World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:7, No:8, 2013

A Dictionary Learning Method Based On EMD for Audio Sparse Representation

Yueming Wang, Zenghui Zhang, Rendong Ying, and Peilin Liu

***Abstract*—**Sparse representation has long been studied and several dictionary learning methods have been proposed. The dictionary learning methods are widely used because they are adaptive. In this paper, a new dictionary learning method for audio is proposed. Signals are at first decomposed into different degrees of Intrinsic Mode Functions (IMF) using Empirical Mode Decomposition (EMD) technique. Then these IMFs form a learned dictionary. To reduce the size of the dictionary, the K-means method is applied to the dictionary to generate a K-EMD dictionary. Compared to K-SVD algorithm, the K-EMD dictionary decomposes audio signals into structured components, thus the sparsity of the representation is increased by 34.4% and the SNR of the recovered audio signals is increased by 20.9%.

***Keywords*—**Dictionary Learning, EMD, K-means Method, Sparse Representation.

I. Introduction

SPARSE representation has been studied since 1990s, which is aimed to represent signals with only a few elementary components. The components are called atoms. A dictionary is formed with a number of atoms. Whether a dictionary can successfully represent signals with sparse decompositions depends on the dictionary used and whether it matches the signal features.

To obtain a dictionary within a sparse decomposition, there are two main methods: dictionary selection and dictionary learning. Dictionary selection is to choose an existing dictionary which can best match the signal features. Such existing dictionaries include Fourier basis, modified discrete cosine basis, wavelet basis and constructed redundant or overcomplete dictionaries (Gabor dictionary, union bases formed by Fourier basis and wavelet basis and so on). Dictionary learning, on the other hand, aims to deduce the dictionary from training signals so that the learned dictionary matches the features of the training set. The sparse representation results are obtained through an alternating optimization strategy and the sparse decomposition is fixed once the dictionary is learned. The uniqueness of the result guarantees the stability of the method.

Compared to dictionary selection, dictionary learning method is more flexible and can meet different specific demands; therefore dictionary learning method has been widely studied ever since early dictionary learning methods are proposed. Research in dictionary learning can be categorized into three main directions: the probabilistic learning methods, the clustering or vector quantization based methods, and the particular construction based learning methods.

Early dictionary methods which are based on a probabilistic learning method are proposed by Olshausen and Field [1] and Lewicki and Sejnowski [2]. In [1], Olshausen and Field proposed a maximum likelihood dictionary learning method for natural images, which is called sparse coding. In [2], Lewicki and Sejnowski clarify the relation between sparse coding methods and independent component analysis (ICA). The method of vector quantization (VQ) achieved by K-means clustering is a typical example of clustering based sparse represent method. Schmid-Saugeon and Zakhor [3] proposed the VQ approach for dictionary learning in MP based video coding. In 2006, Aharon et al [4] proposed the K-means Singular Value Decomposition algorithm (the K-SVD algorithm), which shows excellent performance in practice and is widely used in image denoising. Many applications use dictionaries generated by a set of functions and such functions have similar forms but different parameters. M. Yaghoobi [5] proposed the parametric method and applied it to a Gammatone generating function. In [5], an optimized Gammatone parametric dictionary close to an equiangular tight frame (ETF) is given and it has better coherence properties than the original Gammatone filter bank.

Dictionary learning methods have been widely studied, especially those applied to images. However, dictionary learning methods appropriate for audio signals are not as well developed as those for images. Meanwhile, dictionaries mentioned above are not provided with audio structures. Gabor redundant dictionary, Gammatone parameter dictionary [5], K-SVD algorithm [4] and the latest proposed Greed Adaptive Dictionary (GAD) [6] are suitable for audio signals, but cannot achieve a satisfying recovery result.

Thus, in this paper we propose a dictionary learning algorithm employing Empirical Mode Decomposition (EMD) to decompose audio signals into trend signals and Intrinsic Mode Functions (IMFs) which possess structural property. To reduce the size of the IMF dictionary, a K-means method is applied.

The advantage of proposed approach is that it can obtain better sparse decomposition of audio signals compared to

This work was supported by the National Natural Science Foundation of China under Grant No. 61171171, and the 973 National Basic Research Program of the Ministry of Science and Technology of China under Grant No. 2010CB731904.

Yueming Wang is with the Departmentof Electronic Engineering, Shanghai Jiao Tong University, Shanghai, CO 200240 PRC (phone: +86-188-1751-9397; e-mail: wymatcn@sjtu.edu.cn).

Zenghui Zhang, Rendong Ying and Peilin Liu are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, CO 200240 China. (e-mail:zenghui.nudt@gmail.com, RDying@sjtu.edu.cn, liupeilin@sjtu.edu.cn).

