GTTS Systems for the Albayzin 2012 Audio Segmentation Evaluation*

Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, and Germán Bordel

GTTS (http://gtts.ehu.es), Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country UPV/EHU
Barrio Sarriena, 48940 Leioa, Spain
luisjavier.rodriguez@ehu.es

Abstract. This paper briefly describes the audio segmentation systems presented by GTTS to the Albayzin 2012 Audio Segmentation Evaluation (ASE). The same systems presented to the Albayzin 2010 ASE have been applied in a blind fashion, that is, with no specific tunings to the Aragon radio archive recordings from which test signals have been extracted. The primary system consists of an ergodic Continuous Hidden Markov Model with 5 states and 512 mixtures per state, each state representing a mix of audio sources: (1) music, (2) clean speech, (3) speech with music in the background, (4) speech with noise in the background and (5) other (noise, long silence fragments, etc.). The emission distributions corresponding to the HMM states were estimated on segments extracted from the Catalan broadcast news (3/24 TV) database provided for development, and transition probabilities were heuristically fixed. Given an input signal, this model produces an optimal decoding (and segmentation) according to the maximum likelihood criterion. The contrastive system consists of five 1024-mixture Gaussian Mixture Models (one per class, for the five classes mentioned above), estimated independently using 3/24 TV segments and applied on a frame-by-frame basis to get a sequence of smoothed log-likelihoods. The class yielding the maximum likelihood is chosen at each frame, and finally a mode filter is applied to smooth the sequence of decisions. The output of both systems was filtered so that it consisted of 3-class (speech, music and noise) possibly overlapping segments, as required in this evaluation, assuming that at least one of the categories must appear at any given frame.

Index Terms: Audio Segmentation, Gaussian Mixture Models, Hidden Markov Models

* This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06, by the Government of the Basque Country, under program SAIOTEK (project S-PE11UN065), and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially financed by FEDER funds). Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country.

1 Introduction

Audio segmentation and classification is a preprocessing step required by many applications, typically to filter out signals coming from undesired sources. Our approach to audio segmentation was motivated by the need for a speech detector in a speaker diarization system, which would allow us to discard non-speech segments (containing music, noise, etc.), so that clustering was only performed on speech segments.

The two systems presented to this evaluation were originally developed for the Albayzin 2010 Audio Segmentation Evaluation (ASE), and therefore they deal with five types of audio sources: (1) music, (2) clean speech, (3) speech with music in the background, (4) speech with noise in the background and (5) other (noise, long silence fragments, etc.). The acoustic models (one per class) and the classification approaches have been built and tuned based exclusively on the 3/24 TV channel recordings distributed for development, without any adaptation/tuning to the Aragon radio archive recordings used as test signals in this evaluation. Note also that, according to the above described motivation, our systems were not optimized for audio classification but for speech detection.

The first approach applied a 5-class ergodic HMM and performed maximum-likelihood Viterbi decoding. This approach yielded quite good performance in the speech detection task. The second approach, based on GMM frame-by-frame scoring, was developed with the aim to improve performance on multiclass audio segmentation tasks. It yielded better results than the ergodic HMM system on the Albayzin 2010 ASE datasets, but it heavily relied on parameter tunings, so it is presented as contrastive system in this evaluation. In both cases, the system output was filtered in order to produce 3-class (speech, music and noise) possibly overlapping segments (in RTTM formatted files), as required in this evaluation. We assumed that at least one of the classes must appear at any given frame, so no frame was left unassigned. The mapping of classes is presented in Table 1.

Table 1. Mapping classes from Albayzin 2010 ASE into Albayzin 2012 ASE.

	Speech	**Music**	**Noise**
Music		√	
Clean speech	√		
Speech + Background Music	√	√	
Speech + Background Noise	√		√
Other			√

All speech processing, HMM/GMM estimation, Viterbi decoding and GMM likelihood computations were performed with the Sautrela toolkit [1]. Text processing and file manipulation were all performed by means of UNIX utilities and applications (awk, SoX, etc.).

2 Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features. The choice of MFCC is based on the fact that historically there have been no features specifically designed for audio segmentation, and the MFCC are the most commonly used parameters for speech processing applications.

The audio was analysed in frames of 32 milliseconds (512 samples) at intervals of 10 milliseconds. A Hamming window was applied and a 512-point FFT computed. The FFT amplitudes were then averaged in 24 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. A Discrete Cosine Transform was finally applied to the logarithm of the filter amplitudes, obtaining 13 Mel-Frequency Cepstral Coefficients (MFCC), including the zero (energy) coefficient. Cepstral Mean Subtraction (CMS) was not applied, in order to keep channel and background information that may be relevant for audio classification.

3 Audio segmentation based on HMM decoding (primary system)

For development purposes, only the first 16 sessions of the 3/24 TV channel dataset were used. Sessions 3, 7, 11 and 13 were used for tuning purposes. The remaining 12 sessions were used to estimate model parameters, by splitting them (by means of SoX) into five subsets of segments (one per class), according to reference segmentations. A single-state HMM was estimated for each class, using the Baum-Welch algorithm on the corresponding set of segments. An ergodic Continuous Hidden Markov Model was then built by composing the five single-state HMMs under the Layered Markov Model framework defined in Sautrela [2]. Given an input sequence of feature vectors, the optimal decoding (and segmentation) was obtained by applying the Viterbi algorithm to get the optimal sequence of states in the ergodic HMM.

The number of mixtures per state (512) and the transition probabilities (auto-transitions fixed to 0.999999, transitions between states and final state transitions fixed to $2 \cdot 10^{-7}$) were optimized on audio segmentation experiments over the 4 tuning sessions mentioned above. Though system performance was quite poor for the multiclass setup defined in the Albayzin 2010 ASE, when applying it under a speech detection setup, the false alarm error rate was 1.16% and the miss error rate was 1.78% for the speech class.

