SOUND EVENT CLASSIFICATION BASED ON FEATURE INTEGRATION, RECURSIVE FEATURE ELIMINATION AND STRUCTURED CLASSIFICATION

Huy Dat Tran, Haizhou Li

Institute for Infocomm Research, A*STAR
1 Fusionopolis Way, Singapore 138632
hdtran@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

This paper proposes a novel system for sound event classification based on Feature Integration, Recursive Feature Elimination Support Vector Machine (RFESVM) and Structured Classification. The key points of the proposed method can be summarized as follows: 1) the integration of various feature extraction methods coming from different research communities in one system; 2) the use of feature selection to analyze and select the optimal subset of the integrated features; 3) the adoption of a knowledge-based taxonomic structured classification scheme. Particularly, six groups of features including temporal shape, spectral shape, spectrogram, perceptual cepstral coefficients, harmonic and rhythmic feature sets are investigated in this paper. For the feature selection, the employed RFESVM method enables to select the optimal feature subset taking into account their mutual information. We further develop different feature elimination strategies for RFESVM depending on the requirements of complexity. The RFESVM is combined with a structured classification designed for our task in surveillance and security applications. The proposed method is tested in two realistic environments and the experimental results show good improvements of the classification performance compared to the conventional method.

Index Terms— Sound Event Classification, Surveillance, Security, Structured Classification, Feature Integration, Feature Selection, Recursive Feature Elimination, Support Vector Machine.

1. INTRODUCTION

The Sound Event Classification is a research direction with wide potential for applications in music search, automatic broadcasting, meeting transcription, surveillance, and security [1]-[3]. The Sound Event Classification is a cross-discipline area but has been investigated separately by speech recognition, musical, hearing, and information retrieval communities. The popular systems often employ the well-established speech recognition framework which adopts suitable-for-speech MFCC features and HMM/GMM/SVM classifiers [4]. However, given the differences in the physical natures of sound events which include both speech, music, environment sounds, it is expected that the useful information should not be limited by only speech-related features.

We have been working on the way to integrate useful features from different research communities to a common classification system. In this paper, we propose a method based on Feature Integration, Recursive Feature Elimination Support Vector Machine(RFESVM) and taxonomic structured classification (TCS). The method integrates six feature extraction methods including temporal shape, spectral shape, spectrogram, perceptual cepstral coefficients, harmonic and rhythmic features. The RFESVM feature selection method is then employed. This wrapper feature selection method, originally designed for gene classification problems, is able to select the optimal feature subset when maximizing the classification accuracy. We develop several selection strategies for balancing the trade-off between accuracy and the computation cost and therefore could make the system reliable in realistic applications. Furthermore, we combine the RFESVM to a knowledge-based taxonomic structured classification and their combination show significant improvements on the classification accuracy. The organization of the rest of the paper is as follows. Section 2 describes the integrating feature extraction methods. Section 3 presents the classification system based on RFESVM and structured classification scheme. Section 4 reports experimental results evaluated for sound event classification in realistic environments. Finally, Section 5 summarizes the work.

2. FEATURE INTEGRATION

In this section, we highlight and categorize the feature extraction methods integrated in our system. More details on how to calculate them can be found in [5]-[7].

2.1. Temporal Shape Features

These features carry information on waveform shape and have been used as music descriptors in MPEG-7 [5] and its extension [6].

1. Log-Attack Time: the logarithm of the difference between the start point of event and the time it reaches steady segments. These times are set by fixed thresholds of 20% and 90% on the energy envelop, respectively.
2. Temporal attenuation: the approximated exponential order of the energy envelop attenuation from its peak in time domain and can be fitted by a linear regression to the log-energy.
3. Signal strength: the time period when the energy envelop is more than 40% of its peak.
4. Temporal centroid: the time average over the energy envelop.
5. Temporal kurtosis.
6. Zero Crossing Rate.

2.2. Spectral Shape Features

Spectral Shape Features are other popular features using in MPEG-7 [5] and its extension [6].

