FEATURE ANALYSIS AND SELECTION FOR ACOUSTIC EVENT DETECTION

Xiaodan Zhuang, Xi Zhou, Thomas S. Huang and Mark Hasegawa-Johnson

Beckman Institute of Advanced Science & Technology
Department of Electrical & Computer Engineering
University of Illinois Urbana-Champaign, Urbana, IL 61801, USA

ABSTRACT

Speech perceptual features, such as Mel-frequency Cepstral Coefficients (MFCC), have been widely used in acoustic event detection. However, the different spectral structures between speech and acoustic events degrade the performance of the speech feature sets. We propose quantifying the discriminative capability of each feature component according to the approximated Bayesian accuracy and deriving a discriminative feature set for acoustic event detection. Compared to MFCC, feature sets derived using the proposed approaches achieve about 30% relative accuracy improvement in acoustic event detection.

Index Terms— Acoustic event detection, Feature Selection, Bayesian Accuracy, Hidden Markov Models

1. INTRODUCTION

Acoustic Event Detection (AED), a subtask of audio scene analysis [1, 2, 3, 4, 5, 6], has wide applications. In particular, information about non-speech sounds, i.e. (non-speech) acoustic events, reveals human and social activities. Examples include a chair moving or door noise when the meeting has just started [4], cheering of audience in a sports event [7], a gunshot in the street [8] and hasty steps in a nursing home. Such information is very helpful in applications such as surveilence, multimedia information retrieval and intelligent conference rooms. Some of the events are comparatively consistent and salient, such as cheering, while others are subtle, such as steps in a carpeted meeting room, laptop keyboard typing and paper wrapping.

Previously reported works have focused on the problems of segmenting audio into a small number of categories [2, 3], segregating a few audio sources [1, 9], and detecting a few highlight acoustic event [5]. Computers In the Human Interaction Loop (CHIL) & National Institute of Standards and Technology (NIST) held AED evaluation in CLEAR 2006 [4] and 2007, attempting to identify both the temporal boundaries and labels of twelve acoustic events in a real seminar environment. Many of the acoustic events are either subtle (low SNR, e.g. steps, paper wrapping, keyboard typing), or/and overlapping with speech, making the task particularly challenging. Although various system architectures and feature sets have been explored [4], even the top rated AED system has rather low performance [10].

A suitable feature set plays an important role for AED. Various audio perceptual features have been proposed for different analysis tasks [1, 11, 5]. In recent CLEAR Evaluations for AED, the most popular features are complete sets of speech perception features [4, 6], such as Mel-Frequency Cepstral Coefficients (MFCC) and log frequency filter bank parameters, which have been proven to represent 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. The speech features designed according to the spectral structure of speech might be far from optimal for AED 2) The Signal-to-Noise Ratio (SNR) is low for AED especially when the overlapping speech can be seen as noise. Therefore, analysis of the spectral structure of acoustic events and design of suitable feature sets are important for AED.

We propose quantifying the discriminative capability of each feature component according to the approximated Bayesian accuray and deriving a discriminative feature set for acoustic event detection. All feature components in a feature pool are first decorrelated by Principal Component Analysis (PCA). Then we apply nonparametric distribution estimation on all decorrelated feature components for each acoustic event. With these estimated distributions, we can approximate the Bayesian accuracies for the classification of events, which quantify the discriminative capability of the decorrelated feature components and guide our feature selection. We demonstrate that this proposed feature analysis and selection framework can conveniently derive feature sets for acoustic event detection (the new task) from a conventional speech feature pool engineered for speech recognition (the original task). HMM-based AED systems using the derived feature sets outperform the baseline system using MFCC with identical number of parameters.

This research is 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.

2. SPEECH PERCEPTUAL FEATURES IN ACOUSTIC EVENT DETECTION

Over the past decades, a lot of research has been done on speech perceptual features [12, 13]. These 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 MFCC [12] use triangle bandpass filters 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 non-uniform critical bands, providing high resolution in the low frequency part.

Speech perceptual features have been widely used in audio analysis [4, 6]. However, the spectral structure of acoustic events is different from that of speech as shown in Figure 1.

