HMM-based Acoustic Event Detection with AdaBoost Feature Selection

Xi Zhou, Xiaodan Zhuang, Ming Liu, Hao Tang,
Mark Hasegawa-Johnson and Thomas Huang

Beckman Institute
Department of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA

Abstract. Because of the spectral difference between speech and acoustic events, we propose using Kullback-Leibler distance to quantify the discriminant capability of all speech feature components in acoustic event detection. Based on these distances, we use AdaBoost to select a discriminant feature set and demonstrate that this feature set outperforms classical speech feature set such as MFCC in one-pass HMM-based acoustic event detection. We implement an HMM-based acoustic events detection system with lattice rescoring using a feature set selected by the above AdaBoost based approach.

1 Introduction

There is a growing research interest in Acoustic Events Detection (AED). Although speech is the most informative auditory information source, other kinds of sounds may also carry useful information, such as in surveillance systems [3]. In a meeting room environment, a rich variety of acoustic events, either produced by the human body or by objects handled by humans, reflect various human activities. Detection or classification of acoustic events may help to detect and describe the human and social activity in the meeting room. Examples include clapping or laughter inside a speech discourse, a strong yawn in the middle of a lecture, a chair moving or door noise when the meeting has just started [12]. Detection of the nonspeech sounds also help improve speech recognition performance [8, 1].

Several papers have reported work on acoustic events detection for different environments and databases [13, 4]. AED as a task of CLEAR Evaluation 2006 [12] was carried out by the three participant partners from the CHIL project [2]: The UPC system is based on the Support Vector Machine (SVM) [10] discriminative approach and uses log Frequency Filter bank parameters and four kinds of perceptual features. Both the CMU and ITC systems are based on the Hidden Markov Model (HMM) generative approach using Mel-Frequency Cepstral Coefficients (MFCC) features [6]. In these works, we can see that Hidden Marcov Model (HMM) based Automatic Speech Recognition (ASR) framework worked better for detection task while the discriminative SVM approach was more successful for classification task. The main features for acoustic event detection are

still complete sets of speech perception features (critical band integration simulated by Mel/Bark filter bank or simple log frequency filter bank parameters) which have been proven to represent the speech spectral structure well. However, these features are not necessarily suitable for AED for the following reasons: 1) Limited work has been done in studying the spectral structure of acoustic events which is obviously different from that of speech. The speech features (such as filter bank parameters and MFCC) are designed according to the spectral structure of speech. Those features neglect the frequency parts that contain less speech discriminant information which may contain much discriminant information for acoustic events. 2) The Signal Noise Ratio (SNR) is low for AED. In the meeting room environment, the speech that co-occurs with the acoustic events most of the time should be seen as noise. Therefore, analysis of the spectral structure of acoustic events and design of suitable features are very important for AED task.

In this study we proposed a new front-end feature analysis and selection framework for AED. We characterize the features by quantifying their relative discriminant capabilities using Kullback-Leibler Distance (KLD) [7]. Adaboost [9, 5] based algorithm is used to select the most discriminant feature set from a large feature pool. The acoustic event detection experiments show that the discriminant feature set extracted by the data-driven methods significantly outperform the MFCC features without increasing the parameter number.

This paper is organized as follows: Section 2 analyzes the spectral correlates of acoustic events, and particularly quantifies the discriminant capabilities of all speech feature components in AED task by a KLD based criterion. In Section 3, the new AdaBoost based feature selection algorithm is proposed. Section 4 introduces the HMM-based system architecture for AED task. The experiment results are shown in Section 5, followed by the conclusion and the discussion of future work.

2 Spectral correlates of acoustic events

Currently, the speech features are designed mainly based on the properties of speech production and perception. The envelope of spectrogram (formant structure) instead of the fine structure of spectrogram (harmonic structure) is believed to hold most information for speech. Both log frequency filter bank parameters and Mel Frequency Cepstra Coefficients (MFCC) [6] use triangle bandpass filter to bypass the fine structure of spectrogram. Moreover, to simulate the non-uniform frequency resolution observed in human auditory perception, these speech feature sets adopt -uniform critical bands, providing high resolution in the low frequency part. However, the spectral structure of acoustic events is different from that of speech as shown in Figure 1, questioning the validity of using exactly a speech feature set for AED.

