UPM-UC3M system for music and speech segmentation

Ascensión Gallardo Antolín

Dept. of Signal Theory and Communications
Universidad Carlos III de Madrid
gallardo@tsc.uc3m.es

Rubén San Segundo Hernández

Speech Technology Group
Universidad Politécnica de Madrid
lapiz@die.upm.es

Abstract

This paper describes the UPM-UC3M system for the Albayzín evaluation 2010 on Audio Segmentation. This evaluation task consists of segmenting a broadcast news audio document into clean speech, music, speech with noise in background and speech with music in background. The UPM-UC3M system is based on Hidden Markov Models (HMMs), including a 3-state HMM for every acoustic class. The number of states and the number of Gaussian per state have been tuned for this evaluation. The main analysis during system development has been focused on feature selection. Also, two different architectures have been tested: the first one corresponds to an one-step system whereas the second one is a hierarchical system in which different features have been used for segmenting the different audio classes. For both systems, we have considered long term statistics of MFCC (Mel Frequency Ceptral Coefficients), spectral entropy and CHROMA coefficients. For the best configuration of the one-step system, we have obtained a 25.3% average error rate and 18.7% diarization error (using the NIST tool) and a 23.9% average error rate and 17.9% diarization error for the hierarchical one.
Index Terms: music and speech segmentation, Chroma features, HMMs

1. Introduction

The problem of distinguishing speech signals from other audio signals (e.g., music) has become increasingly important as automatic speech recognition (ASR) systems are applied to more realworld multimedia domains, such as the automatic transcription of broadcast news, in which speech is typically interspersed with segments of music and other background noise. A pre-processing stage that segments the signal into periods of speech and non-speech is very important for improving recognition accuracy.

In another way, automatically detecting music parts from audio signals in TV or radio broadcasts is becoming a basic and important task to meet the increasing demands for multi-media indexing systems and music copyright management systems. In such audio signals, music is often overlapped by narration, conversation, or other environmental sounds.

Previous works on speech and music segmentation have been focused on features analysis or system architecture. About feature analysis, we can remark [1] where authors combine Mel Frequency Cepstral Coefficients (MFCCs) with other features like 4-Hz modulation energy, percentage of low energy frames, spectral centroid, spectral roll-off point, spectral flux, zero-crossing rate and spectral edge. In [2] histogram equalization-based features are proposed for speech, music, and song discrimination. In [3] an artificial neural network (ANN) trained on clean speech only (as used in a standard large vocabulary speech recognition system) is used as a channel model at the output of which the entropy and ‘‘dynamism’’ is measured every 10 ms. These features are then integrated over time through an ergodic 2-state (speech and non-speech) hidden Markov model (HMM) with minimum duration constraints on each HMM state. Finally, in [4], authors propose root mean square (RMS), and zero-crossings features for speech/music discrimination.

In respect of system architecture, [5] proposes a decision-tree-based algorithm for speech/music segmentation. [6] presents a comparison between two different techniques for speech/music discrimination. The first method is based on zero crossing rate and Bayesian classification. The second method uses more features and is based on neural networks (specifically a multi-layer Perceptron).

In work [7], authors propose a hierarchical system segmenting Broadcast News audio files. Finally, in [8] a system that segments an audio signal as speech and music by using posterior probability based features is proposed and implemented using Sphinx. This system uses Hidden-Markov-Model based acoustic models that are trained in Sphinx for posterior probability calculations. Acoustic models are trained with the HMM-states that are associated with the context-independent phones.

This paper presents the UPM-UC3M system for the Albayzín evaluation 2010 on Audio Segmentation. This system is based on Hidden Markov Models (HMMs), including a 3-state HMM for every acoustic class. For feature extraction, we have considered long term statistics of MFCC (Mel Frequency Ceptral Coefficients), spectral entropy [9] and CHROMA coefficients. Chroma features are a powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave [10].

2. Evaluation on Audio segmentation

The proposed evaluation task consists of segmenting a broadcast news audio document into a few acoustic classes (ACs):

- Speech [sp]. Clean speech in studio from a close microphone.
- Music [mu]. Music is understood in a general sense.
- Speech with noise in background [sn]. Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).
- Speech with music in background [sm]. Overlapping of speech and music classes or speech with noise in background and music classes.

There is another class that is not evaluated: Other [ot]. This class refers to any type of audio signal (including noises) that does not correspond to any other class.

2.1. Database description

The database consists of a Catalan broadcast news database from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporació Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, that includes around 87 hours of sound (24 files of approximately 4 hours long), has been split into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). The distribution of classes within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3%. The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

During system training and development, we used 16 files (sessions) that were divided randomly into two sets: 14 files for HMM training and 2 files (sessions 4 and 14) for testing the system performance.

