the properties of generic sound signals. Therefore its ability to characterize *music* signals is better as can be observed by the corresponding recognition rate. Although DWT-based sound parameterization results in the worst average performance, it can recognize *speech* samples with 98.1 % accuracy.

At this point we propose the combined usage of ASP and DWT statistical features. Incoming sounds are processed by both feature generation methodologies. In order our system to be practical and to overcome the problem of the different frame sizes, class predictions are made every 0.3 seconds (or 4800 samples). This portion of sound is divided according to each extraction algorithm. Their outcome is given as input to the already trained HMMs and a simple majority voting procedure takes place. Subsequently the above decisions are fed to the following process:

```
If (Decision_ASP_t=Music) then
    Final_Decision_t=Music;
Else
    If (Decision_Wavelet_t=Speech) then
        Final_Decision_t=Speech;
    Else
        Final_Decision_t= Final_Decision_t-1;
    End
End
```

From the above pseudocode it can be seen that the decisions are enhanced if they are made by the most accurate for each category model (ASP for music and DWT for speech). Otherwise the specific portion, *t* is categorized according to the previous one, *t-1*. We obtain better recognition rates by adopting this process. Average recognition rate reaches 98.5 % with 99 % and 98 % for speech and music classes respectively. We conclude that majority voting improved

Table 1. *Recognition rates achieved by all feature sets.*

Feature Set	Music	Speech	Average
Mel-filterbank	93.4 %	96 %	94.7 %
ASP	97.6 %	91.1 %	94.35 %
DWT	75 %	98.1 %	86.55 %

system's performance including both partial classification tasks.

4. Conclusions

We presented and evaluated the usage of two well known feature sets with psychoacoustic background and a new one, in combination with probabilistic models for automatic classification of broadcast audio. We exploited the ability of wavelet decomposition to model non-stationary signals with a small number of coefficients due to their asymmetric nature. A data driven approach based on PCA was adopted for the derivation of each set while the recognition rates on a test set consisting of 9 hours of radio recordings, are more than promising. The team of descriptors based on Mel filterbank demonstrated the best performance. A fusion scheme emphasizing on MPEG-7 ASP descriptor for modeling music, and wavelets for modeling speech was proposed and shown improved results. A non-redundant fusion of all the above feature sets will be a matter of our future work. Conclusively, our implementation is characterized by low complexity, thus able to run in real time and provide efficient broadcast audio analysis.

5. Acknowledgements

This work was supported by the EC FP 7th grant Prometheus 214901 "Prediction and Interpretation of human behaviour based on probabilistic models and heterogeneous sensors".

6. References

[1] Pinquier, J., Rouas, J. L. and Andre-Obrecht, R., "A fusion study in speech/music classification", in ICASSP-03, 2003.

[2] Didiot, E., Illina, I., Mella, O., Fohr, D. and Haton, J.-P., "A wavelet-based parameterization for speech/music segmentation", in INTERSPEECH-06, 2006.

[3] Senac, C., Ambikairajh, E., "Audio classification for radio broadcast indexing: feature normalization and multiple classifiers decision", Advances in Multimedia Information Processing - PCM 2004, Springer Berlin/Heidelberg, 2004.

[4] Kashif Saeed Khan, M., Al-Khatib, W. G. and Moinuddin, M., "Automatic classification of speech and music using neural networks", 2nd ACM international workshop on Multimedia databases, 2004.

[5] Casey, M., "MPEG-7 sound recognition tools", IEEE Trans. on Circuits and Systems for Video Technology, 11(6):737-747, 2001.

[6] Torrence, C. and Compo, G. P., "A practical guide to wavelet analysis", Bulletin of the American Meteorological society, 79(1):61-78, 1998.

[7] Sohn, J., "A statistical model-based voice activity detection", IEEE Signal Processing Letters, 6(1):1-3, 1999.

[8] http://www.torch.ch/

[9] Saunders, J., "Real-time discrimination of broadcast speech/music", in ICASSP-96, 1996.

[10] Bugatti, A., Flammini, A. and Migliorati, P., "Audio classification in speech and music: a comparison between a statistical and a neural approach", EURASIP Journal on Applied Signal Processing, 4:372-378, 2002.

[11] Allamanche, E., Herre, J., Hellmuth, O., Froba, B., Kastner, T. and Cremer, M., "Content-based identification of audio material using MPEG-7 low level description", in ISMIR-01, 2001.

[12] Cowling, M., "Non-speech environmental sound classification system for autonomous surveillance", PhD Thesis, Griffith University, 2004.

[13] Kim, H.-G., Moreau, N. and Sikora, T., MPEG-7 Audio and Beyond: audio content indexing and retrieval, Wiley Publishers, 2005.

[14] Rabiner, L. R., "A tutorial on Hidden Markov Models and selected applications in speech recognition", in Proc. IEEE, 77(2):257-286, 1989.

[15] Mertins, A., Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications, Wiley Publishers, 1999.