Spectral Patch Based Sparse Coding for Acoustic Event Detection

Xugang Lu[1]*, Yu Tsao*[2]*, Peng Shen*[1]*, Chiori Hori*[1]

1. National Institute of Information and Communications Technology, Japan
2. Research Center for Information Technology Innovation, Academic Sinica, Taiwan

Abstract

In most algorithms for acoustic event detection (AED), frame based acoustic representations are used in acoustic modeling. Due to lack of context information in feature representation, large model confusion may occur during modeling. We have proposed a feature learning and representation algorithm to explore context information from temporal-frequency patches of signal for AED. With the algorithm, a sparse feature was extracted based on an acoustic dictionary composed of a bag of spectral patches. In our previous algorithm, the feature was obtained based on a definition of Euclidian distance between input signal and acoustic dictionary. In this study, we formulate the sparse feature extraction as l_1 regularization in signal reconstruction. The sparsity of the representation is efficiently controlled via varying a regularization parameter. A support vector machine (SVM) classifier was built on the extracted sparse feature for AED. Our experimental results showed that the spectral patch based sparse representation effectively improved the performance by incorporating temporal-frequency context information in modeling.

Index Terms: Acoustic event detection, sparse coding, support vector machine.

1. Introduction

In real acoustic environments, many types of acoustic events exist, for example, in a lecture room, rather than only speech events made in presentations, other acoustic events, such as background music event, audience's applause and laugh events occur. For automatic speech recognition (ASR), the non-speech events must be removed before doing recognition. The aim of acoustic event detection (AED) is to associate the underlying audio streams to their semantic event categories, and locate their time segments. Designing an efficient AED algorithm is important not only for ASR, but also for many application fields such as audio content analysis and retrieval [1, 2, 3].

The classical framework for AED is to extract the Mel frequency cepstral coefficient (MFCC) feature of audio signal and model the feature with either a hidden Markov model (HMM) or support vector machine (SVM) [4, 5, 6]. This framework is directly inspired by ASR techniques. In ASR, the HMM or SVM is trained with frame based feature representations (e.g, 20 ms frame length) [3, 4, 5]. And speech inherent structure, such as phoneme and word structure, is further modeled to reduce the model confusion caused by frame based feature representation. However, in AED, we do not have knowledge of such kind of inherent structure. Directly mapping the frame based representation to their semantic categories inevitably cause model confusion.

Considering acoustic events have well organized temporal-frequency structure spanned in many continuous frames (as spectral patches), we have proposed a spectral patch based sparse representation for AED [9]. The work was inspired in image pattern classification [10]. In our study, we first learned an acoustic dictionary in which the acoustic words in the dictionary are a bag of spectral patches. Then a sparse feature representation was extracted based on a similarity measurement of the input signal to the learned acoustic dictionary. The sparse feature was used to train SVMs for AED. In our study, the similarity measurement was defined as an Euclidian distance between acoustic signal and acoustic dictionary. And the feature sparsity was controlled by manually setting a threshold to zero-out the small values in the similarity measurement (corresponding to large Euclidian distances).

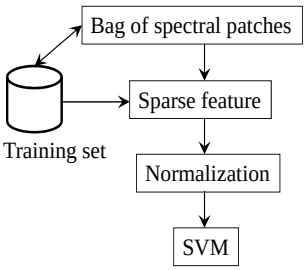

Figure 1: Framework of spectral patched based sparse feature extraction and modeling.

In this study, we further investigate the sparse feature extraction for AED. As mostly used in sparse coding for machine learning [7, 8], we formulate the feature extraction as a l_1 regularization in signal reconstruction. The advantage of using this regularization sparse coding framework is that the feature extraction is well embedded in a mathematical optimization framework that is easy to analyze. And the sparsity and signal reconstruction accuracy can be well controlled via variation of a regularization parameter.

The paper is organized as follows. Section 2 introduces the algorithm of learning acoustic dictionary from spectral patches, and sparse feature extraction and modeling based on the learned acoustic dictionary. Section 3 describes the AED experiments based on the algorithm. Discussion and conclusion is given in last section.

