The Albayzin 2012 Audio Segmentation Evaluation

Alfonso Ortega, Diego Castan, Antonio Miguel, Eduardo Lleida

Vivolab, Aragon Institute for Engineering Resarch (I3A), University of Zaragoza, Spain

{ortega,dcastan,amiguel,lleida}@unizar.es

Abstract

This document describes the 2012 Albayzin Audio Segmentation Evaluation that will be conducted as part of the Iberspeech 2012 conference. Audio Segmentation is a very important task for some speech technologies applications like Automatic Speech Recognition or Spoken Document Retrieval. This evaluation consists of segmenting and labeling broadcast audio documents to indicate which segments contain speech, music and/or noise. The Segmentation Error Rate will be used as scoring metric as Diarization Error Rate is used in the Diarization evaluations organized by NIST as part of the RT evaluations.

Index Terms: audio segmentation, broadcast data, speech, music and noise.

1. Introduction

In some applications of speech technologies like Automatic Speech Recognition systems for Broadcast shows or Spoken Document Retrieval in very large multimedia repositories, Audio Segmentation is considered a very important task. Speech is usually found along with music or environmental noise, and the presence of each one of these acoustic classes must be accurately labeled, since the accuracy of these labels is critical for the subsequent systems to be successful. Thus, the development of accurate Audio Segmentation Systems is essential to allow applications like ASR or SDR to perform adequately in real-world environments

2. Description of the Evaluation

The proposed evaluation consists of segmenting a broadcast audio document and assign labels for each segment indicating the presence of speech, music and/or noise. That is, two or more classes can be found simultaneously in audio segments and the goal is to indicate if one, two or the three aforementioned classes are present for a given time instant. For example, music can be overlapped with speech or noise can be in the background if someone is speaking. In this evaluation, we consider that Speech is present every time that a person is speaking but not in the background. Music is understood in a general sense and Noise is considered every time some acoustic content is present different than speech and music (including speech in the background).

2.1. Changes from the 2010 Albayzin Audio Segmentation Evaluation

As in the 2010 Albayzin Audio Segmentation Evaluation, the goal is segmenting and labeling audio documents indicating where speech, music and/or noise are present. Nevertheless, in this evaluation, no prior classes are defined (*speech, music, speech with noise in background, speech with music in background, other*) and a multiple layer labeling is proposed. Therefore in the 2012 evaluation the goal is to segment the incoming audio into three possibly overlapped acoustic classes: Speech, Music and Noise.

3. Database Description

The Catalan broadcast news database from the 3/24 TV channel proposed for the 2010 Albayzin Audio Segmentation Evaluation will be used for training segmentation systems [1, 2]. This database was recorded by the TALP Research Center from the UPC in 2009 under the Tecnoparla project [3] funded by the Generalitat the Catalunya. The Corporació Catalana de Mitjans Audiovisuals (CCMA), owner of the multimedia content, allows its use for technology research and development. The database consists of around 87 hours of recordings in which speech can be found in a 92% of the segments, music is present a 20% of the time and noise in the background a 40%. Another class called *others* was defined which can be found a 3% of the time. Regarding the overlapped classes, 40% of the time speech can be found along with noise and 15% of the time speech along with music.

For this new evaluation, the Corporación Aragonesa de Radio y Televisión (CARTV) has donated part of the Aragón Radio archive. As the owner of the audio content, Aragón Radio and the Corporación Aragonesa de Radio y Televisión allow the use of these data for research purposes. Around four hours of the Aragón Radio database will be used for development and another sixteen hours of the Aragón Radio database will be used for testing.

All the data that will be supplied in PCM format, mono, little endian 16 bit resolution, and 16 kHz sam-

pling frequency.

4. Segmentation Scoring

As in the NIST RT Diarization evaluations [4], to measure the performance of the proposed systems, the segmentation error score (SER) will be computed as the fraction of class time that is not correctly attributed to that specific class (speech, noise or music). This score will be computed over the entire file to be processed; including regions where more than one class is present (overlap regions).

