Audio Engineering Society

Convention Paper 10395

Presented at the 149th Convention
Online, 2020 October 27–30

Overlapping Acoustic Event Detection via Perceptually Inspired the Holistic-based Representation Method

Hyeonsik CHOI[1], Keunsang LEE[1], Minseok KEUM[2], David Han[3], Hanseok KO[4]

[1]*Artificial Intelligence Lab., Future Tech center, LG Electronics., South Korea*

[2] *SELVAS AI, South Korea*

[3] *Army Research Laboratory, Adelphi, MD USA*

[4] *Korea University, Seoul, South Korea*

Correspondence should be addressed to Author (hyeonsik.choi@lge.com)

ABSTRACT

A novel dictionary learning approach that utilizes Mel-scale frequency warping in detecting overlapped acoustic events is proposed. The study explored several dictionary learning schemes for improved performance of overlapping acoustic event detection. The structure of NMF for calculating gains of each event was utilized for including in overlapped signal for its low computational load. In this paper, we propose a method of frequency warping for better sound representation, and apply dictionary learning by a holistic-based representation, namely nonnegative K-SVD (NK-SVD) in order to resolve a basis sharing problem raised by part-based representations. By using Mel-scale in a dictionary learning, we show that the information carried by low frequency components more than high frequency components and dealt with a low complexity. Also, the proposed holistic-based representation method avoids the permutation problem between another acoustic events. Based on these benefits, we confirm that the proposed method of Mel-scale with NK-SVD delivered significantly better results than the conventional methods.

1 Introduction

Real acoustic environments are generally filled with complex sound fields generated by multiples of different sources at different locations. Often in acoustic signal processing, such as event detection or event classification, separating signals from different sources is an important step. One of the earliest source separation methods is of spectral type [1]. Performance of these frequency-based methods often degrades dramatically when these sources overlap in their frequency regions.

In recent years, effectiveness of matrix decomposition methods for dictionary learning in overlapping acoustic event detection has sparked interests among researchers [2-3]. By using these methods, a given data matrix can be decomposed into a product of basis and weight matrices forming a salient structure from raw data. Among the matrix decomposition methods, such as principal component analysis and vector quantization, Non-negative Matrix Factorization (NMF) is widely used to detect overlapping acoustic events [4-6]. NMF, however, may not converge in cases when the number of components present in the sound is not accurately matched with the number of NMF bases. K-SVD is an effective alternative in cases when the number of acoustic components is not well known in advance [7]. However, according to Bertin et al, computational load of K-SVD is about an order of magnitude higher than that of NMF [8]. This is obviously a special case since these techniques were applied to musical sounds of a piano. The number of bases present in such sounds is finite and NMF can be effective in generating sufficient number of bases for capturing key characteristics of the sounds. When

applied to non-musical sounds, NMF may not be as effective since the number of bases required can be quite large and difficult to estimate. Since each basis generated by NMF represents a certain part of the spectrogram, a threshold number of bases is necessary to adequately characterize energy structure of a spectrogram. When the number of NMF bases is less than the threshold, the decomposition would result incomplete representation of the sound. To resolve this part-based representation issue, we propose a holistic-based approach for dictionary learning, namely the nonnegative K-SVD (NK-SVD) [7]. Additionally, by using Mel-scale features, the human auditory system was exploited to be well known for its ability of easily distinguishing or interpreting a variety of acoustic events in different acoustic backgrounds [9].

2 Proposed Method for Overlapping Acoustic Event Detection

The overall system for detecting overlapping acoustic event consists of three parts as depicted in Fig. 1.

In the first part, a nonnegative dictionary from training data was constructed for each acoustic event as a training procedure. As observed earlier, an auditory system recognizes acoustic events non-linearly in frequency domain. Therefore, we propose a nonnegative dictionary learning method based on Mel-scale which extracts relevant features nonlinearly in frequency domain. Thus, higher resolution is applied in lower frequency regions that contain more salient frequency changes. It was based on the known evidence that the information carried by low frequency components of the speech signal is phonetically more important than carried by high frequency components. However, unlike a speech signal, a high frequency signal could cause the performance deterioration.

Next, gain or weight is calculated for a given mixed acoustic input by NMF method as a source decomposition procedure.