2. Spectral patch based sparse representation and modeling

The whole system for sparse feature extraction and modeling is showed in Fig. 1. In this figure, two steps are involved, one is sparse feature extraction processing, the second is SVM based modeling. In the first step, an acoustic dictionary composed of a bag of spectral patches is learned from a training data set. The learned spectral patches are regarded as prototypes of acoustic dictionary for acoustic event analysis. Based on the dictionary,

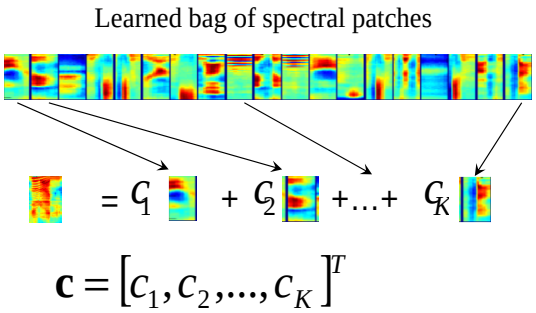

Figure 2: Linear approximation point of view for feature extraction.

sparse coding is applied for sparse feature extraction. The following three subsections describe the procedures in details.

2.1. Learning a bag of spectral patches

In order to reduce confusion caused by feature representation for AED, long-term temporal-frequency structure is incorporated in learning the acoustic dictionary. We have proposed an k-means clustering algorithm on randomly selected spectral patches for learning the acoustic dictionary. The cluster centroids were selected as acoustic words in the dictionary [9]. For completeness, we briefly describe the algorithm here. A training set of spectral patches was first randomly selected from acoustic spectrum. Each spectral patch is composed of several consecutive frames of Mel band spectra [12]. Each spectral patch is concatenated to be a long vector and is used in k-means clustering. Before doing clustering, a normalization is applied on the training data set to remove the difference of each spectral patch caused by absolute density. In addition, a zero-phase component analysis (ZCA) is applied on each spectral patch for data whitening. By this processing, the learning is supposed to focus on high-order statistical structure of the signal. The learned acoustic dictionary is supposed to span a representative space for acoustic event classification.

2.2. Sparse coding with l_1 regularization

After learned the acoustic dictionary, any input signal (spectral patches) can be represented as a linear combination of the acoustic words. In our previous study, we defined a similarity measurement between the input signal and acoustic words based on an Euclidian distance, and set a distance threshold to control the representation sparsity. Although the sparse representation achieved good performance, it is not easy to explain the representation in a solid mathematical way. In this study, we adopt an alternative method to obtain the sparse representation. The idea is to regard the input signal as a linear function of the acoustic words. The feature representation then is regarded as a linear regression (or approximation) problem. In addition, a regularization is added in the function for controlling the representation sparsity and approximation accuracy. The approximation point of view is illustrated in Fig. 2. And the mathematical formulation is given in Eq. 1.

$$\mathbf{c}^* \triangleq \arg\min_{\mathbf{c}} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\mathbf{c}\|_2 + \lambda |\mathbf{c}|_1 , \quad (1)$$

where $\mathbf{x}$ is an input spectral patch, $\mathbf{D}$ is the dictionary composed of a bag of spectral patches (i.e. codebook), $\mathbf{c}$ is the parameter vector which is used as a representation of $\mathbf{x}$ on $\mathbf{D}$. In Eq. 1, the first term is concerned with approximation accuracy, the second term is concerned with sparsity of the representation, and λ is used to control the tradeoff between the approximation accuracy and sparsity. When $\lambda = 0$, the problem is a classical regression problem. The advantage of using Eq.1 for sparse feature extraction is that sparse representation is well coincided in an optimization framework with many efficient algorithms to solve it [7, 8].

2.3. Training classifiers

As shown in Fig. 2, the dimension of feature representation depends on the number of acoustic words. With many acoustic words in the dictionary, the representation is with a high dimensional space. For such a high dimensional feature modeling, a linear SVM is used. For training data pairs as $(\mathbf{c}_i, l_i)$, with $i = 1, 2, ..., N$, where l_i is the label, and $\mathbf{c}_i$ is the sparse feature vector. Multi-class SVMs are built, and each SVM for each acoustic event is constructed as one-against-all with parameter $\mathbf{w}_j$ (the j-the SVM) as [11]:

$$\min_{\mathbf{w}_j} \sum_{i=1}^{N} \left(\max \left\{ 0, 1 - l_i \left(\mathbf{w}_j^T \mathbf{c}_i \right) \right\} \right)^2 + \alpha \|\mathbf{w}_j\|_2^2 \quad (2)$$

The classification can be done by picking up the one which gives the maximum value from all the SVMs as:

$$\hat{l} = \arg \max_{j \in \{1,2,\cdots M\}} \mathbf{w}_j^T \mathbf{c}, \quad (3)$$

where M is the total event number.

