SEGMENTAL MODELING FOR AUDIO SEGMENTATION

Hagai Aronowitz

IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, U.S.A
haronow@us.ibm.com

ABSTRACT

Trainable speech/non-speech segmentation and music detection algorithms usually consist of a frame based scoring phase combined with a smoothing phase. This paper suggests a framework in which both phases are explicitly unified in a segment based classifier. We suggest a novel segment based generative model in which audio segments are modeled as supervectors and each class (speech, silence, music) is modeled by a distribution over the supervector space. Segmental speech classes can then be modeled by generative models such as GMMs or can be classified by SVMs. Our suggested framework leads to a significant reduction in error rate.

Index Terms— Speech segmentation, voice activity detection, music detection, segmental modeling, GMM supervectors

1. INTRODUCTION

Speech/non-speech segmentation and music detection have been important unsolved problems for decades. Many approaches were suggested with considerable success, but the problem is still challenging due to the need for robust and extremely accurate solutions. This work was originally motivated by the requirement of an extremely accurate speech/non-speech segmentation algorithm as a preprocessing stage preceding broadcast speech transcription. For this application, every 1% of missed speech may account to 1% of WER (word error rate). Taking into account that WER for state-of-the-art broadcast transcription systems is in the order of 10% [1], a suitable solution must have much less than 1% speech misdetection. Accurate detection and removal of non-speech may improve speaker change detection and clustering, feature normalization and speaker adaptation, all leading to lower WER.

Speech/non-speech segmentation is also important for speaker and language recognition. For these applications it is more important to reject most of the non-speech audio while failing to detect some speech is tolerable.

Reviewing relevant prior work reveals that the most popular approach for audio segmentation is based on modeling the likelihood of a frame given a certain class with a Gaussian Mixture Model (GMM) using Mel Frequency Cepstral Coefficients (MFCC) features [2-4]. The likelihood is usually smoothed with the likelihood of neighboring frames. In [5] LDA was applied to the MFCC features and a 5-state automaton was used for smoothing. In [6] a 3-state HMM was used instead of a GMM to represent each class, and the frame length was adjusted to 60ms instead of 20ms. In [7, 8] the MFCC features were replaced by high-level features but modeling and classification were still done on a frame basis.

In [9], classification was done using SVMs. The classification was done on a segment level. In order to parameterize a segment, high-level features and MFCCs were extracted on a frame basis, and their mean and standard deviation were used as an input to an SVM.

In [10], an initial segmentation stage was done followed by rule based classification. The classification rules were based on frame-level features such as energy, zero-crossing-rate, pitch and spectral peaks.

In this paper we introduce a novel framework for modeling and classification of audio. In particular, we focus on classification of speech, silence and music. Our framework is based on a more general concept of intra-class inter-entity variability modeling which we have previously successfully applied for speaker recognition [11, 12] and for language identification [13].

The remainder of this paper is organized as follows. In section 2 we review the concept of intra-class variability modeling. In Section 3, we present our novel framework for audio classification. In Section 4, we describe the experimental setup, the baseline systems and the empirical results comparing the baseline systems to our proposed algorithm. Finally, section 5 concludes.

2. INTRA-CLASS INTER-ENTITY VARIABILITY MODELING

Let X be an entity we want to classify such as a speech segment for audio classification, an audio file for speaker recognition, a text for topic classification, etc. We assume X is represented by a sequence (not necessarily ordered) of feature vectors of length $n(x)$: $X=\{x_1,\ldots,x_{n(X)}\}$. The training set T is defined as a set of entities labeled with their class identity: $T=\{X_i, C_i\}$. The goal is given a test entity $Y=\{y_1,\ldots,y_{n(Y)}\}$, to classify it.

A common approach for computing the Maximum-Likelihood (ML) class for a test entity is assuming that the feature vectors are independent given a class identity C_j (equation 1):

$$\Pr(y_1,..,y_{n(Y)} \mid C_j) = \prod_{i=1}^{n(Y)} \Pr(y_i \mid C_j) \quad (1)$$

This approach was taken by the GMM-based algorithm for speaker recognition reported in [14], and by the GMM-based algorithms for speech/non-speech and audio classification reported in [2-4]. The frame-independence approach is also taken by other speech/non-speech algorithms [5-8, 10], though it is somewhat relaxed by using a state machine in [5], and using larger frames and an HMM in [6].

The weakness of the approach described above is that it does not capture intra-entity dependency. For example, for speech/non-speech modeling, there is a strong dependency between the frames of a segment. The dependency can be accounted to background noise, channel, speaker characteristics, and practically all other factors which are relativity constant during a segment. In order to capture such dependency, we use a modified generative model to model the generation of feature vectors, as follows:

Generate entity Y for class C_j:

1. Generate entity-distribution f_Y using class prior distribution over the entity-distribution space $\Pr(f_Y | C_j)$.
2. Generate a sequence of frames using distribution f_Y.