2.2. Evaluation metrics

The proposed metric is inspired on the NIST metric for speaker diarization. The metric is defined as a relative error averaged over all ACs:

$$Error = \underset{i}{average}\left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}\right)$$

where

- $dur(miss_i)$ – the total duration of all deletion errors (misses) for the ith AC.
- $dur(fa_i)$ – the total duration of all insertion errors (false alarms) for the ith AC.
- $dur(ref_i)$ – the total duration of all the ith AC instances according to the reference file.

Note that incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. In the case when the system output is Other (non-Other) and the corresponding reference label is non-Other (Other), the audio segment is computed as a deletion (insertion) error for only non-Other AC. A forgiveness collar of 1 sec (both + and -) will not be scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an acoustic class begins/ends.

In this paper, we also report, for all the experiments, the NIST metric for speaker diarization. The NIST metric is similar to the error computed with the previous formula but doing a weight average: considering every AC duration for weighting each error.

3. Baseline

The baseline is a one-step system based on HMM. In particular, we have considered a 3-state HMM model for each acoustic class, considering 16 Gaussians per state. The HMM topology can be seen in Figure 1. The number of states has been adjusted from preliminary experiments.

The features considered in this system have been statistics over 1 second window (with an overlapping of 0.5 seconds) of the 15-MFCCs (Mel Frequency Cepstral Coefficients) and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and standard deviation. In total, there are 96 features every 0.5 seconds. In preliminary experiments, several windows lengths and overlapping were considered.

For all the experiments, we have used HTK software [11] for training and testing the HMMs. For feature extraction, we have considered the OpenSMILE tool [12].

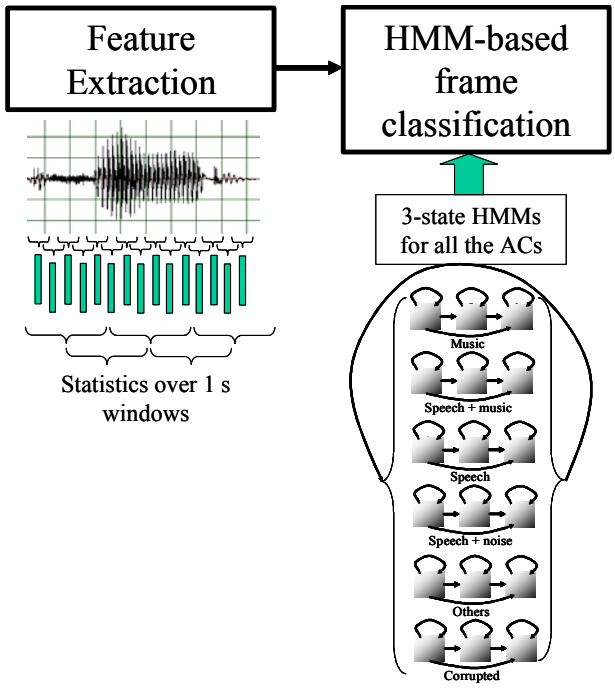

Figure 1. System diagram with details about feature extraction and HMM topology

4. Feature analysis

The main analysis during system development has been focused on feature selection. During system development, we have evaluated an important amount of features used in speech and speaker recognition. The best features for this task have been:

- MFCC15_E_D_A (mean+var): 15-MFCCs and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and variance computed along a 1 second with 0.5s overlapping.
- MFCC15_E_D_A (mean+std). Similar to the previous one but considering as statistics: mean and standard deviation.
- MFCC15_E_D_A (mean+std+skew): Similar to the previous one but considering as statistics: mean, standard deviation and skewness.
- MFCC15_E_D_A (mean+std+skew+kurt): adding kurtosis as a new statistic.
- MFCC15_E_D_A (mean+std+kurt): same that previous one removing skewness.
- PLP7_E_D_A (mean+std): 7-PLP (Perceptual Linear Prediction Coefficients) and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and standard deviation along a 1 second with 0.5s overlapping
- PLP5_E_D_A (mean+std): Same to previous one but considering 5 PLP instead of 7.

- PLP5_E_D_A (mean+std+skew+kurt): adding skewness and kurtosis as new statistics along 1 second window.
- MFCC15CHR_E_D_A (mean+std): 15-MFCCs, local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta and 12 CHROMA coefficients computed every 50 ms. Statistics are mean and standard deviation along a 1 second with 0.5s overlapping.
- MFCC15CHR+SpectralFeatures_E_D_A (mean+std): same to the previous one adding the statistics (mean and standard deviation) of several spectral features computed at 50ms frames (flux, centroid, entropy and band energies).
- MFCC15CHR+Entropy_E_D_A (mean+std): same to the previous one adding only the mean and standard deviation of the spectral entropy.

Table 1 presents the results for different features: FALA error for each AC, the average (AVG) and the NIST error.