1. Spectral centroid: the frequency average over the frame power spectrum.

2. Spectral skewness and kurtosis are calculated on the frame power spectrum distribution.
3. Spectral attenuation: frequency-domain alternative measurement to temporal attenuation.
4. Spectral roll-off: the point at which 95% of frame energy is under this frequency.
5. Spectral flatness measurement(SFM): the ratio between geometric mean to arithmetic mean of power spectrum expressed in dB scale.
6. Tonality indicates how close the given frame to a tonal. This measurement can be derived from SFM as follows

$$tonality = \min\left(\frac{SFM}{60}, 1\right) \quad (1)$$

7. Spectral range is the difference between the maximum and minimum of a frame power spectrum
8. Spectral divergence is the difference between the first and second norms of power spectrum over frames.

2.3. Critical Band-Powers Features

Critical Band-Powers (in logarithm) have been used in studies on human auditory systems. The features are calculated by multiplication of the linear-frequency-scale power spectrum to the filterbank responses centralized in perceptual frequency scales. In this paper we investigate the features in Mel, Bark, and one-third Octave frequency scales which are effectively used in speech recognition, auditory approximation and acoustic measurements, respectively. In addition, we also investigate the auditory-driven Specific Loudness Sensation (SLS) [7] which simulates Terhardt's model of outer-and-middle ear response and Schroeder masking effect.

2.4. Perceptual Cepstral Coefficients

The cesptral features have been demonstrated as the most powerful features for speech recognition. These features are calculated from logarithm of Band-Powers using dct, - a fast empirical decorrelator. Here we investigate the cepstral coefficients in Mel, Bark, and one-third Octave frequency scales. In addition, the Perceptual Linear Prediction Cepstral Coefficients (PLPC) are also investigated. These features differ from MFCC by that the Mel Power Spectrum is smoothed by LPC before computing the cepstral coefficients.

2.5. Harmonic Features

Harmonic features are calculated using the Sinusoidal Harmonic Modeling of speech.

1. Fundamental frequency: estimated using YIN method [8].
2. Harmonicity: the ratio between the energy of the non-harmonic parts to the total energy.
3. Harmonic attenuation: similar to temporal attenuation but calculated using the harmonic spectral amplitudes.
4. Harmonic Spectral Deviation: Deviation of the amplitude harmonic peaks from a global spectral envelop.
5. Harmonic Energy Ratio is calculated from the ratio between the odd harmonic energy to the even one.

	Set features	#	Components
Spectral shape features	Spec	8	Spectral shape features
Band-Powers	MBP	24	Mel-band powers
	BBP	24	Bark-band powers
	OBP	24	Octave-band powers
	SLS	24	Specific Loudness Sensation
Cepstral coeff	MFCC	24	12 MFCC vector &delta
	BFCC	24	12 BFCC vector &delta
	OFCC	24	12 OFCC vector &delta
	PLPC	24	12 PLPC vector &delta
Harmonic feature	Harmon	6	Harmonic features
Temporal features	Temp	6	Temporal shape features
Rhythm features	Flux	14	7 Rhythm features plus delta

Fig. 1. Summary of features

2.6. Rhythm Features

The rhythm features are calculated from the modulation spectrogram (MS), which describes modulations in the loudness calculated in each Mel-frequency band. Modulations around 4Hz are emphasized using a model of perceived fluctuation strength. From MS, 7-rhythm features are calculated in frame-by-frame manner. The features indicate max, sum, bass, aggressiveness, low-frequency domination, gravity, focus properties of MS. The details of rhythm features calculation are given in [7].

The summary of investigated feature sets is illustrated in Fig.1. Excluding the temporal shape features, all features are calculated in frame base. For each feature component, the long-term information over frames is summarized in each clip by the median value. Totally, each audio sample is represented by a 226-dimension feature vector.