As the third step, the decomposed signal for each event is integrated and the final decision is delivered. A more detailed description is given in the following section.

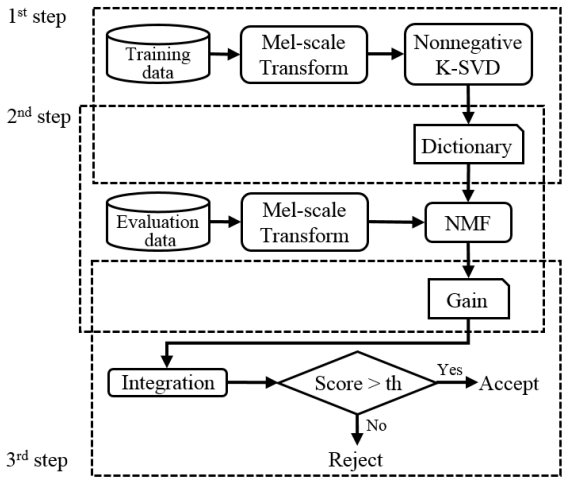

Figure 1. Proposed system diagram.

2.1 Nonnegative dictionary learning methods

In overlapping acoustic event detection, NMF is used mainly for constructing a dictionary representing characteristics of the input data. It delivers excellent results in extracting an arbitrary number of sources from monophonic signals. Given a nonnegative data matrix $\mathbf{V}$ ($m \times n$), NMF finds an approximate factorization of nonnegative factors $\mathbf{D}$ ($m \times r$) and $\mathbf{X}$ ($r \times n$) as

$$\mathbf{V} \approx \mathbf{DX} \tag{1}$$

where $r \le m$ while the reconstruction error is minimized. A cost function measures the reconstruction error between the original matrix and the product of the NMF factors. One form of cost function is generalized Kullback-Leibler divergence, which is denoted as follows.

$$c(\mathbf{D},\mathbf{X}) = \left(\sum_{i=1}^{m} \sum_{t=1}^{n} \mathbf{V}_{i,t} \log \frac{\mathbf{V}_{i,t}}{(\mathbf{DX})_{i,t}} - (\mathbf{V}-\mathbf{DX})_{i,t} \right) \tag{2}$$

For minimization of cost function, NMF iteratively modifies $\mathbf{D}$ and $\mathbf{X}$ using multiplicative update rules.

As it can be seen from the above, NMF process does not allow negative entries in the matrix factors. These non-negativity constraints only permit additive combinations. For these reasons, it called that NMF is a method for finding a part-based representation [10].

However, due to its so-called part-based representation, resultant dictionary of each acoustic event shares common basis that degrades detection performance [10]. To resolve this common basis sharing problem, a holistic-based representation method was incorporated, namely the Spherical K-Means (SKM) [11] and NK-SVD [7].

SKM is represented as the closest dictionary, K-means algorithm used by the cosine distance instead of Euclidean distance [11]. The objective of SKM clustering is to maximize the average cosine similarity

$$\min_{\mathbf{D},\mathbf{X}} \left\{ \left\| \mathbf{V} - \mathbf{D}\mathbf{X} \right\|_F^2 \right\} \text{ subject to } \forall_i,\ \mathrm{x}_i = \mathrm{e}_i \text{ for some } k. \quad (3)$$

Also, K-SVD algorithm, which is a generalization of K-means algorithm, adjusts sparsity, T_0, for controlling relative magnitudes of the two decomposed matrices as shown in Eq. (4).

2.2 The importance of holistic based representation in acoustic signal Page Headers

Like NMF, K-SVD is an iterative method that alternates between sparse coding and update process for the dictionary elements to better fit the data. In each iteration there are two stages: one for sparse coding that evaluates **X** and one for updating the dictionary **D**. In the sparse coding stage, any pursuit algorithm can be used to compute the representation vectors x_i for each example v_i, by an approximate solution of

$$\min_{\mathbf{D},\mathbf{X}} \left\{ \left\| \mathbf{V} - \mathbf{D}\mathbf{X} \right\|_F^2 \right\} \text{ subject to } \forall_i,\ \left\| \mathrm{x}_i \right\|_0 \le T_0. \quad (4)$$