To analyze the spectral structure of acoustic events and design suitable features for AED, we carry out KLD based feature discriminant capability analysis. This helps us to understand the salient feature components of speech feature sets in the AED task. Intuitively, a discriminative feature component should sepa-

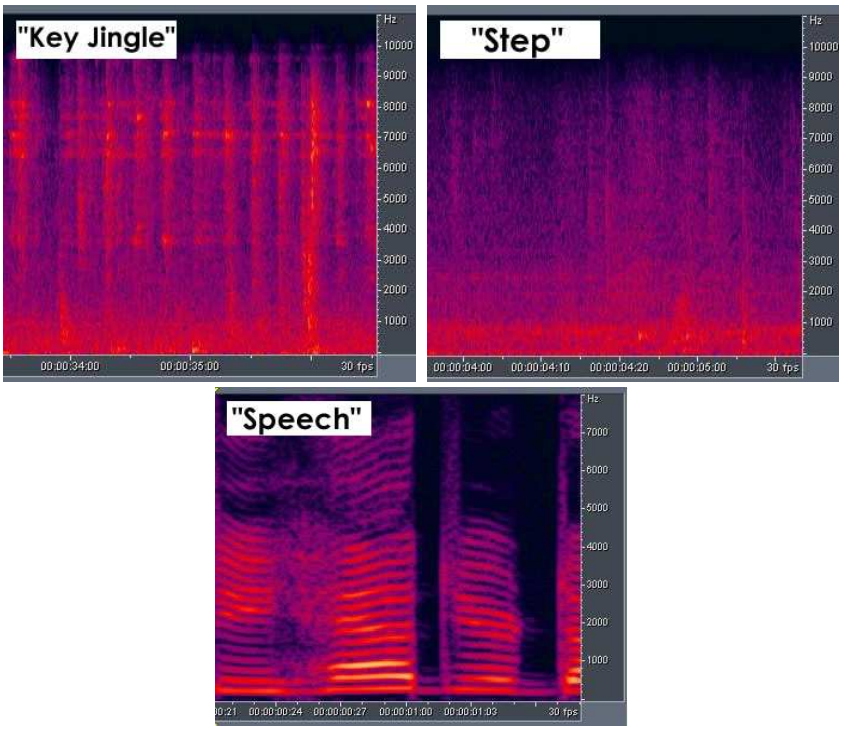

Fig. 1. Spectrograms of the acoustic events "Key Jingle", "Step" and human speech

rate an acoustic event from the other audio events (other events and speech). From a statistical point of view, more difference between the distributions of an acoustic event and the other audio parts results in smaller Bayesian error rates. The distance between the distributions of an acoustic event and the other audio parts reveals the discriminant capability of the feature for that acoustic event. Therefore, we introduce a KLD based analysis method to quantify the discriminant capability of feature components.

KLD ($D(p||q)$) is a measure between two distributions, p and q, and is defined as the cross entropy between p and q minus the self entropy of p.

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

We adopt KL distance to measure the discriminant capability of each feature component for each acoustic event, $d_{ij} = D(p_{ij}||q_i)$,where p_{ij} denotes the distribution of i^{th} feature component given the j^{th} acoustic event and q_i denotes the distribution of i^{th} feature component given all the audio parts. Then the global discrimant capability of i^{th} feature component is defined by

$$d_i = \sum_j P_j d_{ij} \tag{2}$$

where P_j is the prior probability for the j^{th} acoustic event.

Obviously, a larger global KL distance d_i means that the distributions in i^{th} component have larger difference between different acoustic events, thus having greater discriminant capability. In Figure 2, we show the global KL distances for different log frequency filter bank parameters for AED are different from those for speech digit recognition. The KL distances for speech digit recognition are calculated in the same way as described above, having speech digits in the place of acoustic events. All global KL distances in Figure 2 are mean normalized.