3. RECURSIVE FEATURE ELIMINATION SVM AND STRUCTURED CLASSIFICATION

Gathering all features, particularly in realistic noisy environments, should not guarantee the improvements of the classification performance as many features may be "noisy". Moreover, this is impractical as the cost is highly expensive. Here we adopt the RFESVM feature selection method to analyze the features in order to select their optimal subset to be used in the implementation phase.

3.1. Recursive Feature Elimination SVM

We start with the classical binary SVM classifier. consider the input vector X as a concatenation of feature sets described in sec.2,

$$X = [s_1, s_2, ..., s_k]^T. \quad (2)$$

SVM will try to separate data $X \subset \mathbf{R}^k$ from two classes by finding a weight vector $\mathbf{w} \in \mathbf{R}^k$ and an offset $b \in \mathbf{R}$ of a hyperplane

$$\begin{aligned} H : \mathbf{R}^k &\rightarrow \{1, 2\} \\ X &\mapsto \text{sign}(\mathbf{w}.X + b) \end{aligned} \quad (3)$$

which can be found by a constrained optimization expressed by

$$\mathbf{w} = \underset{\substack{y_i(\mathbf{w}x_i + b) \geq 1 - \xi_i^+ \\ (\mathbf{w}x_i + b) \leq -1 + \xi_i^-}}{\arg\min} \left[\frac{\|\mathbf{w}\|^2}{2} + C^+ \sum_i \xi_i^+ + C^- \sum_i \xi_i^-\right] \quad (4)$$

where x_i and y_i are feature and label given from a training database, C and $\mathbf{x}_i$ are regularization parameters.

Fundamental idea of feature ranking is to use the change in objective function when one feature is removed as a ranking criterion. For SVM, the change in the cost function caused by removing a given feature is calculated by setting its weight to zero. Since SVM minimizes $J = \|\mathbf{w}\|^2$ under certain constraints, the optimum J for removing feature index-i is equivalent to minimize w_j^2 and this can be used as feature ranking criterion. The single higher-ranked features may not create a good feature subset. This problem can be overcome by using an iterative procedure called RFE [9]:

1. Train the classifier (optimize the weights respect to J)
2. Compute the feature with smallest ranking criterion;
3. Remove the feature with smallest ranking criterion.

The final number of features are normally determined experimentally. RFESVM method can be generalized for multi-class SVM. The ranking criterion for component$-i$ is then defined as follows

$$J_i = \sum_k \left(\frac{w_i^{(k)}}{\max_j \left[w_j^{(k)} \right]} \right)^2, \qquad (5)$$

where k is index of binary SVM. The number of binary SVMs for OAO-SVM (One-Against-One) and OAA (One-Against-All) are $N(N-1)/2$ and N, respectively.

3.2. Feature Selection Strategies

For computational reasons, it may be more efficient to remove several features at a time at the expense of possible classification performance degradation. Depending on the requirements of the system complexity, two selection strategies are developed as follows:

1. Strategy-1 (optimal but expensive): Elimination is done in feature component level.
2. Strategy-2 (sub-optimal and cheaper): Elimination is first done in feature type level (to down the number of feature sets to a given number, says 3). After that run the strategy-1 to optimize the subset.

3.3. Taxonomic Structured Classification

The tree-structured classification has been widely used in music identification applications [10]. Given the wide variety in the physical natures of sound event creations we also expect that the structured classification could improve the accuracy as well as simplify the system solution in each classification node. Typically, two types of tree-structures can be considered: the automatic generated tree based on mutual information analysis like C4.5 and the knowledge-based taxonomic structure. We will leave the comparison of structures to a future work and this paper just adopts a taxonomic structure scheme considered by physical natures of given sound events and the requirements of the task which aims to recognise aggressive events from audio records. The proposed structured classification for a 10-class sound event classification is illustrated in Fig.2. The audio clip is first classified into voiced and unvoiced. Then the voiced sample is classified into normal and aggressive, respectively. The normal voiced-class consists of speech and laugh events while the aggressive-class can be further classified into cry or scream. Similarly, the aggressive unvoiced is separated into short (dropping, door slam) and non-short (explosion, breaking). An advantage of proposed scheme compared to automatic decision tree approaches is the suitability in the incorporation with RFESVM.