In the dictionary update stage, both **X** and **D** are assumed to be fixed, and a penalty term of an objective function defined as follows is minimized.

$$\begin{aligned} \left\| \mathbf{V} - \mathbf{D}\mathbf{X} \right\|_F^2 &= \left\| \mathbf{V} - \sum_{j=1}^{K} \mathrm{d}_j \mathrm{x}_T^j \right\|_F^2 = \left\| \left(\mathbf{V} - \sum_{j \neq k} \mathrm{d}_j \mathrm{x}_T^j \right) - \mathrm{d}_k \mathrm{x}_T^k \right\|_F^2 \\ &= \left\| \mathbf{E}_k - \mathrm{d}_k \mathrm{x}_T^k \right\|_F^2 \end{aligned} \quad (5)$$

where K is the number of bases, d_k is the kth column of the dictionary, x_T^k is the kth row of **X**, and E_k stands for the error for all the N frames when the kth element is removed. Revised d_k and x_T^k can be acquired by SVD applied to E_k. In NK-SVD, any negative value calculated in the error matrix is set to zero in the iteration. Note that when the sparsity T_0 is set to 1, the decomposition becomes K-means.

K-SVD is an algorithm for training an overcomplete dictionary that best suits a set of given signals [7]. Due to the dictionary being overcomplete, this algorithm represents holistic bases that composed dictionary of events.

$$\mathbf{D}_l = \left[\mathbf{d}_{l:1}, \mathbf{d}_{l:2}, ..., \mathbf{d}_{l:r} \right] \text{ subject to } \mathbf{d}_l \neq \mathbf{d}_i + \mathbf{d}_j \quad (6)$$

where $\mathbf{D}_l$ is the dictionary of the l th event, $\mathbf{d}_l$ is basis vectors of the l th event and $\mathbf{d}_i$, and $\mathbf{d}_j$ is an arbitrary basis vector.

For acoustic event detection, discernibility between dictionaries of different events is essential. Naturally, if a dictionary of a particular event shares basis with a different event, detection performance may degrade. To validate the holistic-based representation, it is shown in Fig. 2 that the basis vector is in conformity with the dictionary learning methods. The first basis vector of NK-SVD based dictionary fully represents harmonics of a speech for a given frames. On the other hand, the NMF based dictionary exhibits the basis sharing problem raised by the part-based representation in the first and ninth basis vector of NMF. This constitutes a problem that includes other dictionaries that have the same basis, and is a phenomenon that reduces performance. The discriminability of a dictionary between dictionaries of different events is very important.

SKM as a holistic-based representation technique shows better performance than the part-based representation, although its performance is with some limitations. SKM generates basis vectors by the voiceless and microphone mute, we can be seen by the seventh and eighth basis vector of SKM in Fig. 2 (b). Because these basis vectors are poor at representing harmonics, SKM can be regarded as worse basis vector than NK-SVD.

Figure 2 shows a dictionary obtained by applying each dictionary learning algorithm with respect to a speech of a single speaker. Test samples were used for about 8 seconds of Women's reciting voice DB of ETRI 2002, the number of a basis vector for every algorithm was set to 10. Also, the basis vector was

arranged in the order of high contribution to configure a large energy and occurrence.

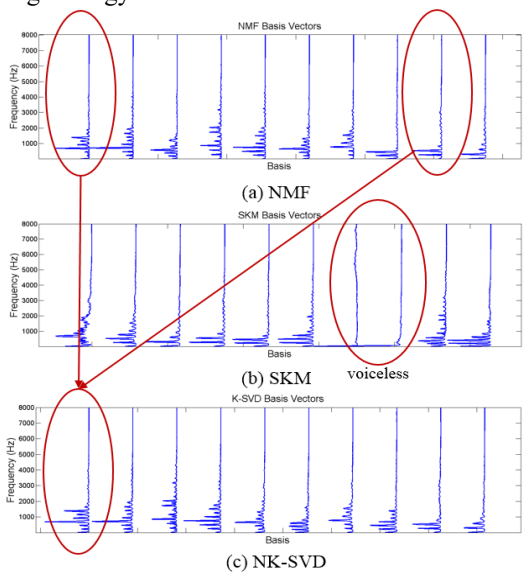

Figure 2. Basis vectors obtained by dictionary learning methods of (a) NMF, (b) SKM and (c) NK-SVD.