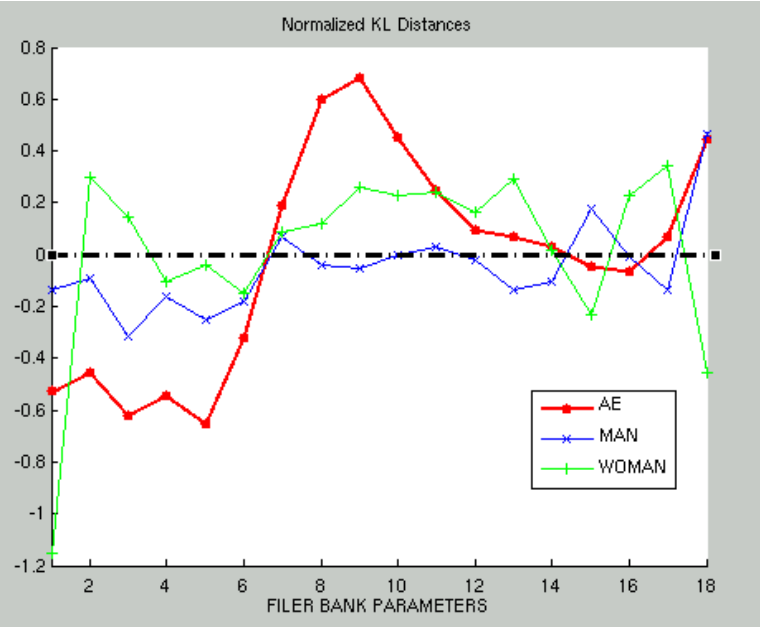

Fig. 2. Global KL distances for acoustic event detection, speech digit recognition (Men / Women).

3 Adaboost based feature selection

3.1 Adaboost algorithm

The basic Adaboost algorithm [5] deals with a 2 class classification problem. It iteratively selects and combines several effective classifiers among lots of weak

classifiers. For each iteration one weak classifier is chosen from the weak classifier pool and the error rate is required to be less than 0.5.

The basic steps of Adaboost are:

1. Given a set of sample $x_1, x_2, ..., x_m$, and the corresponding labels $y_1, y_2, ..., y_m$, where $y_i \in Y = \{-1, +1\}$ for negative and positive examples respectively.
2. Initialize weights $D_1(i) = \frac{1}{m}$ where m is the total number of positive and negative examples.
3. For $t = 1, ..., T$:
   (a) Find the classifier h_t that minimizes the error with respect to the weight D_t. The error of h_t is given by $\epsilon_t = \sum_{i=1}^{m} D_t(i)\,(h_t(x_i) \neq y_i)$
   (b) Choose $\alpha_t \in \mathbf{R}$, typically $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$
   (c) Update weights D_t:

$$D_{t+1}(i) = D_t(i) \frac{\exp\{-\alpha_t y_i h_t(x_i)\}}{Z_t}$$

   where Z_t is a normalization constant, such that $\sum_{i=1}^{m} D_{t+1}(i) = 1$
4. Output the final classifier

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Details about AdaBoost algorithm can be found in [9, 5].

3.2 Adaboost based feature selection

As described in the earlier section, we need to choose a set of features that can best separate each acoustic event from the other audio part. In this paper, AdaBoost is used to select the feature set but not to linearly combine several classifers. In our framework, each audio utterance in the development set is segmented to several acoustic event instances (as well as silence and speech) according to the labels. These event instances together with their labels serve as the labeled examples in AdaBoost. The weak classifiers in AdaBoost are of just one type: if the log likelihood of a particular example on the one-feature-component correct-label GMM is larger than that on the one-feature-component global GMM, this example is correctly classified.

4 HMM-based AED System Architecture

For the detection and classification of acoustic events, we implement a hidden Markov model (HMM)-based system with lattice rescoring using features selected by AdaBoost (Figure 3).