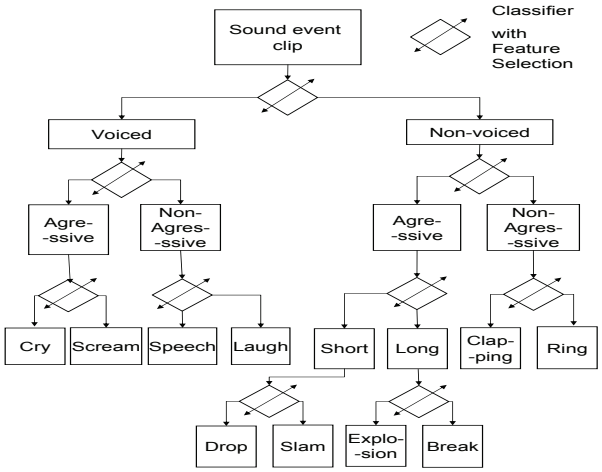

Fig. 2. Flowchart of the proposed classification system

4. EXPERIMENTS

We test the proposed system with a sound event classification task designed for surveillance and security applications. The task is to classify sound clips from 10 classes of normal speech, cry, scream, laugh, knock, break, door slamming, phone ring, and clapping.

4.1. Data collection

Our database consists of about 2.5 hours audio taken from [11]. The audio clips are manually balanced with approximate 2-second lengths. One hour data is used to test and the remaining for training and system calibration. To analyze the methods in realistic environments, the sounds were played and recorded back in two environments: the office environment with A/C noise (about 12-18dB SNR) and the university canteen (about 5dB-10dB SNR).

4.2. Reference methods

The following methods are implemented and compared:

1. Conventional MFCC-SVM: 26-dimensional MFCC (12-MFCC, log-energy from 24-filterbank system and deltas)

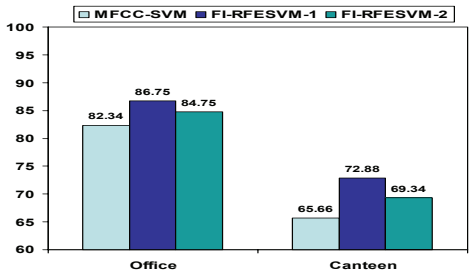

Fig. 3. Effect of RFE on classification accuracy

Method	Office	Canteen
FI-RFESVM-1	MFCC(4), MBP(3), OFCC(4), SLS(2), Temp(4), Spec(5), Harmon (2)	OFCC(8), OFB(4), MFCC(4), MBP (2) Spec(2), Temp(4)
FI-RFESVM-2	MFCC(10), MBP(8), Harmon (6)	OFCC(10),OFB (8), Spec (4)

Fig. 4. Selected features in FI-RFESVM methods

2. Feature Integration plus multi-class RFESVM on two strategies (FI-RFESVM-1 and FIRFESVM-2, respectively);
3. Feature-Integration plus Structured Classification Scheme and RFESVM implemented on two above strategies (FI-RFESVM-S-1 and FI-RFESVM-S-2, respectively).

4.3. Experimental results

4.3.1. Overall results: Effect of Recursive Feature Elimination

Fig.3 shows overall classification accuracies (the ratio between the total number of correctly classified in final step to the total number of clips) of the MFCC-SVM and FI-RFESVM methods. We see improvements up to 6% of overall accuracy in the case of FI-RFESVM. Fig.4 summarizes the selected feature components in FI-RFESVM systems. It is interesting to note that the optimal subset of features for the canteen noise environment does not include the MFCC. The OFCC features look very promising and should be investigated more in future.