3. Experiments

We test our algorithm on audio data of the TED (technology, entertainment, and design) talks [14]. In the TED talks, besides speech, other acoustic events such as music, applause and laugh events exist. For data preprocessing either for ASR task or audio content analysis, acoustic event detection and segmentation must be done. 50 TED talks are chosen as training data set, and 10 TED talks as testing set. On average, each TED talk has about 15 minutes audio data with 16k Hz sampling rate. In training data set, original event labels were manually transcribed. In the transcription, besides pause event, nine event categories are with semantic labels as {speech, applause, cough, laugh, audience, video, music, mix, other}. Among them, "mix event" is a collection of overlapped acoustic events, e.g., applause mixed with laugh. "other event" is a collection of events we have not defined well in our application, such as some moving of chairs in a lecture, or natural environment sounds. As a detection task, performance evaluation metrics are related to false alarm rate and hitting rate. For audio data, these metrics can be frame based, event based or class-wise event based evaluation [1, 3]. In this study, frame based evaluation is used, i.e., frame based Rec (recall), Pre (precision) and F evaluation metrics are defined as follows which are the same as used in [5].

$$\begin{aligned} \text{Rec} &\triangleq \frac{N_{TPR}}{N_R} \\ \text{Pre} &\triangleq \frac{N_{TPE}}{N_E} \\ \text{F} &\triangleq \frac{2 * \text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}} \end{aligned} \quad (4)$$

where N_{TPR} is the number of true positive corresponding to reference events, N_R is the total number of reference events, N_{TPE} is the number of true positive estimated events, and N_E is the total number of estimated events.

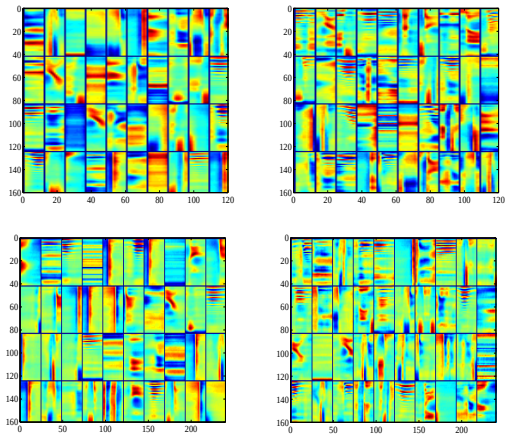

Figure 3: Acoustic words (prototypes of spectral patches) in the learned acoustic dictionary. Patch size: 11 (upper panels) and 23 (lower panels) frames. Codebook size: 128 (left column) and 1024 (right column).

Similarly as we did before, we investigate many factors that may affect the performance of the proposed representation. Before quantitative evaluation, we show some examples of learned acoustic words.

3.1. Visualization of the learned acoustic words

The structure of learned acoustic words depends on the size of spectral patch and codebook size. 40 such acoustic words are shown in Fig. 3. In this figure, each spectral patch is one acoustic word with horizontal axis as time index (in frames) and vertical axis as frequency (Mel frequency band index). These acoustic words can be regarded as prototype temporal-frequency structure to represent acoustic event patterns. From these acoustic words, we can see that acoustic event patterns are composed of many temporal-frequency structure with harmonics, formant frequency transitions, stops and fricatives.

3.2. Effect of number of acoustic words

The number of acoustic words in the learned dictionary is also known as codebook size as used in vector quantization (VQ) [13]. Intuitively, a large number of acoustic words can be much more accurate to represent acoustic event space than using a small number of ones. We did experiments for AED to test the effect of increasing the number of acoustic words in the learned dictionary, and show the results in Fig. 4. In this figure, the patch size is fixed as 11 frames, and regularization parameter in Eq. 1 is $\lambda = 0.5$. From this figure, we see a continuous improvement with increasing of the codebook size. However, the improvement is smaller and smaller with codebook size becomes larger and larger. In real applications, the tradeoff between number of acoustic words and computational complexity must be considered.