Features	FALA error					NIST
	mu	sm	sp	sn	avg	
MFCC15_E_D_A (mean+var)	16.8	34.6	44.1	54.5	37.5	26.3
MFCC15_E_D_A (mean+std)	18.6	33.7	32.1	45.8	32.6	23.2
MFCC15_E_D_A (mean+std+skew)	14.7	33.0	38.0	51.8	34.4	24.8
MFCC15_E_D_A (mean+std+skew +kurt)	13.6	27.5	40.1	47.0	32.1	22.8
MFCC15_E_D_A (mean+std+ kurt)	15.2	31.8	42.5	49.8	34.8	24.6
PLP7_E_D_A (mean+std)	17.6	36.1	39.8	52.4	36.5	25.6
PLP5_E_D_A (mean+std)	13.6	38.1	40.9	53.5	36.5	26.3
PLP5_E_D_A (mean+std+skew +kurt)	12.3	34.2	41.6	50.4	34.6	25.1
MFCC15CHR_E_ D_A (mean+std)	19.1	24.9	29.4	41.3	28.7	20.9
MFCC15CHR+Sp ectralFeatures_E_ D_A (mean+std)	19.1	31.0	30.4	44.6	31.3	22.1
MFCC15CHR+ Entropy_E_D_A (mean+std)	15.7	27.7	28.0	38.9	**27.5**	**20.1**

Table 1. Results for different features: FALA error for each AC and the average (AVG) and the NIST error

From these experiments, the main conclusions are:

- Including CHROMA coefficients allows reducing significantly the error for all ACs from 32.6% to 28.7%.
- Including several spectral features frames (flux, centroid, entropy and band energies) do not improve the results. However, when only the spectral entropy is considered, the average error is reduced to 27.5%.
- For music segmentation, PLP features perform better than MFCC (considering even less number of features: 5 instead of 15) obtaining a 12.3% error.

5. Increasing the number of Gaussians per state

The feature study (in previous sections) was performed considering 16 Gaussians per state in order to reduce the time needed for the experiments. After this analysis, we decided to increase the number of Gaussians per state obtaining a good number around 128 or 256. Because of this, we decide to repeat the experiment with the most promising features but, in this case, considering 128 Gaussians per state, except in the case of MFCC15_E_D_A (mean+std) for which better results were achieved with 256 Gaussians. The considered features were:

- MFCC15_E_D_A (mean+std): 15-MFCCs and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and standard deviation along a 1 second with 0.5s overlapping.
- MFCC15_E_D_A (mean+std+skew+kurt): adding skewness and kurtosis as a new statistics.
- MFCC15CHR+Entropy_E_D_A (mean+std): 15-MFCCs, local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta, 12 CHROMA coefficients computed every 50 ms, and the 50ms frame spectral entropy. Statistics are mean and standard deviation along a 1 second with 0.5s overlapping.
- PLP5_E_D_A (mean+std+skew+kurt): 5-PLP and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean, standard deviation, skewness and kurtosis along a 1 second with 0.5s overlapping.

Table 2 presents the results for different features: FALA error for each AC, the average (AVG) and the NIST error.

Features	FALA error					NIST
	mu	sm	sp	sn	avg	
MFCC15_E_D_A (mean+std)	20.7	33.9	24.2	32.5	27.8	19.9
MFCC15_E_D_A (mean+std+skew+ kur)	17.3	29.7	26.6	37.9	27.9	19.8
PLP5_E_D_A (mean+std+skew+ kurt)	15.6	34.3	31.4	40.2	30.4	22.4
MFCC15CHR+ Entropy_E_D_A (mean+std)	14.7	25.7	26.8	34.0	**25.3**	**18.7**

Table 2. Results for different features and 128 or 256 Gaussians per state: FALA error for each AC and the average (AVG) and the NIST error

The main conclusions are:

- In all cases, when increasing the number of Gaussians the results are better. Best results are obtained when using MFCC+CHROMA+Entropy features.
- For music segmentation, PLP and MFCC+CHROMA+Entropy features perform better than only MFCC, whereas MFCC+CHROMA+Entropy are the best features for extracting speechmusic segments. However, speech and speechnoise are better segmented using only MFCC. Following these observations we consider the possibility to implement a hierarchical system in which different features are used for segmenting different acoustic classes. This system is described in next section.

6. Hierarchical system

According to the results presented in the previous section we designed a hierarchical system whose structure is shown in Figure 2. Two alternatives were considered:

- HS1: in this case, music and speechmusic are segmented using, respectively, PLP and MFCC+CHROMA

+Entropy features and finally, speech and speechnoise segments are extracted using MFCC parameters.

- HS2: MFCC+CHROMA+Entropy features are used for extracting music and speechmusic whereas MFCC are considered for segmenting speech and speechnoise.