4 Audio segmentation based on frame-by-frame GMM scoring (contrastive system)

This system was also developed using 16 sessions of the 3/24 TV channel dataset, as for the HMM-based system (12 sessions for training, 4 sessions for tuning). A GMM was estimated for each class, starting from the corresponding subset of

training segments. Given an input sequence of feature vectors, the set of GMMs was applied to compute frame-by-frame log-likelihoods. A smoothing window of length N was then applied, so that the log-likelihood of each class i was replaced by the arithmetic mean computed in that window, as follows:

$$\hat{ll}(i,t) = \frac{1}{N} \sum_{k=-N/2}^{N/2} ll(i,t+k)$$

At each frame, the class yielding the highest smoothed likelihood was chosen, and a frame-level sequence of class labels was produced. Finally, a mode filter of length M was applied to smooth the sequence of decisions. The number of mixtures of the GMMs (1024), the length of the score smoothing window ($N = 100$) and the length of the mode filter ($M = 200$) were optimized on audio segmentation experiments over the 4 tuning sessions. This system yielded better results than the HMM-based system in both the multiclass setup defined in the Albayzin 2010 ASE and the speech detection setup on which we were interested (false alarm error rate = 1.14% and miss error rate = 1.32%, for the speech class). However, since the performance of this system was quite sensitive to parameter tunings (and these were not optimized for the Aragon radio archive recordings from which test signals have been extracted in the Albayzin 2012 ASE), we have presented it as contrastive system.

5 Results on the Aragon radio development datasets

Tables 2 and 3 show the performance of the two audio segmentation systems described above on the Aragon radio development datasets (dev1 and dev2) provided for this evaluation. Besides the segmentation error score in the 3-class audio segmentation task, the segmentation error score in the corresponding speech detection task (computed by considering only the speech segments in both reference and system segmentations) is shown too. In both cases, miss, false alarm and class labeling error rates are presented in parentheses. Finally, to provide a performance ground, we provide the segmentation error scores for a *trivial* system which outputs a single segment including the full signal as containing both speech and music (the most common classes).

Table 2. Performance of the GTTS primary and contrastive audio segmentation systems on the Aragon radio development dataset *dev1*.

	% SER (miss, false alarm, labeling error)	
	3-class audio segmentation task	**speech detection task**
Primary	35.45 (12.8, 8.9, 13.8)	3.02 (2.5, 0.5, 0.0)
Contrastive	34.94 (11.8, 8.9, 14.2)	1.96 (1.1, 0.9, 0.0)
Trivial	35.74 (4.2, 15.6, 16.0)	7.33 (0.0, 7.3, 0.0)

Table 3. Performance of the GTTS primary and contrastive audio segmentation systems on the Aragon radio development dataset *dev2*.

	% SER (miss, false alarm, labeling error)	
	3-class audio segmentation task	**speech detection task**
Primary	38.70 (17.8, 9.4, 11.6)	3.03 (1.2, 1.8, 0.0)
Contrastive	37.29 (16.3, 8.8, 12.1)	2.83 (0.3, 2.5, 0.0)
Trivial	41.26 (3.4, 20.6, 17.2)	11.44 (0.0, 11.4, 0.0)

Experiments were carried out on a Dell PowerEdge 1950, equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 4GB of RAM. CPU times (in terms of real-time factor, $\times$RT) are shown in Table 4, considering three separate operations: (1) feature extraction, (2) model estimation and (3) audio segmentation. In the latter case, I/O operations and all the secondary computations needed to carry out the audio segmentation task are counted. Note that the primary system employs more time than the contrastive system for model estimation, but is faster for audio segmentation, providing only slightly worse performance in the speech detection task. The total CPU time, computed by adding CPU times for feature extraction and audio segmentation, falls below 0.05$\times$RT in both cases.

Table 4. CPU time (real-time factor, $\times$RT) employed in feature extraction, model estimation and audio segmentation for the GTTS primary and contrastive systems.

	Primary	**Contrastive**
Feature extraction	0.0033	
Model estimation	0.4819	0.1205
Audio segmentation	0.0375	0.0458

6 Conclusions

For the Albayzin 2012 Audio Segmentation Evaluation, GTTS has applied two fast systems (CPU time requirements falling below 0.05$\times$RT): a primary system based on a five-class ergodic HMM which outputs the optimal Viterbi-based sequence of states (and thus of classes) given an input signal, and a contrastive system based on frame-by-frame GMM scoring, followed by smoothing and mode filtering. The development effort has been almost null, since we simply recycled the two systems developed for the Albayzin 2010 Audio Segmentation Evaluation, applied them to the Aragon radio development and test signals (with no tunings), and adapted the output to the 3-class RTTM formatted output required in this evaluation, assuming that speech, music and/or noise must appear at any given frame.

When applied to the 3-class audio segmentation task, error rates were above 35%, only slightly better that those attained by a trivial system providing a fixed *speech and music* output. On the other hand, the same systems provided remarkably good performance in the corresponding speech detection task (between 2% and 3% detection error rates). This makes them suitable as speech detectors in many applications, such as speaker diarization, language recognition, ASR, etc.

References

1. M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, (San Juan, Puerto Rico), pp. 386–391, December 2005.
2. M. Penagarikano and G. Bordel, "Layered Markov Models: A New Architectural Approach to Automatic Speech Recognition," in *Proceedings of the MLSP Workshop*, (Sao Luis, Brasil), pp. 305–314, October 2004.