2.3 **Signal decomposition and final decision**

NMF was utilized as a framework for calculating gain of each event included in overlapped signal. Although there are various sparse coding methods for calculating gain in given signal, such as basis pursuit [12], and orthogonal matching pursuit [13], their computation load is very demanding compared to NMF. It has been shown that their computation time is roughly 10 times greater than that of the NMF which utilizes simple matrix multiplication [8]. A dictionaries of each event are concatenated by column-wise where each column contains a spectrum corresponding to one of the sources. A fixed dictionary obtained by different dictionary learning methods as mentioned earlier can be readily applicable to NMF for calculating gain. Then, update rule is applied for gain while fixing the concatenated dictionary matrix. Final decision scheme is presented in

$$p(w_l \mid \mathbf{X}_l)\sum_{j=1}^{r}\mathbf{x}_{l:j} > \theta_l \qquad (7)$$

where $\mathbf{x}_{l:j}$ is a gain of l th event, j is the number of basis, θ_l is threshold of l th event and $p(w_l|\mathbf{X}_l)$ is conditional probability of weighting factor w_l given event X_l. The gain represents activations of basis for each event along temporal line, and score of each event is calculated by summation of the gains that belong to each event category. When the score of an event in a frame exceeds a predefined threshold, the frame is regarded as that the event has occurred.

3 Experimental result and analysis

3.1 **Database**

Database from the Audio and Acoustic Signal Processing (AASP) Challenges was used [14]. There were 16 event classes that may occur in office environments. (Table 1) All of the acoustic signals were recorded at a sampling rate of 16 kHz using 16-bit stereo-channel format, and equally mixed two channels to create mono-channel signal. The training dataset contains 20 examples for each acoustic event. The total recording length of an event ranges from 1 to 3 minutes. The evaluation dataset consists of 9 artificial samples created by concatenated overlapping acoustic events (subtask_OS). The dataset contains signals with various SNRs (-6dB, 0dB, 6dB) with respect to the background noise and different levels of density (high, med, low) of acoustic events. Experiments was conducted on the high density evaluation data with all of SNRs.

Alert	Clear throat	Cough	Door slam
Drawer	Keyboard	Keys	Knock
Laughter	Mouse	Page turn	Pen drop
Phone	Printer	Speech	Switch

Table 1. Sixteen classes of acoustic event.

3.2 **Experimental setting and result analysis**

Experiments were conducted on overlapping event detection by the proposed Mel-scale transformed and original spectrograms. Also linear-scale transform was conducted in order to verify the effect of dimension reduction by using the filterbank.

		Original			Linear-scale transform			Mel-scale transform		
		F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall
NMF	SNR 6	23.6833	25.1246	22.3985	22.1397	23.6482	20.8122	26.8709	29.7610	24.4924
	SNR 0	22.3322	22.8916	21.7995	21.7802	23.6712	20.1691	25.7143	29.4393	22.8261
	SNR -6	20.5484	21.2329	19.9066	18.7161	20.6629	17.1045	22.4591	25.2012	20.9574
SKM	SNR 6	24.5264	26.8274	22.5888	22.3370	22.7332	21.9543	27.6456	30.2642	25.4442
	SNR 0	23.1083	24.0680	22.2222	22.6332	23.5869	21.5869	26.5610	27.9520	25.3019
	SNR -6	21.2363	21.5228	20.9574	19.8413	20.7933	18.9726	24.4591	25.2012	23.7595
NK-SVD	SNR 6	**27.8473**	**26.2655**	**29.6320**	**23.0847**	**24.6575**	**21.7005**	**32.2413**	**32.9794**	**31.5355**
	SNR 0	**24.1955**	**23.8824**	**24.5169**	**22.6525**	**24.3243**	**21.1957**	**30.0236**	**26.5155**	**34.6014**
	SNR -6	**22.6066**	**23.1098**	**22.1249**	**20.9688**	**23.0877**	**19.2061**	**26.3021**	**23.7424**	**29.4804**

Table 2. Comparison of Performance.