4.3.2. Overall results: Effect of Structured Classification

The results of the proposed system with structured classification are shown in Fig.5. We see significant improvements up to 15% of overall accuracy compared to baseline MFCC-SVM. The RFESVM method is well complementary to the proposed taxonomic scheme and this can can be explained by two factors: 1) taxonomy increases the discrimination capacity; 2) the RFESVM is more effective for binary SVM. Fig.6 summaries the contributions of feature sets on the final selected subset for important nodes in structured classification for the FI-RFESVM-S-2 system. This system balances the trade-off between accuracy and the complexity and reliable in the real-time implementation. We can see that these contributions are varying from the sub-tasks to the noise environments. In most cases, the cepstral and band-powers features are most important. However, the spectral shape, temporal shape and modulation-spectrogram-driven features are good complements for particular classification sub-tasks.

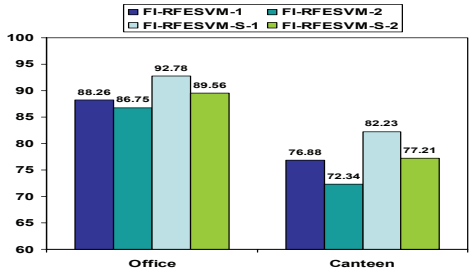

Fig. 5. Effect of Structured Classification on classification accuracy

Method	Office	Canteen
Voiced /Unvoiced	Harmon (4), Spec (4)	OFB (5), Harmon (3)
Voiced aggressive /Voiced non aggressive	MFCC (8), MFB (4), Spec(4)	OFCC (6), OFB (6), Spec(4)
Unvoiced aggressive/ Unvoiced aggressive	MFCC (8), Spec(4), Flux(4)	OFCC (6), OFB (6) Flux(4),
Speech /Laughter	MFCC (10), Harmon (4), Flux (2)	OFCC (6), OFB (6), Flux (4)
Dropping /Explosion &Breaking	MFCC (10), MFB (4), Temp(2)	OFCC (8), OFB (4), Temp(4)
Clapping /Phone ring	MFCC (10), MFB (4), Temp(2)	OFCC(8), OFB(5), Temp(3)

Fig. 6. Selected features in FI-RFESVM-S-2 method

5. CONCLUSION

Here, we present a sound event classification system based on Feature Integration, Recursive Feature Elimination and Structured Classification. The method enables to select the quasi-optimal subset from a huge number of feature extraction methods to improve the classification accuracy. The system is further improved by adopting a knowledge-based taxonomic structured classification.

6. REFERENCES

[1] Hain, T., et al., Segment Generation and Clustering in the HTK Broadcast News Transcription System, *in Proc. DARPA BNTU,* 1998.

[2] J. Foote, Content-based retrieval of music and audio, *Multimed. Storage Archiv. Syst. II*,1997, pp. 138-147.

[3] A. Harma, M.F. McKinney, J. Skowronek, Automatic surveillance of the acoustic activity in our living environment, *in Proc of 2005 IEEE ICME*, pp.634-637, 2005.

[4] A. Temko at. el, CLEAR Evaluation of Acoustic Event Detection and Classification systems, *Lecture Note in Computer Science, Multimodal Technologies for Perception of Humans* edited by Stiefelhagen, R., Garofolo, J., Vol. 4122, 2007.

[5] MPEG-7 Multimedia Content Description Interface Standard, Part 4: Audio, 2002.

[6] G. Peeters, A large set of audio features for sound description (similarity and classification) in the cuidado project, CUIDADO Project Report,2003.

[7] E. Pampalk, Computational Models of Music Similarity and their Application to Music Information Retrieval, Doctoral Thesis, Vienna University of Technology, Austria, 2006.

[8] A. Cheveigne and H. Kawahara. YIN, a fundamental frequency estimator for speech and music, *J. Acoust. Soc. Amer.* v.111, is.4, pp. 1917-1930, 2005

[9] Guyon, I. at. el, Gene selection for cancer classification using support vector machines, *Machine Learning*, vol. 46, pp. 389-422, 2002.

[10] K. D. Martin, Sound-Source Recognition: A Theory and Computational Model, , *PhD thesis, MIT,* 1999.

[11] Sound Effect Collections at http://www.sound-ideas.com/