An entity-distribution is the distribution used to generate the frames of a single entity. Each class is modeled as a prior distribution over entity-distributions. In order to derive the likelihood of an observed entity Y given class C_j the product of the prior entity-distribution likelihood and the posterior entity likelihood should be integrated over the entity-distribution space:

$$\Pr(y_1,..,y_{n(Y)} \mid C_j) = \int_{f_Y} \Pr(y_1,..,y_{n(Y)} \mid f_Y) \Pr(f_y \mid C_j) df_Y \qquad (2)$$

Let us define f_Y^* as in equation (3):

$$f_Y^* = \arg\max_{f_Y} \Pr(y_1,..,y_{n(Y)} \mid f_Y) \qquad (3)$$

In order to develop simple and tractable training and classification algorithms we approximate the integral in equation (2) by noting that $\Pr(y_1,\ldots,y_{n(Y)} \mid f_Y)$ as a function of f_Y is concentrated sharply around f_Y^*. Therefore:

$$\Pr(y_1,..,y_{n(Y)} \mid C_j) \cong \Pr(y_1,..,y_{n(Y)}) \Pr(f_Y^* \mid C_j) \qquad (4)$$

We therefore can approximate the likelihood of entity Y given class C_j by first estimating a parameterization (distribution f_Y^*) for entity y and then computing the likelihood of distribution f_Y^* given class C_j. The prior probability of Y used in (4) cancels out for likelihood ratio (LR) testing. We define the described concept as intra-class inter-entity variability modeling.

In [11, 12] we used this approach in the framework of speaker recognition. In this case entity Y is an audio file, f_Y^* is assumed to be a GMM, and Pr $(f_Y^* | C_j)$ is modeled by constructing a supervector from the concatenated means of GMM f_Y^* and assuming its distribution to be multivariate normal with a shared covariance matrix across all speakers. Using this approach was beneficial for capturing intra-speaker intra-session variability such as channel and changing speaker characteristics, which can be assumed constant during the recognition entity (audio file) but variable for given a class (speaker).

3. INTER-SEGMENT VARIABILITY MODELING

It is clear from [2-10] that the natural classification entity for audio classification and segmentation is a sequence of frames (a segment) and not a single frame. It is difficult and probably impossible to distinguish with high accuracy between speech, silence and music given a single frame. However, speech cannot be assumed to be stationary during a long segment. Therefore, most of [2-10] choose to apply acoustic modeling on the frame level. Our novel approach is to apply the method of intra-class inter-entity variability modeling described above to audio classification and segmentation. We define the classification entities as uniformly spaced overlapping audio segments of length L (300ms) and parameterize each segment by adapting a universal background model GMM (UBM) to the segment's feature vector. The parameterization (GMM) is then modeled and classified by a segment based classifier. Using UBM adaptation for segment parameterization ensures that the parameterizations of different segments are aligned and comparable. The modified generative model is therefore the following:

Generate segment Y for class speech/silence/music:

1. Generate segment-GMM f_Y for current segment using class prior distribution over the GMM space $\Pr(f_Y |$ class$)$.
2. Generate a sequence of frames using GMM f_Y (assuming frame conditional independence given f_Y).

3.1. Proposed algorithm

The outline of the proposed algorithm is as follows:

Training

1. Train a UBM from the entire training corpus.
2. For each class:
   a. Define evenly spaced overlapping segments of length L.
   b. Estimate a GMM for each segment by adapting the UBM to the segment's frames.
   c. Construct a supervector from each GMM.

For GMM training:

3. For each class:
   Estimate a GMM for the class using the supervectors as training data.

For SVM training:

3. For each pair of classes:
   Train an SVM to classify between the supervectors of both classes.

Speech segmentation

1. Define evenly spaced overlapping segments of length L.
2. Estimate a GMM for each segment by adapting the UBM to the segment's frames.
3. Compute classification scores for each segment using either SVMs or GMMs.
4. Compare scores to pre-tuned thresholds (on a development dataset) and classify whole segments.
5. A frame is classified as music if it is part of any segment that was classified as music.
   Otherwise, it is classified as silence if it is part of any segment that was classified as silence.
   Otherwise, it is classified as speech.

3.2. Segment parameterization

We extract 24-order MFCCs + derivatives every 10ms with the mean normalized over the entire audio file. Segments were defined as sequences of 30 frames and are extracted every single frame for

the training dataset (to maximize the amount of training data) and every 15 frames for the test dataset.