Different settings of the dictionary learning method and SNR value were used for evaluating performance of the proposed method. The FFT size was set to 1024 with 50% overlap, which results in 513 dimensional spectrogram. Each frame is multiplied by a Hamming window before applying FFT. The number of filters was empirically set to 80 as a trade-off between representation power and generalization. The number of basis for each event was set to 40.

F-score is used as a measure for performance evaluation, which is the harmonic mean of precision and recall. Because of the trade-off relation between precision and recall, F-score can vary by different setting of the threshold. For evaluating the effectiveness of the proposed method regardless of setting the threshold, we set the threshold for oracle performance, i.e. the best result that a given system can achieve.

Table 2 shows the performance of the proposed method compared to two other conventional dictionary learning methods, namely NMF and SKM. Regardless of the methods, the dictionary acquired by Mel-scale transform gives best detection performance. Comparing with the linear-scale transform, it can be inferred that the performance improvement is not due to the dimensionality reduction but because of the non-linear frequency warping by the Mel-scale. When comparing the proposed NK-SVD to the other methods, the proposed performed best among the three methods. The poor performance by the NMF method was attributed mainly due to the common basis sharing problem caused by part-based representation nature intrinsic in NMF method. The proposed NK-SVD performed better over SKM because it can discern frames with higher energy (containing more relevant information) more effectively.

4 Conclusions

In this paper, we proposed a nonnegative dictionary learning method inspired by human auditory perception. We confirmed that Mel-frequency warping applied to spectrogram gives consistent performance improvement over the conventional dictionary learning methods. Thus, higher resolution is applied in lower frequency regions that contain more salient frequency changes. However, unlike a speech signal, a high frequency signal could cause the performance deterioration.

The proposed NK-SVD based dictionary learning showed the best performance because it can discern frames with higher information content more effectively and also it does not lead to the part-based representation issue inherent in NMF. The NMF structure based gain calculation led to the overall lower computational requirements. Also, the proposed holistic-based representation method avoids the permutation problem between another acoustic events.

References

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “A blind source separation technique using second-order statistics,” IEEE Trans. Signal Processing, vol. 45, pp. 434—444, (1997).

[2] I. Tosic and P. Frossard, “Dictionary learning,” IEEE Signal Process. Mag., vol. 28, no. 2, pp. 27-38, (2011).

[3] Bae, Soo Hyun, Inkyu Choi, Hyung Yong Kim, Kang Hyun Lee, and Nam Soo Kim. "Overlapping acoustic event classification

based on joint training with source separation." In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1730-1734. IEEE, (2017).

[4] J. F. Gemmeke, L.Vuegen, P. Karsmakers, and B.Vanrumste, “An exemplar-based NMF approach to audio event detection,” In Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on pp. 1—4, (2013).

[5] J. Ludeña, and A. Gallardo-Antolín. “NMF-based temporal feature integration for acoustic event classification,” INTERSPEECH, p.2924—2928, (2013).

[6] S. Innami and H. Kasai, “NMF-based environmental sound source separation using time-variant gain features,” Computers & Mathematics with Applications, vol. 64, no. 5, pp. 1333—1342, (2012).

[7] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for Designing Overcomplete Dictionaries for Sparse Representation.” Signal Processing, IEEE Transactions on, vol.54, no.11 pp. 4311—4322, (2006).

[8] N. Bertin, R. Badeau, and G. Richard, “Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark,” in Proc. Int. Conf. Acoust., Speech, Signal Process, Honolulu, pp. 65——68, (2007).

[9] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli. “Summary statistics in auditory perception,” Nature neuroscience 2013, vol.16, no.2 pp.493—498, (2013).

[10] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, pp. 788—791, (1999).

[11] S. Zhong, “Efficient online spherical k-means clustering”. In Proceedings of IEEE Int. Joint Conf. Neural Networks, Montreal, Canada, pp. 3180-3185, Aug. (2005).

[12] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” SIAM Rev., vol. 43, no. 1, pp. 129—159, (2001).

[13] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in Proc. 27th Annu. Asilomar Conf. Signals, Systems and Computers, Pacific Grove, CA, pp. 40-44, (1993).

[14] G. Dimitrios. B. Benetos, “Detection and classification of acoustic scenes and events: An IEEE AASP challeng,”, 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, (2013).