We formulate the goal of acoustic event detection in a way similar to speech recognition: to find the event sequence that maximizes the posterior probability of the event sequence $W = (w_1, w_2, ..., w_M)$, given the observations $O = (o_1, o_2, ..., o_T)$:

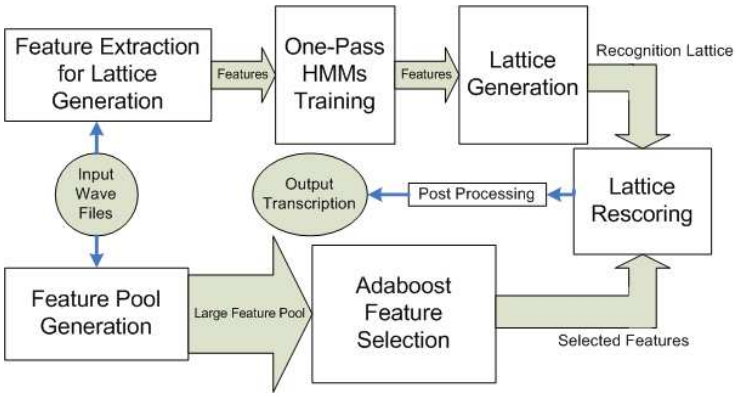

Fig. 3. AED System Architecture

$$\hat{W} = \arg\max_{W} P(W|O) = \arg\max_{W} P(O|W)P(W) \tag{3}$$

The acoustic model $P(O|W)$ is one HMM for each acoustic event, with three emitting states and left-to-right state transitions. To account for silence and speech, we use a similar HMM, but with additional transitions between the first and third emitting states. The structure of HMMs can model some of the nonstationarity of acoustic events. The observation distributions of the states are incrementally-trained Gaussian mixtures. Each HMM is trained on all the segments labeled as this event in the development seminar data. The language model, which is a bigram model, acounts for the probability of a particular event in the event sequence conditioned on the previous event: $P(w_1 w_2 \cdots w_m) = P(w_1) \prod_{i=2}^{m} P(w_i|w_{i-1})$.

Such a language model in acoustic event detection favors those recognized acoustic event sequences that are more similar to the sequences in the development data. Although the language model here does not have those linguistic implications as in speech recognition, it does improve performance, one of the possible reasons being to suppress a long sequence of identical event labels so that the internal structure of HMM can try to fit the temporal structure of the event segment.

In our HMM-based AED System Architecture, the upper part of Figure 3 generates a recognition lattice, which is a compact representation of N-best recognition outputs. Each edge in the lattice is annotated with acoustic model score and language model score. The acoustic features used to train HMMs for lattice generation consists of 26 frequency-filtered log filter-bank parameters or 26 MFCC, their overall energy, delta and acceleration, calculated on 25 ms Hamming windows with 10ms shifts.

The lower part of the architecture in Figure 3 selects the most discriminative feature set according to the AdaBoost approach described in the previous section.

This selected feature set is a subset of a feature pool having much more feature components. New HMMs trained using this feature set assign new acoustic scores to each arc in the recognition lattice. The best path in this updated lattice is output as the recognized event sequence.

5 Experiments

Our acoustic event detection experiments use about 3 hours development data with event boundaries and labels to train our systems, and test these systems on about 2 hour testing data, which is the official testing data for CLEAR 2007 AED Evaluation [11]. Both the development data and testing data are seminar style, having both speech and acoustic events with possible overlap. The performances are measured using AED-ACC (the first metric in CLEAR 2007 AED Evaluation). This metric aims to score detection of all acoustic event instances, oriented to applications such as real-time services for smart rooms and audio-based surveillance.

The first experiment is designed to compare the performance of one-pass HMM-based AED systems, using either a set of MFCC or the AdaBoost-selected feature set. The baseline set of MFCC (called MFCC13DA) is the widely-used feature set for speech recognition: 13 parameters calculated on 0Hz - 11000Hz band of the audio files. The delta and acceleration of these parameters are also included, forming a whole MFCC feature set of 39 components. The AdaBoost-selected feature set in this experiment (called SELECTED13DA) includes 13 feature components selected using the AdaBoost-based approach in Section 3 from a feature pool of 26 log frequency filter bank parameters and 26 MFCC parameters on the 0Hz-11000Hz band. The delta and acceleration of these selected feature components are also included, forming a feature set of 39 components.