3.3. Effect of spectral patch size

Spectral patches with long temporal window is supposed to reduce feature representation confusion compared with short temporal window. However, spectral patches with too long temporal window containing multiple acoustic structure of events may bring pattern confusion. A proper selection of temporal window

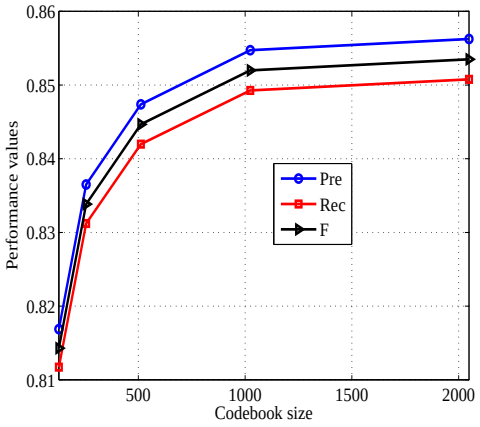

Figure 4: Effect of codebook size.

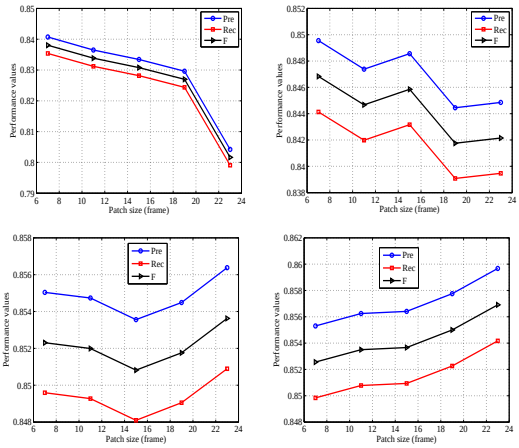

Figure 5: Effect of spectral patch size for different codebook size 256 (upper-left), 512 (upper-right), 1024 (lower-left), and 2048 (lower-right).

should be applied for spectral patch extraction. We did experiments to test the selection of spectral patch size and show results in Fig. 5 (the regularization parameter is set as 0.5). From the figure, we can see that for small codebook size, it seems that small patch size is preferred to get good performance, and for large codebook size, increasing patch size always helps AED performance. We have supposed that there should exist a tradeoff of spectral patch size and AED performance corresponding to small and large codebook sizes. From this figure, however, we can not see a consistent tendency for different selection of codebook size. In the future, we will further check the effect of selection of patch size for AED performance.

3.4. Representation sparsity

The representation sparseness can be controlled by varying λ in Eq. 1. By varying this parameter, the tradeoff between representation accuracy and sparsity can be controlled. An example of the representation is shown in Fig. 6 with codebook size as 128. From this figure, we can see that with increasing of the regularization parameter λ, most of the values of the coefficients approximate to zeros. We did experiments with different λ values for AED, and showed the results in Fig. 7. From this figure, we can see that with different codebook size and tem-

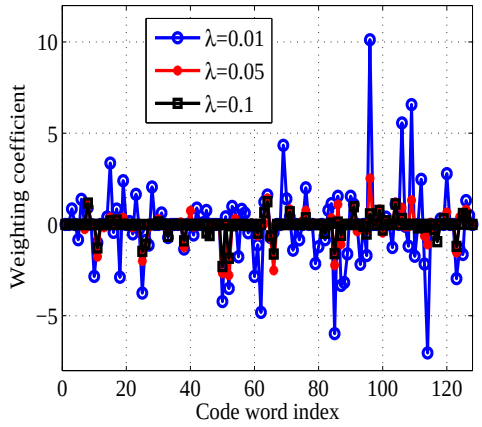

Figure 6: Representation sparsity: coefficient corresponding to each acoustic word (codebook size 128).