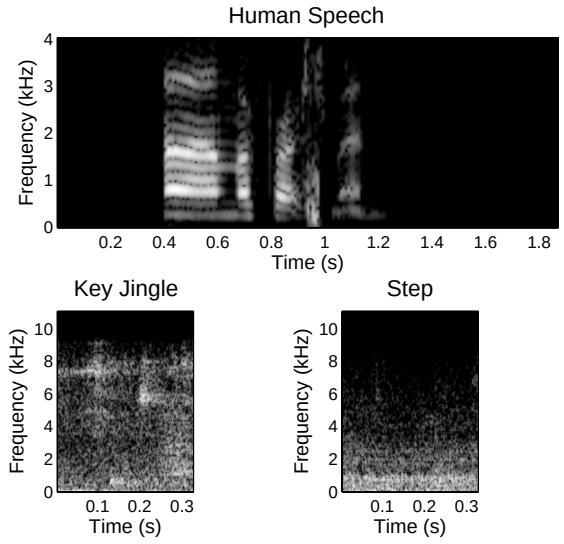

Fig. 1. Spectrograms of the acoustic events "Key Jingle", "Step" and human speech

Therefore, the speech feature sets designed according to the spectral structure of speech could be far from optimal for AED, questioning the validity of using exactly speech feature sets for AED. For example, they might neglect the frequency parts that contain less speech discriminative information which may contain much discriminative information for acoustic events. On the other hand, they are designed to particularly emphasize the acoustics that differentiate speech phonemes.

3. MEASURE OF DISCRIMINATIVE CAPABILITY

We propose quantifying the discriminative capability of each feature component according to the approximated Bayesian accuray. Intuitively, the feature components with more discriminative capability should have higher Bayesian accuracy. This will help us to understand the salient feature components of speech feature sets in the AED task and design suitable feature sets for AED.

For multi-class case, Bayesian accuracy is defined as:

$$\begin{aligned} P(correct) &= \sum_{i=1}^{c} P(x \in \mathcal{R}_i, \omega_i) \\ &= \sum_{i=1}^{c} P(x \in \mathcal{R}_i | \omega_i) P(\omega_i) \\ &= \sum_{i=1}^{c} \int_{\mathcal{R}_i} P(x|\omega_i) P(\omega_i) dx \end{aligned} \quad (1)$$

where $P(\omega_i)$ is the prior probability for i^{th} class and $P(x|\omega_i)$ is the likelihood for an observation x of the i^{th} class. Notice $\mathcal{R}_i$ defines a particular region in feature space, where the i^{th} class gives the highest likelihood:

$$\mathcal{R}_i = \left\{ x | \arg\max_k P(x|w_k) = i \right\} \quad (2)$$

Therefore we can approximate the Bayesian accuracy on a data set $X = x_1, x_2, \ldots, x_T$ as

$$P(correct) \approx \frac{1}{T} \sum_{t=1}^{T} \delta \left(\arg\max_k P(x_t|w_k) - l(t) \right) \quad (3)$$

where $l(t)$ denotes the true label for the t^{th} instance, and $\delta(*)$ is the Dirac delta function.

To calculate the Bayesian error rate for a feature component and without prior knowledge for each feature component's distribution, we adopt nonparametric density estimation. Parzen window density estimation is a technique for nonparametric density estimation [14]. Given a kernel function, the distribution of a given training set is approximated by a linear combination of kernels centered at the observed data points. In this study, we use parzen windows with Gaussian kernel function to estimate the distribution on each feature component for each event.

Figure 2 shows the varying Bayesian accuracy for all 52 feature components of a feature pool consisting of 26 log frequency filter bank parameters and 26 MFCCs.

4. FEATURE SET DERIVATION

We propose to select feature components in a feature pool according to their Bayesian accuracy on training dataset. To reduce the correlation between different feature components, we first apply Principal Component Analysis to the feature

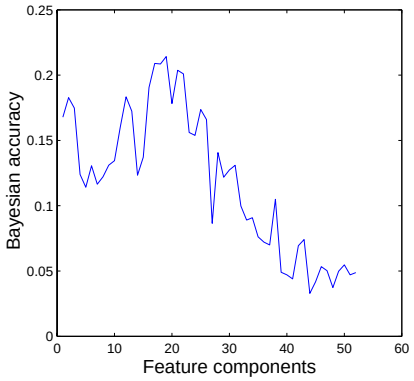

Fig. 2. Bayesian accuracy for different feature components

pool. Then two Bayesian accuracy based approaches are proposed to quantify the discriminative capability for decorrelated feature components.