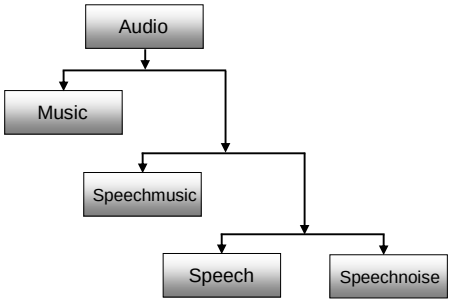

Figure 2. Block diagram of the hierarchical system

As it can be observed in Table3, results achieved by HS1 and HS2 systems are very similar. In comparison with the one-step system (with MFCC+CHROMA+Entropy features), the hierarchical one improves the results, yielding an error reduction of about 1.4% absolute.

System	FALA error					NIST
	mu	sm	sp	sn	AVG	
HS1	13.8	26.1	23.9	31.1	23.7	18.0
HS2	14.7	25.7	23.9	31.2	**23.9**	**17.9**

Table 3. Results for different hierarchical systems: FALA error for each AC and the average (AVG) and the NIST error

7. Final results

For the final evaluation, we presented two systems. The main system considering the best non-hierarchical system: using MFCC15CHR+ Entropy_E_D_A (mean+std) and the best hierarchical system HS1 described in previous section. The results are presented in Table 4.

System	FALA error				
	mu	sm	sp	sn	AVG
Main	19.2	25.0	39.5	37.2	30.2
Alternative	19.2	25.1	39.3	37.2	**30.2**

Table 4. Final results with the evaluation set: FALA error for each AC and the average (AVG)

8. Conclusions

This paper describes the UPM-UC3M system for the Albayzín evaluation 2010 on Audio Segmentation. The proposed system is based on Hidden Markov Models (HMMs), including a 3-state HMM for every acoustic class. The number of states and the number of Gaussian per state have been tuned for this evaluation.

The main analysis during system development has been focused on feature selection. From the experiments, we can conclude that MFCC are better complemented with CHROMA coefficients and spectral entropy than with other spectral features like flux, centroid and band energies in a one-step system.

Also a hierarchical system has been investigated obtaining slightly improvements over the one-step one when using MFCC features for segmenting speech and speechnoise and MFCC+CHROMA+Entropy for music and speechmusic.

In summary, for the best configuration of the one-step system, we have obtained a 25.3% average error rate and 18.7% diarization error (using the NIST tool) and a 23.9% average error rate and 17.9% diarization error for the hierarchical one.

9. Acknowledgements

This work has been supported by MA2VICMR (2010-2013), INAPRA (DPI2010-21247-C02-02), SD-TEAM (MEC ref: TIN2008-06856-C05-03), and TEC2008-06382 projects. Authors also want to thank all the comments from their colleagues at GTH-UPM and UC3M.

10. References

[1] Tomonori Izumitani, Ryo Mukai, and Kunio Kashino. "A background music detection method based on robust feature extraction". Proc ICASSP 2008.

[2] Ascensión Gallardo-Antolín and Juan M. Montero, "Histogram Equalization-Based Features for Speech, Music, and Song Discrimination". IEEE Signal processing letters, Vol. 17, No. 7, 2010.

[3] Jitendra Ajmera, Iain McCowan, Herve Bourlard. "Speech/music segmentation using entropy and dynamism features in a HMM classification framework". Speech Communication 40 (2003) 351–363.

[4] Costas Panagiotakis and George Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings". IEEE Trans. On Multimedia, Vol. 7, No. 1, Feb 2005.

[5] Yizhar Lavner1 and Dima Ruinskiy, "A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation". EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[6] Alessandro Bugatti, Alessandra Flammini, PierangeloMigliorati "Audio Classification in Speech and Music: A Comparison Between a Statistical and a Neural Approach". EURASIP Journal on Applied Signal Processing 2002:4, 372–378.

[7] Mateu Aguilo, Taras Butko, Andrey Temko, Climent Nadeu "A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task". Proc I SLtech 2009. Lisbon.

[8] Cemil Demir; Erdem Ünal, Mehmet Ugur Dogan, "A Sphinx Based Speech-Music Segmentation Front-End For Improving The Performance Of An Automatic Speech Recognition System In Turkish". CMU Sphinx Workshop 2010.

[9] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky, "Spectral Entropy Based Feature for Robust ASR", Proc. ICASSP 2004, pp. 193–196.

[10] M. A. Bartsch & G. H. Wakefield, "To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing," WASPAA, Mohonk, 2001.

[11] S. Young et al., HTK-Hidden Markov Model Toolkit (Ver 3.2). Cambridge, MA: Cambridge Univ., 2002.

[12] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Firenze, Italy, 25.-29.10.2010.