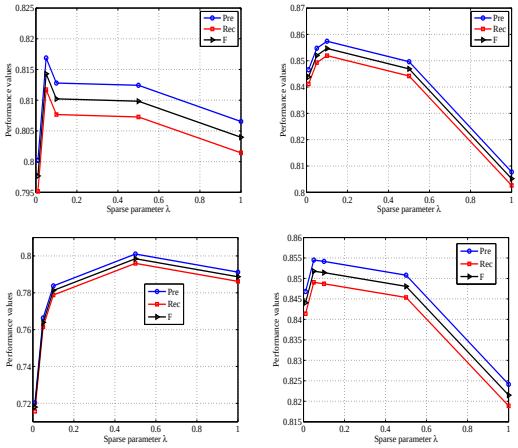

Figure 7: Sparsity with AED performance. Spectral patch size 11 (upper row) and 19 (lower row); codebook size 128 (left column) and 1024 (right column).

poral window of spectral patch, the best performance is with different sparsity parameter. This also suggest that the AED performance depends on the a combination of codebook size, temporal window of selected spectral patch and sparsity parameter.

4. Conclusion and discussion

In this study, we explored acoustic pattern distributed in temporal-frequency of the signal for AED. In order to extract long-temporal information, an acoustic dictionary composed of a bag of spectral patches were learned. Based on the learned acoustic dictionary, a sparse representation was extracted as an l_1 regularization for signal reconstruction. The sparseness of the representation was well controlled with consideration of spectral reconstruction accuracy. With the sparse representation, we built an SVM based AED system. Based on the system, we discussed several important factors that affect the AED performance.

The problem of acoustic event detection is how to extract efficient representations that have discriminative information among different patterns. Differently from speech, currently we are lack of knowledge of how acoustic events are organized from basic acoustic units to their semantic categories. Directly mapping the low level acoustic feature to their high level abstraction may cause large confusion due to representation confusion in low level representation. The automatic representational feature learning is suitable in finding an appropriate middle level representation for AED. In addition, classification is different from reconstruction, some factors important for reconstruction are not necessarily important for classification. Based on this consideration, the learning can be constraints with some criterion that may bring advantages for AED. In this study, the sparsity constraint is applied for that purpose. Although, currently we are not clear how sparse of the representation is optimal for AED, we believe that sparse constraint may help to keep major signal structure that is relevant for AED. In addition, with different size of temporal window and number of acoustic words, the sparsity may have different effects on AED. In the future, we will further investigate the relationship between these factors for AED.

5. References

[1] D. Giannoulisy, E. Benetosx, D. Stowelly, M. Rossignolz, M. Lagrangez and M. Plumbley, "Detection and Classification of Acoustic Scenes and Events: an IEEE AASP Challenge," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[2] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 1, pp. 1-13, 2013.

[3] X. Zhuang, X. Zhou, M. A. Hasegawa-johnson, T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543-1551, 2010.

[4] C. Zieger, "An HMM based system for acoustic event detection," *Multimodel technologies for perception of humans*, pp. 338-344, 2008.

[5] A. Temko, C. Nadeu, and J. I. Biel, "Acoustic Event Detection: SVM-Based System and Evaluation Setup in CLEAR'07," *Multimodel technologies for perception of humans*, pp. 354-363, 2008.

[6] Z. Huang, Y. Cheng, K. Li, V. Hautamaki, C. Lee, "A Blind Segmentation Approach to Acoustic Event Detection Based on I-Vector," in *Proc. Interspeech*, pp. 2282–2286, 2013.

[7] M. Aharon, M. Elad and A. Bruckstein, K-SVD,"An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol 54, no. 11, pp. 4311-4322, 2006.

[8] J. Mairal, F. Bach, J. Ponce and G. Sapiro, "Online Dictionary Learning for Sparse Coding," *International Conference on Machine Learning*, Montreal, Canada, 2009

[9] X. Lu, Y. Tsao, S. Matsuda, C. Hori, "Sparse representation based on a bag of spectral exemplars for acoustic event detection," ICASSP, Italy, 2014.

[10] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. the 14-th International Conference on AI and Statistics*, 215-223, 2011.

[11] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871-1874, 2008.

[12] X. Lu, Y. Tsao, S. Matsuda, C. Hori, "Speech Enhancement Based on Deep Denoising Autoencoder," INTERSPEECH, Aug. 26, 2013, Lyon, France.

[13] F. Soong, A. Rosenberg, L. Rabiner, B. Juang, "A vector Quantization approach to Speaker Recognition," ICASSP, 1: 387-390, 1985.

[14] TED http://www.ted.com/