The first approach adopts the approximation of Bayesian accuracy defined in Equation 3 as the objective function. The second approach adopts the negative sum of the likelihood $Rank$ of the true label $l(t)$ on each data point x_t as the objective function $\mathcal{F}$, as defined in Equation 4. We refer to these two approaches as $Hard\_Bayesian$ and $Soft\_Bayesian$ respectively.

$$\mathcal{F} = -\sum_{t=1}^{T} Rank_t\left(l(t)\right); \tag{4}$$

The derived feature set consists of those decorrelated feature components with high scores on the objective function.

We summarize the process of deriving a feature set for AED in Figure 4.

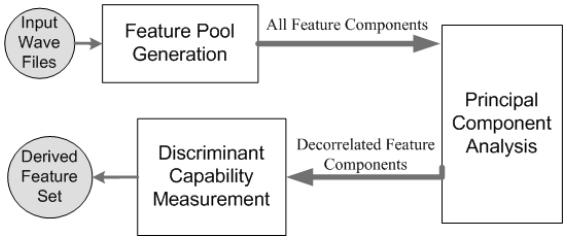

Fig. 3. Feature Analysis & Selection Framework

5. EXPERIMENTS

5.1. HMM-based AED system

For the detection and classification of acoustic events, we implement a hidden Markov model (HMM)-based system, where each acoustic event is modelled by an HMM with three emitting states and left-to-right state transitions. The observation distributions of the states are incrementally-trained Gaussian Mixtures Models with five mixtures. More detailed description of our CLEAR Evaluation HMM-based AED system is available at [10].

5.2. Dataset & metric

Our acoustic event detection experiments use the official data for CLEAR 2007 AED Evaluation [15]: about 3 hours for system development and 2 hours for system evaluation. All data are seminar style, having both speech and acoustic events with possible overlap. Many of the events are subtle and have low SNR compared to background noise or speech. The performances are measured using AED-ACC [15], which is defined as the F-score (the harmonic mean between precision and recall) on system output acoustic event (AE) labels and reference AE labels. AED-ACC aims to score detection and classification of all acoustic event instances, oriented for applications such as real-time services for smart rooms and audio-based surveillance.

5.3. Experiment setup

These experiments compare the performance of single-pass HMM-based AED systems using either one of the derived AED feature or the baseline set MFCC. The AED feature sets are derived using the approaches in Section 4, from a pool of conventional speech perceptual features, i.e. MFCC and log frequency filter bank parameters. All feature sets have identical number (78) of components and all systems have identical number of parameters.

The baseline set MFCC is widely-used in speech recognition and other audio applications. We use 26 MFCCs calculated on 0Hz - 11000Hz band along with their first order regression (delta) coefficients and second order regression (acceleration) coefficients (called MFCC26DAZ).

The first two derived feature sets (DERIVE26DAZ_hard, DERIVE26DAZ_soft) each consists of 26 components derived using $Hard\_Bayesian$ or $Soft\_Bayesian$ approaches in Section 4 from a feature pool of 26 log frequency filter bank parameters and 26 MFCCs. The delta and acceleration coefficients of the above derived feature components are also included. The second two derived feature sets (DERIVE78_hard,DERIVE78_soft) each consist of 78 feature components derived from a feature pool of 26 log frequency filter bank parameters, their delta and acceleration coefficients and all 78 MFCC26DAZ components, using either $Hard\_Bayesian$ or $Soft\_Bayesian$ approaches.

5.4. Experiment results

When training the system, we reserve one third of the three hour development data as Dev set, used to tune some system parameters. Figure 4 shows that the systems using any of the derived feature sets outperform the baseline system both on

Dev and test sets, with a relative AED-ACC improvement of about 30%.