Three GMM configurations were investigated for UBM training. The first configuration named GMM1 is a single Gaussian trained on the entire training dataset. The second configuration named GMM3 consists of three Gaussians, each of them trained separately for a different class – speech, silence and music. The third configuration named GMM128 is a 128-order GMM trained on the entire dataset. All GMMs have diagonal covariance matrices.

For the GMM1 configuration a supervector is created for a segment by adapting the single Gaussian UBM with the feature vectors of the segment and concatenating the mean of the Gaussian with the diagonal of the covariance matrix of the Gaussian into a single supervector. We apply the log function on the covariance components of the supervector in order to partly Gaussianize them.

For the GMM3 configuration a supervector is created similarly as for GMM1. The weights of the Gaussians are concatenated to the supervector after applying the log function in order to partly Gaussianize them.

Last, for the GMM128 configuration we only concatenate the weights after applying the log function (no mean and covariance components).

3.3. Segment classification

We have tested two different classifiers for segment classification. The first one is a GMM ML classifier and the second one is an SVM [15]. We used an SVM classifier with a Radial-Basis-Functions (RBF) kernel. For the SVM classifier the supervectors were pre-processed by dividing each supervector component by the corresponding standard deviation of that component in the entire training corpus.

4. EXPERIMENTS AND RESULTS

We tested the algorithms described in section (3) on Arabic broadcast news speech. The training and development data was obtained from the Linguistic Data Consortium. Four shows were segmented and labeled internally at IBM and used for training. 130 shows were annotated automatically using forced alignment and used as development data. The test data consists of 12 shows collected, segmented and labeled internally at IBM. The test set was recorded from five different broadcasting networks. The signal-to-noise (SNR) ratio for the test shows varies significantly from 40db to 10db.

4.1. Baseline systems

In order to evaluate the potential of our approach, we compared it under both the GMM and SVM frameworks. The GMM classifier is similar to those described is [2-4]. The SVM classifier uses an RBF kernel. Our baseline (GMM and SVM) classifiers use the same front-end as described in subsection (3.2). We optimized the parameters of the classifiers using the development dataset. Both classifiers produce scores on a frame level which are smoothed over evenly spaced overlapping audio segments with correspond to the segments used by our approach. The decision logic used is identical to the one used for the proposed approach.

4.2. Evaluation method

We report the accuracy of the various algorithms by measuring EER (Equal Error Rate), false alarm rate at low rejection rate (0.5%) and false rejection at low false alarm rate (1%). The various measures were chosen according to the potential applications for speech segmentation. For selected experiments we present DET curves [16] which we claim are significantly clearer than the usual receiver operating characteristic (ROC) curves for presenting speech classification results.

We have run many experiments in order to optimize the parameters of the GMM baseline system. We report only the results of the best configuration chosen on the development data. The SVM baseline was not heavily optimized. We have tested the 3 proposed segment parameterizations GMM1, GMM3 and GMM128 defined in subsection (3.2) using the classification systems described in table (1).

System	Segment parameterization	Segment classifier
GMM1+GMM	GMM1	GMM
GMM1+SVM	GMM1	SVM
GMM3+GMM	GMM3	GMM
GMM128+GMM	GMM128	GMM

Table 1: Speech classification systems evaluated

The results for the GMM128+GMM were not as good as the other systems and are not reported. This issue is discussed in section 5.

4.3. Speech / silence classification results

We report in table (2) the recognition results for speech/silence classification, and present in figure (1) selected DET curves.

System	EER	FA @ FR=0.5%	FR @ FA=1%
GMM baseline	2.92%	7.9%	29.6%
SVM baseline	2.51%	6.8%	14.6%
GMM1+GMM	**1.72%**	**5.1%**	**2.7%**
GMM1+SVM	1.96%	5.5%	5.4%
GMM3+GMM	2.21%	4.1%	24.6%

Table 2: Speech / silence classification results

We can see that both the GMM1+GMM and the GMM3+GMM systems reduce the error rate dramatically compared to the baseline GMM. For the SVM framework, we see a significant improvement for the GMM1+SVM system compared to the baseline SVM.

4.4. Speech / music classification results

Recognition results for speech/music classification are presented in table (3), and selected DET curves are presented in figure (1).

System	**EER**	**FA @ FR=0.5%**	**FR @ FA=1%**
GMM baseline	1.43%	3.4%	3.2%
SVM baseline	1.82%	9.4%	5.0%
GMM1+GMM	1.81%	4.6%	4.3%
GMM1+SVM	1.70%	3.4%	3.7%
GMM3+GMM	**1.27%**	**2.0%**	**1.9%**

Table 3: Speech / music classification results

We can see that the GMM1+GMM system degrades accuracy compared to the baseline, while the GMM3+GMM system reduced the error rate significantly. For the SVM framework, we see a dramatic improvement for the GMM1+SVM system compared to the SVM baseline.