Table 1. AED-ACC score of one-pass recognition using MFCC or selected feature set

AED-ACC	Dev	Test
MFCC13DA	38.92	25.27
SELECTED13DA	39.8	26.9

Table 2. AED-ACC score of complete system and one-pass system using 81 dimension MFCC

AED-ACC	Test
MFCC27DA (one-pass)	29.51
Selected 26DA	31.44** 33.6*

Table 1 shows that the AdaBoost-selected feature set outperform the MFCC feature set in recognition on both development data and test data. This indicates that for the AED task, using a complete set of features designed for speech recognition is far from optimal, and a feature set extracted using data-driven approach could yield better performance in AED without parameter increase.

The second experiment is designed to compare the performance of a one-pass HMM-based AED systems using MFCC and our complete system as described in Section 4. Reasonably increased parameter size would lead to better system performance. Therefore, the MFCC feature set used in this experiment (called MFCC27DA) consists of 27 MFCC parameters along with their delta and acceleration, forming 81 feature components. The complete system uses either the feature set MFCC27DA or a similar 81-dimension log frequency filter bank feature set to generate a recognition lattice. Then this system uses 26 feature components selected by AdaBoost from the same feature pool as in the first experiment, together with their delta and acceleration, forming a feature set of 78 dimensions (called SELECTED26DA), to rescore the recognition lattice and obtain the final recognition result.

Table 2 indicates that given a recognition lattice, using the selected feature set could benefit the system by updating the acoustic scores of the lattice and finding the optimal path in the updated lattice.

6 Conclusion

In this study, we use KLD to quantify the discriminant capability of all speech feature components in acoustic event detection. Global KLD shows that the speech feature components have different discriminant capabilities for speech recognition and acoustic event detection. The most discriminant feature set is extracted from a large feature pool using AdaBoost based approach. The acoustic event detection experiments show that the discriminant feature set extracted by the data-driven methods significantly outperform the MFCC features without increasing the parameter number in one-pass HMM-based acoustic event detection. And additional performance improvement is achieved by using speech feature set to generate recognition lattice and using the AdaBoost-selected feature set to rescore this lattice.

7 Acknowledgement

This research was supported in part by the U.S. Government VACE Program; and in part by National Science Foundation Grants 04-14117 and 05-34106. Findings and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Government or NSF.

References

1. F. Beaufays, D. Boies, M. Weintraub, and Q. Zhu. Using speech/non-speech detection to bias recognition search on noisy data. In *ICASSP03*, pages I: 424–427, 2003.
2. CHIL. Computers in the human interaction loop. http://chil.server.de/, 2006.
3. C. Clavel, T. Ehrette, , and G. Richard. Events detection for an audio-based surveillance system. In *ICME05*, pages 1306–1309, 2005.
4. R. Cui, L. Lu, H.-J. Zhung, and L.-H. Cai. Highlight sound effects detection in audio stream. In *ICME03*, pages III: 37–40, 2003.
5. Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.
6. H. Hermansky. Mel cepstrum, deltas, double deltas, .. -what else is new? In *Proc. Robust Methods for Speech Recognition in Adverse Condition*, 1999.
7. V. Krishnamurthy and J. Moore. On-line estimation of hidden markov model parameters based on the kullback-leibler information measure. *IEEE Trans. on Signal Processing*, 41(8):2557–2573, 1993.
8. A. Martin and L. Mauuary. Voicing parameter and energy based speech/non-speech detection for speech recognition in adverse conditions. In *Interspeech03*, pages I: 3069–3072, 2003.
9. G. Ratsch, T. Onoda, and K.-R. Muller. Soft margins for adaboost. *IEEE Trans. on Signal Processing*, 42:287–320, 2001.
10. B. Schòlkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, US, 2002.
11. A. Temko. Clear 2007 AED evaluation plan. http://isl.ira.uka.de/clear07, 2007.
12. A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of chil project systems. *Cough*, 65:5–11, 2006.
13. A. Temko and C. Nadeu. Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering. In *ICASSP05*, pages V: 505–508, 2005.