Albayzin 2014 Evaluations
TES-UAB Audio Segmentation System

Héctor Delgado, Javier Serrano

CAIAC, Autonomous University of Barcelona, Barcelona, Spain
{hector.delgado, javier.serrano}@uab.cat

IberSPEECH2014

**IberSPEECH, 19-21 November 2014,
Las Palmas de Gran Canaria (Spain)**

caiac

UAB
Universitat Autònoma de Barcelona

Introduction

Audio segmentation: segmenting an audio document into a given set of acoustic classes

Applications:
- Automatic indexing of audio documents for Spoken Document Retrieval
- Supporting ASR, speaker diarization and speaker recognition
- Improving ASR accuracy by means of adaptation

Evaluation: Segmenting a set of broadcast audio documents according to a series of audio classes: speech, music and noise. Two or more classes can be present in a given time instant (overlapping classes). Therefore, a multiple layer labeling is required.

Changes from 2012 audio segmentation evaluation: To increase difficulty, the input audio is composed of different databases that can be merged or even overlapped.

TES-UAB submission: The proposed audio segmentation system is based on the recently introduced Binary Key modeling, which has been successfully applied to speaker verification, speaker diarization, emotion recognition, and speech activity detection.

Database

Database: It consists of around 37 hours (21 hours for training/development and 15 for testing) of audio from the following databases:
- **Catalan broadcast news database**
- **Aragón Radio database**
- **Freesound.org**
- **HuCorpus**

Acoustic classes
- Speech
- Music
- Background noise

Two or more classes may be present at any time

Evaluation metric

Segmentation error time for each segment *n*

$$\Xi(n) = T(n)[\max(N_{ref}(n), N_{sys}(n)) - N_{correct}(n)]$$

$T(n)$: Duration of segment *n*
$N_{ref}(n)$: Number of reference classes present in segment *n*
$N_{sys}(n)$: Number of system classes present in segment *n*
$N_{correct}(n)$: Number of reference classes in segment *n* correctly assigned by the system

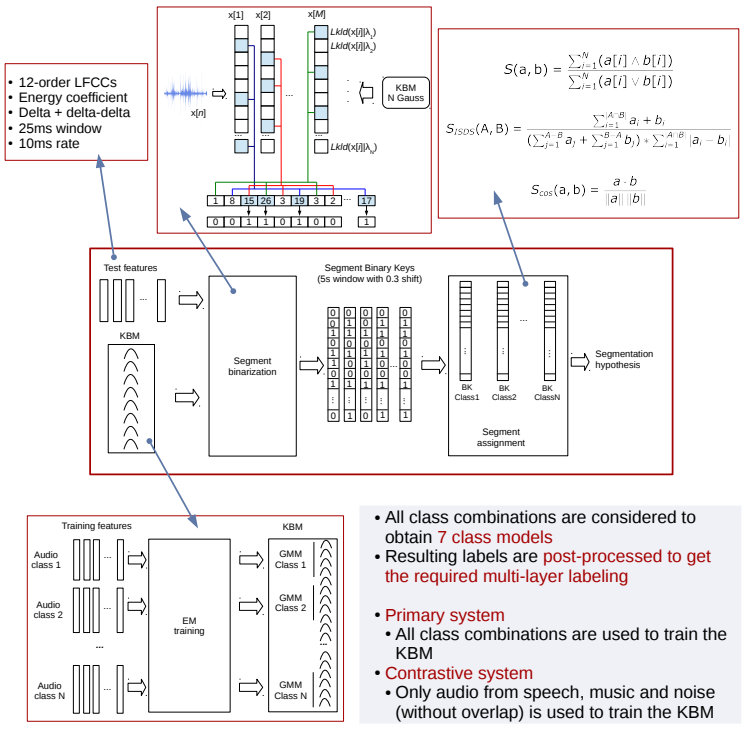
Segmentation Error Time (SER)

$$SER = \frac{\sum_{n\in\Omega}\Xi(n)}{\sum_{n\in\Omega}(T(n)N_{ref}(n))}$$

SER may be decomposed as:

- ***Class Error Time.*** Amount of time assigned to incorrect classes
- ***Missed Class Time.*** Amount of time that a class is present but not labelled
- ***False Alarm Class Time.*** Amount of time that has been assigned to a class which is not present

Audio segmentation system

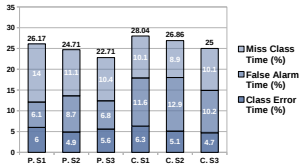


- All class combinations are considered to obtain 7 class models
- Resulting labels are post-processed to get the required multi-layer labeling

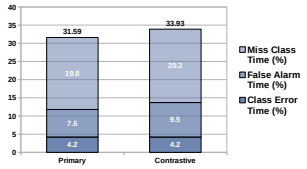
- Primary system
  - All class combinations are used to train the KBM
- Contrastive system
  - Only audio from speech, music and noise (without overlap) is used to train the KBM

Results

System results



SER of primary and contrastive system on development data, by using different binary Key similarity measures



SER of submitted primary and contrastive systems on the test data

- Cosine similarity (S3) is the best performing in both primary and contrastive systems
- Very high miss error time rate. Too much class time is not labeled by the system

- In the test data, class and false alarm errors keep similar to the results with development data.
- However, miss error is much higher (around 10% absolute).

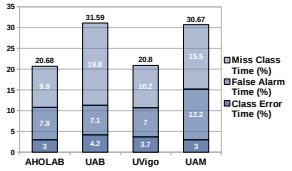
Execution time

	Primary system		Contrastive system	
Task	**Time**	**xRT**	**Time**	**xRT**
Feature extraction	00:02:17	0.002	00:02:17	0.002
Audio segmentation	07:13:02	0.462	03:11:38	0.204
Overall	07:15:20	0.464	03:13:55	0.207

Execution time taken by primary and contrastive systems when processing the test data (around 15 hours of data). Real time factor (xRT) is also provided

- Primary system execution time, although near 2 times faster than real time, could be too slow for processing big quantities of data
- Contrastive system provides a faster alternative at the cost of a decrease of performance

All participant results



All participant primary systems results

- Generally, the main challenge is to reduce the number of miss errors
- AHOLAB and UVigo perform very similarly, getting the best results of the evaluation (around 20% SER)
- UAM and UAB performances are very close but far from the top systems (around 30% SER)

Conclusions

- Audio segmentation system based on the Binary Key modeling
- Two different approaches to get the KBM in the primary and contrastive system
- Best result of experiments on development data is 22.71% SER
- Official result in the Albayzin evaluation is 31.58% SER
- Obtained performance is far from the top-performing systems ones (differences of around 10% absolute SER)
- Official evaluation top-performing systems have achieved around 6% absolute gain in performance with respect to 2012 evaluation
- The proposed audio segmentation task is still challenging, as obtained SER rates are still high (around 20% SER)

Acknowledgments

This work is part of the project “Linguistic and sensorial accessibility: technologies for voiceover and audio description”, funded by the Spanish Ministerio de Econonomía y Competitividad (FFI2012-31023)

This article is supported by the Catalan Government Grant Agency Ref. 2014SGR027