This score will be defined as the ratio of the overall segmentation error time to the sum of the durations of the segments that are assigned to each class in the file.

Given the dataset to evaluate Ω, each document is divided into contiguous segments at all class change points[1] and the segmentation error time for each segment n is defined as

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

where $T(n)$ is the duration of segment n, $N_{ref}(n)$ is the number of reference classes that are present in segment n, $N_{sys}(n)$ is the number of system classes that are present in segment n and $N_{Correct}(n)$ is the number of reference classes in segment n correctly assigned by the segmentation system.

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

The segmentation error time includes the time that is assigned to the wrong class, missed class time and false alarm class time:

- **Class Error Time**: The Class Error Time is the amount of time that has been assigned to an incorrect class. This error can occur in segments where the number of system classes is greater than the number of reference classes, but also in segments where the number of system classes is lower than the number of reference classes whenever the number of system classes and the number of reference classes are greater than zero.

- **Missed Class Time**: The Missed Class Time refers to the amount of time that a class is present but not labeled by the segmentation system in segments where the number of system classes is lower than the number of reference classes.

- **False Alarm Class Time**: The False Alarm Class Time is the amount of time that a class has been labeled by the segmentation system but is not present in segments where the number of system classes is greater than the number of reference classes.

A forgiveness collar of one second, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a class begins or ends.

4.1. Segmentation Scoring Tool and Audio Segmentation Systems Output Files

The tool used for evaluating the segmentation system is the one developed for the RT Diarization evaluations by NIST "md-eval-v21.pl", available in the web site of the NIST RT evaluations and directly accessible by clicking here.

The format's definition for the submission of the Audio Segmentation results has been fixed according to the operation of the NIST's tool. Specifically the Rich Transcription Time Marked (RTTM) format will be used for audio segmentation system output and reference files. RTTM files are space-separated text files that contain meta-data 'Objects' that annotate elements of each recording and a detailed description of the format can be found in Appendix A of the 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan [4]. Nevertheless, since in this evaluation, acoustic classes will be considered as if they were speakers in Diarization evaluations, the "SPEAKER" object will be used. Thus, the required information for each segment will be:

SPEAKER File Channel Beginning_Time Duration <NA> <NA> Class_Name <NA> <NA>

Where:

- **SPEAKER**: Is a tag indicating that the segments contains information about the beginning, duration, identity, etc. of a segment that belongs to a certain speaker. In our case, instead of a speaker, an acoustic class will be considered (speech, music or noise).

- **File**: Is the name of the considered file.

- **Channel**: Refers to the channel. Since we are dealing with mono recordings this value will always be 1.

- **Beginning_Time**: The beginning time of the segment, in seconds, measured from the start time of the file.

- **Duration**: Indicates the duration of the segment, in seconds.

[1] A "class change point" occurs each time any reference class or system class stars or ends. Thus, the set of active reference classes and/or system classes does not change during any segment

- **Class_Name**: Refers to the name of the class that is present in the considered segment (sp for speech, mu for music, no for noise).

The tag <NA> indicates that the rest of the fields are not used. The numerical representation must be in seconds and hundredth of a second. The decimal delimiter must be '.'.

As an example, let us consider that we are dealing with a recording contained in a file named session08.wav. Thus, the RTTM file name session08.rttm will contain the required information about that specific file. If the first three rows of the file are:

SPEAKER session08 1 2.67 17.91 <NA> <NA> mu <NA> <NA>
SPEAKER session08 1 11.98 13.30 <NA> <NA> sp <NA> <NA>
SPEAKER session08 1 25.28 76.20 <NA> <NA> no <NA> <NA>

This means that there is one segment containing music that starts at 2.67 sec. with a duration of 7.91 sec. Then in second 11.98 a speech segments starts with a duration of 13.30 sec. and finally, a noise segment starts at 25.28 sec. with a duration of 76.20 sec.

