GTTS Systems for the Albayzin 2010 Audio Segmentation Evaluation

Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
luisjavier.rodriguez@ehu.es

Abstract

This paper briefly describes the audio segmentation systems developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the Albayzin 2010 Audio Segmentation Evaluation. The primary system consists of five Gaussian Mixture Models estimated independently on the reference segmentations provided for development, 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 contrastive system (used as speech/non-speeh detector in the GTTS submission to the Albayzin 2010 Speaker Diarization Evaluation) consists of an ergodic Continuous Hidden Markov Model with 5 states (one per class) and 512 mixtures per state. Independent sets of segments (extracted from the reference segmentations provided for development) are used to estimate the emission distributions corresponding to the HMM states, transition probabilities being heuristically fixed. Given an input signal, this model produces an optimal decoding (and segmentation) according to the maximum likelihood criterion.

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

1. Introduction

Our participation in this evaluation was motivated by our participation in the Albayzin 2010 Speaker Diarization Evaluation, since speaker diarization requires a speech/non-speech detector to discard non-speech segments (containing music, silence, noise, etc.), so that clustering is performed only on speech segments. Therefore, we have not optimized our systems for the classification task proposed in the evaluation, but for a speech/non-speech detection setup. We used the reference segmentations provided for development to estimate five acoustic models, and then applied two simple classification approaches, with two main concerns: rapid development and low computational cost. Our first (and quite obvious) approach consisted in estimating a 5-class ergodic HMM and applying maximum-likelihood Viterbi decoding. This approach yielded quite good performance in the speech/non-speech classification task. However, with the aim to improve performance on the 4-class segmentation task proposed in this evaluation, a second system was developed. The second system, based on GMM frame-by-frame scoring, yielded better results on the development set and is presented as the GTTS primary system. 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 done with UNIX utilities and applications (awk, SoX, etc.).

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

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 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 (contrastive system)

Development data were organized as follows: 4 sessions (3, 7, 11 and 13) were used for tuning purposes; the remaining 12 sessions were used to estimate model parameters. In fact, these latter sessions were splitted (using SoX) into five subsets of segments, according to reference segmentations provided with development data, for the five acoustic classes: (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.).

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 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 tun-

ing sessions. Though system performance was quite poor for the 4-class setup defined in the evaluation, when considering a 2-class speech/non-speech classification setup, the false alarm error rate was 1.16% and the miss error rate was 1.78% for the speech class (gathering the three speech sub-classes mentioned above). This system was used as speech/non-speech detector in the GTTS submission to the Albayzin 2010 Speaker Diarization Evaluation.

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

Development data were organized the same way 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 each log-likelihood 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 for the 4-class setup defined in the evaluation. When considering a 2-class speech/non-speech classification setup, the false alarm error rate was 1.14% and the miss error rate was 1.32% for the speech class.

5. Results

Tables 1 and 2 show the performance of the two audio segmentation systems described above on the development and evaluation sets, respectively. Besides the average segmentation error used to rank systems, miss and false alarm error rates in speech detection are shown too.

Table 1: Performance of the primary and alternative GTTS audio segmentation systems on a development set consisting of sessions 3, 7, 11 and 13.

	primary	**contrastive**
%error (AS)	43.48	48.08
%miss error (speech)	1.32	1.78
%fa error (speech)	1.14	1.16

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 3, 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 4-class audio segmentation task are counted. Note that the contrastive system employs more time than the primary system for model estimation, but is faster for audio segmentation, providing only slightly worse performance in the speech/non-speech segmentation 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 2: Performance of the primary and alternative GTTS audio segmentation systems on the evaluation set (sessions 17-24).

	primary	**contrastive**
%error (AS)	45.10	48.50
%miss error (speech)	1.23	1.55
%fa error (speech)	0.90	0.86

Table 3: CPU time (real-time factor, $\times$RT) employed in feature extraction, model estimation and audio segmentation for the primary and contrastive GTTS systems.

	primary	**contrastive**
Feature extraction	0.0033	
Model estimation	0.1205	0.4819
Audio segmentation	0.0458	0.0375

6. Conclusions

Two naive audio segmentation systems have been developed and evaluated: a primary system based on frame-by-frame GMM scoring and subsequent mode filtering; and a contrastive system based on a five-class ergodic HMM which outputs the optimal Viterbi-based sequence of states (classes) given an input signal. Though their performance on the 4-class audio segmentation task proposed in this evaluation was quite poor, they provided miss and false alarm error rates of around 1% in speech detection. This makes them suitable as speech/non-speech detectors for a speaker diarization system (as we actually intended). Systems have been built and evaluated in two weeks and their CPU time requirements fall below 0.05$\times$RT.

7. 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*, (São Luís,Brasil), pp. 305–314, October 2004.