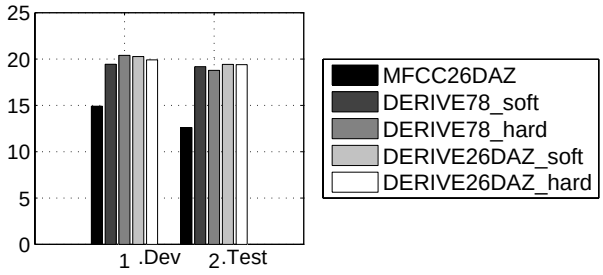

Fig. 4. AED-ACC scores using the baseline feature set MFCC26DAZ and the four derived sets

We also compare the performances when the feature sets are used in our CLEAR Evaluation single-pass HMM-based AED systems [10], trained on all data for system development. Figure 5 shows all derived AED feature sets outperform the baseline. In particular, DERIVE78_soft achieves a relative AED-ACC improvement of over 30%.

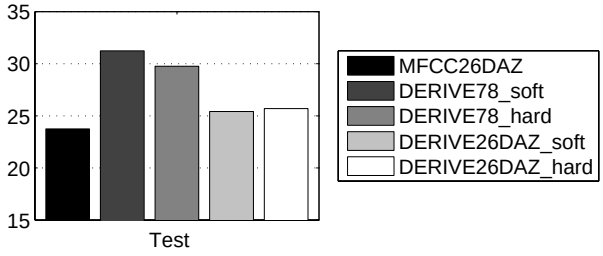

Fig. 5. AED-ACC scores using the baseline feature set MFCC26DAZ and the four derived sets

The above results indicate that for the AED task, the widely-used complete set of MFCCs is far from optimal, and feature sets derived in the proposed approaches could yield better performance in AED without parameter increase.

6. CONCLUSION AND DISCUSSION

In this paper, we propose quantifying the discriminative capability of each (decorrelated) feature component according to the approximated Bayesian accuray and deriving a discriminative feature set for acoustic event detection. We demonstrate the effectiveness of our method on CLEAR AED evaluation task. The proposed feature analysis and selection framework can conveniently derive feature sets for a new task (i.e. acoustic event detection) from a conventional feature pool engineered for a more conventional task (i.e. speech recognition).

7. REFERENCES

[1] Guy J. Brown and Martin Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[2] L. Lu, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 504–516, 2002.

[3] J. Pinquier, "Robust speech / music classification in audio document," in *ICSLP02*, 2002, pp. III: 2005–2008.

[4] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo, "Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems," *IV Jornadas en Tecnologia del Habla, Zaragoza, Spain, November*, 2006.

[5] Rui Cui, Lie Lu, Hong-Jiang Zhung, and Liun-Hong Cai, "Highlight sound effects detection in audio stream," in *ICME03*, 2003, pp. III: 37–40.

[6] Pradeep K. Atrey, Namunu C. Maddage., and Mohan S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *ICASSP06*, 2006.

[7] M. Baillie and J.M. Jose, "Audio-based event detection for sports video," *Lecture Notes in Computer Science*, vol. 2728, pp. 61–65, 2003.

[8] C. Clavel, T. Ehrette, , and G. Richard, "Events detection for an audio-based surveillance system," in *ICME05*, 2005, pp. 1306–1309.

[9] D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT, 1996.

[10] Xi Zhou, Xiaodan Zhuang, Ming Liu, Hao Tang, Mark Hasegawa-Johnson, and Thomas Huang, "HMM-based acoustic event detection with AdaBoost feature selection," in *Classification of Events, Activities and Relationships Evaluation and Workshop*, 2007.

[11] E. D. Scheirer, "Sound scene segmentation by dynamic detection of correlogram comodulation," Tech. Rep. 491, M.I.T Media Laboratory Perceptual Computing Section, Apr. 1999.

[12] H. Hermansky, "Mel cepstrum, deltas, double deltas, .. -what else is new?," in *Proc. Robust Methods for Speech Recognition in Adverse Condition*, 1999.

[13] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Trans. Speech Audio Processing*, vol. 10, pp. 72–83, 1995.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Intersicence, 2004.

[15] Andrey Temko, "CLEAR 2007 AED evaluation plan," http://isl.ira.uka.de/clear07, 2007.