The Albayzin 2012 Audio Segmentation evaluation will use the md-eval version 21 software and the command line will be:
md-eval-21.pl -ac -c 1.0 -r <SPKR-REFERENCE>.rttm -s <SYSTEM>.rttm

5. General Evaluation Conditions

The organizers encourage the participation of all researchers interested in audio segmentation. All teams willing to participate in this evaluation must send an e-mail to

- ortega@unizar.es
- dcastan@unizar.es

Indicating the following Information:

- RESEARCH GROUP:
- INSTITUTION:
- CONTACT PERSON:
- E-MAIL:

with CC to the Chairs of the Albayzin 2012 Evaluations:

- javier.gonzalez@uam.es
- javier.tejedor@uam.es

before July 15, 2012.

All participant teams must submit at least a primary system but they can also submit up to two contrastive systems. Each and every submitted system must be applied to the whole test database. The ranking of the evaluation will be done according to results of the primary systems but the analysis of the results of the contrastive systems will be also processed and presented during the evaluation session at Iberspeech. All participant sites must agree to make their submissions (system output, system description, ...) available for experimental use by the rest of the participants or the organizing team.

Each participant team must provide also the total time required to run the set of tests for each submitted system (specifying the computational resources used). No manual intervention is allowed for each developed system to generate its output, thus, all developed systems must be fully automatic. Listening to the evaluation data, or any other human interaction with the evaluation data, is not allowed before all results have been submitted. The evaluated systems must use only audio signals. Any publicly available data can be used for training together with the data provided by the organization team to train the audio segmentation system. In case of using additional material, the participant will notify it and provide the references of this material.

5.1. Results Submission Guidelines

The evaluation results must be presented in just one RTTM file per submitted system. The file output file must be identified by the following code:

EXP-ID::=<SITE>_<SYSID> where,

- <**SITE**>: Refers to a three letter acronym identifying the participant team (UPM, UPC, UVI, ...)
- <**SYSID**>: Is an alphanumeric string identifying the submitted system. For the primary system the SYSID string must begin with p-, c1- for contrastive system 1 and c2- for contrastive system 2.

Each participant site must send an e-mail with the corresponding RTTM result files alogn with a technical description of the submitted systems to

- ortega@unizar.es
- dcastan@unizar.es

before September 30, 2012.

5.2. System Descriptions

Participants must send, along with the result files, a PDF file with the description of each submitted system. The format of the submitted documents must fullfil the requirements given in the IberSpeech 2012 call for papers at http://iberspeech2012.ii.uam.es. You can use the templates provided for the Iberspeech conference (WORD or LaTeX). Please, include in your descriptions all the esential information to allow readers to understand which are the key aspects of your systems.

5.3. Schedule (Tentative)

- May 18, 2012: Open the evaluation registration
- June 15, 2012: Release of the training and development data.
- July 15, 2012: Registration deadline.
- September 3, 2012: Release of the evaluation data.
- September 30, 2012: Deadline for submission of results and system descriptions.
- October 15, 2012: Results distribute to the participants.
- Iberspeech 2012 workshop: Official public publication of the results.

6. Acknowledgments

The Albayzin 2012 Audio Segmentation Evaluation organizing team would like to thank the Corporación Aragonesa de Radio y Televisión and Aragón Radio for providing the data for the evaluation. Thanks also to Taras Butko and Climent Nadeu who organized the 2010 Albayzin Audio Segmentation Evaluation for their help, support and for providing the training material for this evaluation. And also to the organizing committee of Iberspeech 2012 for their help and support.

7. References

[1] Butko, T, Albayzin Evaluations 2010: Audio Segmentation. Online: http://fala2010.uvigo.es/images/stories/pdfs/albayzinproposalaudiosegmentation_v1.pdf, accessed on 15 May 2012.

[2] Zelenak, M, Albayzin Evaluations 2010: Audio Segmentation. Online: http://fala2010.uvigo.es/images/stories/pdfs/speakerdiarizationevaluationplanfala2010_v2.pdf , accessed on 15 May 2012.

[3] Tecnoparla Project. Online: http://www.talp.upc.edu/tecnoparla , accessed on 15 May 2012.

[4] The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan. Online: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, accessed on 15 May 2012.