Albayzin Evaluation: Audio Segmentation System at CEPHIS-UAB

Héctor Delgado, Javier Serrano, Jordi Carrabina

Software-Hardware Prototypes and Solutions Lab, Autonomous University of Barcelona, Spain
hector.delgado@uab.cat, javier.serrano@uab.cat, jordi.carrabina@uab.cat

Abstract

This paper describes the audio segmentation system developed at the Software-Hardware Prototypes and Solutions Lab (Autonomous University of Barcelona) for the Albayzin 2010 Evaluations at the FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop.

1. Introduction

Audio segmentation problem consists of dividing audio streams into acoustically homogeneous segments. It is usually applied to a series of applications, such as indexing and retrieval, as a previous step to improve ASR accuracy by means of adaptation techniques, and so on. Before segmenting, a categorization of the acoustic classes must be done, depending on the particular domain where it will be applied. Typically, audio streams are segmented into silence, music, background noise, clean speech and speech corrupted with some kind of noise (music or background music).

The paper is structured as follows: section 2 describes the system, particularly the training and test data, feature extraction and acoustic classes modeling. Experimental results are given in section 3. Finally, some conclusion are extracted in section 4

2. System overview

The current section describes the audio segmentation system set-up in detail, particularly the training and test data, the feature extraction and the acoustic classes modeling and configuration.

2.1. Training data

It 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), will be 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.

2.2. Feature extraction configuration

PLP (further explanation in [1]) method has been chosen for the current task. PLP features have been empirically proven to be beneficial for audio segmentation tasks [2]. Firstly, a speech signal processing is made. A 0.97 coefficient pre-emphasis filter is applied, and a 25 ms Hamming window that scrolls each 10 ms is used to obtain signal frames. Then, a feature vector of 12 PLP coefficients is obtained from each frame using a 50 channel filter bank. Finally, the energy coefficient, delta and delta-delta features (time derivatives) are added to the feature vectors.

2.3. Acoustic classes modeling

An HMM for each acoustic class is created. Each HMM has three states in a left-to-right topology. Only the central state has a self-transition and a diagonal covariance matrix single-gaussian mixture model as emitting probability density function.

Then, the single-gaussian models are consecutively split into 2, 4, 8, 16, 32 and 64 mixture gaussians, re-estimating the model parameters using the Baum-Welch algorithm (HERest tool in HTK [3]) before doing each split.

Once the model set is obtained, the audio segmentation is performed through the Viterbi algorithm (HVite tool). Part of the training data has been used to tune some parameters in order to find a compromise between accuracy and computation time.

3. Experimental results

3.1. The metric

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

$$Error = average_i(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}) \quad (1)$$

where

- $dur(miss_i)$ is the total duration of deletion errors (misses) for the *i*th acoustic class.
- $dur(fa_i)$ is total duration of all insertion errors (false alarms) for the *ith* acoustic class.
- $dur(ref_i)$ is the total duration of all the *i*th acoustic class instances according to the reference file.

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.

Table 1: Segmentation error

	Music	Speech	Speech over music	Speech over noise	Average
Error (%)	23.65	45.07	36.95	45.21	**37.72**

3.2. Results

Experimental results are shown in table 1. A results discussion is given in section 4

3.3. Execution time

The experiment has been carried out in a Intel Core2 Duo 6420, 2.13 GHz CPU, 3 GByte RAM system. The operating system is Linux (Ubuntu 10.04).

The total execution time is shown next (output of the 'time' UNIX command).

```
real    22m1.533s
user    20m26.270s
sys     0m7.870s
```

4. Conclusions

Generally, it can be observed that the resulting error rates are considerably high. The used technique has been successfully applied on speaker segmentation tasks. However, other methods should be explored for audio segmentation purposes.

The system presents a better error when classifying music. On the other hand, the error rate increases considerably when the system tries to distinguish between the different kinds of speech. It indicates that the proposed models and feature extraction techniques are not totally suitable in order to classify speech in under different acoustic conditions. Possible improvements would be the use of other kind of acoustic class-based feature extraction instead of traditional features like MFCC. Other further improvement could be the application of a hierarchical architecture instead of classifying each acoustic class independently.

5. Acknowledgements

This work has been partly founded by the Spanish Ministerio de Ciencia y Innovacion, Project Reference TEC2008-03835/TEC. This article is supported by the Catalan Government Grant Agency Ref. 2009SGR700

6. References

[1] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001.

[2] H. Delgado, "Segmentación de Vídeo mediante Reconocimiento de Locutores," Master's thesis, Universitat Autónoma de Barcelona, June 2009.

[3] S. Young and G. Evermann, *The HTK Book*. Cambridge University Engineering Department, March 2009.