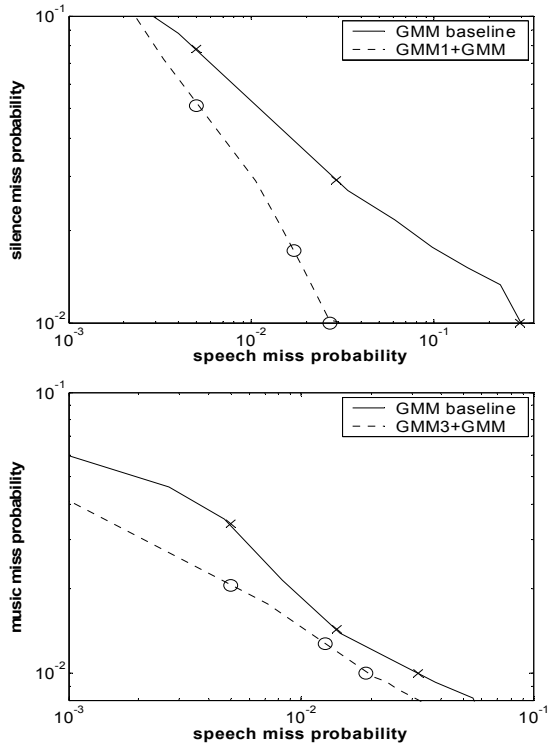

Figure 1: Selected DET curves for speech-silence (top) and speech-music (bottom) classification. The tabulated error rates are designated by 'x' for the baseline and 'o' for our approach.

5. CONCLUSIONS AND FUTURE WORK

In this paper we introduced a concept named intra-class inter-entity variability modeling we have used before for improved speaker recognition and described how it can be used successfully for speech segmentation and classification. We intend to extend our work by exploring better methods to model GMM supervectors. This can be done by trying to capture dependencies between GMM coefficients and coping with non-Gaussian distributions. We also intend to use the intra-class inter-entity variability modeling concept for other speech related classification challenges.

.

6. ACKNOWLEDGEMENTS

This work was partially supported by the Defense Advanced Research Projects Agency under contract No. HR0011-06-2-0001.

7. REFERENCES

[1] Y. Qin, Q. Shi, Y.Y. Liu, H. Aronowitz, S. M. Chu, H-K. Kuo, and G. Zweig, "Advances in Mandarin Broadcast Speech Transcription at IBM under the DARPA GALE Program", to appear in *ISCSLP*, 2006.

[2] C. Barras, X. Zhu, S. Meignier and J. L. Gauvain, "Improving speaker diarization" in Proc. *RT-04F* Workshop, 2004.

[3] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, P. C. Woodland, "The development of the Cambridge University RT-04 diarization system", in Proc. *RT-04F* Workshop, 2004.

[4] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln laboratory RT-04F diarization systems: applications to broadcast news and telephone conversations" in Proc. *RT-04F* Workshop, 2004.

[5] A. Martin, D. Charlet, L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC", Proc. *ICASSP*, 2001.

[6] C. Wooters, J. Fung, B. Peskin, X. Anguera, "Towards robust speaker segmentation: the ICSI-SRI fall 2004 diarization system", in Proc. *RT-04F* Workshop, 2004.

[7] F. Beaufays, D. Boies, M. Weintraub, Q. Zhu, "Using Speech/Non-Speech Detection to Bias Recognition Search on Noisy Data", in Proc. *ICASSP*, 2003.

[8] W.-H. Shin, B.-S. Lee, Y.-K. Lee, and J.-S. Lee, "Speech/Non-Speech Classification Using Multiple Features for Robust Endpoint Detection", in Proc. *ICASSP,* 2000.

[9] L. Lu, SH Li, and J. Zhang, "Content-based audio segmentation using support vector machines", in Proc. *ICME*, 2001.

[10] T. Zhang and J. Kuo., "Audio content analysis for online audiovisual data segmentation and classification", *IEEE Trans. on Speech and Audio Processing*, 9(4), 2001.

[11] H. Aronowitz, D. Burshtein., A. Amir, "A session-GMM generative model using test utterance Gaussian mixture modeling for speaker verification", in Proc. *ICASSP*, 2005.

[12] H. Aronowitz, D. Irony, D. Burshtein, "Modeling intra-speaker variability for speaker recognition", in Proc. *Interspeech*, 2005.

[13] E. Noor, H. Aronowitz, "Efficient language Identification using Anchor Models and Support Vector Machines", in Proc. *Odyssey*, 2006.

[14] D. A. Reynolds, T. F. Quatieri ,R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, Vol. 10, No.1-3, 2000.

[15] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[16] "The NIST Year 2004 Speaker Recognition Evaluation Plan", http://www.nist.gov/speech/tests/spk/2004/.