K-SVD because EMD can effectively decompose audio signal into a small number of IMFs. For example, if we are training the dictionary using audio signal set consist of frames with length of 256 points, only 7 IMFs and 1 residual are required to recover the original signal, which results a compression ratio of about 4%. Though the compression ratio may be larger than 4% in practice, it is still considerable enough compared to existing coding methods.

II. Dictionary Learning Method Based On EMD

The goal of sparse representation is to decompose a given signal $x \in R^N$ into a linear combination of a small number of signals, which is called the dictionary. Signals in a dictionary are a set of unit norm functions, called atoms. Denote the dictionary as $\Phi \in R^{M \times N}$ and the atoms as $\phi_k \in R^N$ where $k = 1, 2, ..., M$. Each row of Φ is an atom. A given signal x can be represented as a linear combination of atoms in the dictionary, i.e.

$$x = \Phi^T \mathrm{A} = \sum_{k=1}^{M} \alpha_k \phi_k \tag{1}$$

where α_k is the kth row of the matrix $\mathrm{A} \in R^M$.

The dictionary is overcomplete ($M > N$) when it spans the signal space and its atoms are linearly dependent. When it is overcomplete, the decomposition is not unique. To achieve efficient and sparse representations, we generally look for a sparse representation with an approximation error η of bounded energy ε instead of finding the exact representation. The purpose of sparse representation is to find decomposition with a small number of significant atoms while the rest of the coefficients are close or equal to zero. The optimization problem can be described as follows:

$$\min_{\alpha} \|\alpha\|_0 \text{ s.t. } x = \Phi^T \alpha + \eta \quad where \|\eta\|_2^2 < \varepsilon. \tag{2}$$

However, the problem is NP-hard. In this paper, the Iterative Hard Thresholding (IHT) algorithm is adopted to achieve the decomposition, thus the optimization problem is changed into follow description:

$$\min_{\alpha} \left\| x - \Phi^T \alpha \right\|_2^2 \text{ s.t. } \|\alpha\|_0 \le s. \tag{3}$$

where s is the number of non-zero coefficients of α .

A. Empirical Mode Decomposition Technique

Empirical Mode Decomposition (EMD) [7] was proposed by N. E. Huang in 1998. It's a method to separate data into different components according to their scales and is used to extract variations form the data by separating the mean from the fluctuations by using spline fits. The EMD method is illustrated as follows,

Algorithm 1: Empirical Mode Decomposition

1: Initialization:
Initialize $k = 1$
Obtain the length of x_0 and denote it by L
Obtain the total number of components of one frame, i.e. $K = floor(\log_2(L))$
2: while $k \le K$ do
3: Identify all the local extrema of x_{k-1}
4: Connect local maxima by a cubic spline line to form an upper envelope en_{upper_k} and the local minima to form a lower envelope en_{lower_k}
5: $m_k = (en_{upper\_k} + en_{lower\_k})/2$
6: $x_{k+1} = x_k - m_{k+1}$
7: $k = k + 1$
8: end while

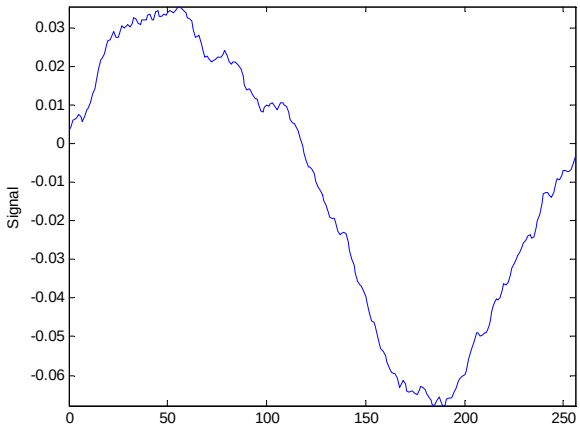

Fig. 1 The original signal

The different components extracted by EMD are Intrinsic Mode Functions (IMFs) and residual components. An IMF is a function satisfies two conditions:

1) The number of extreme points and that of zero crossings must be equal or differ by one;
2) The mean value of the envelope defined by the local maximum value and the envelope defined by the local minimum value is zero.

While the residual components indicate the trend of the signal, which is in most cases smooth and simple.

An example audio signal is shown in Fig. 1, while the components of the audio signal drawn using EMD are given in Fig. 2.

Fig. 2 indicates that the components become smooth with the increase of scale. The first obtained IMF captures the plentiful structure information under the finest scale of signal while the rest IMFs capture structure information under finer scale of signal. The residual component is non-structure information.

B. K-means Method

As we use IMFs and the residual signals as atoms of the learned dictionary, the size of the dictionary can be extremely

larger than the training set, sometimes of triple size or more. The optimization of the dictionary is in desperate need to limit its size thus the recovery time is acceptable. For its universal application and easy usability, we adopt the K-means clustering method to reduce the size of the learned dictionary.

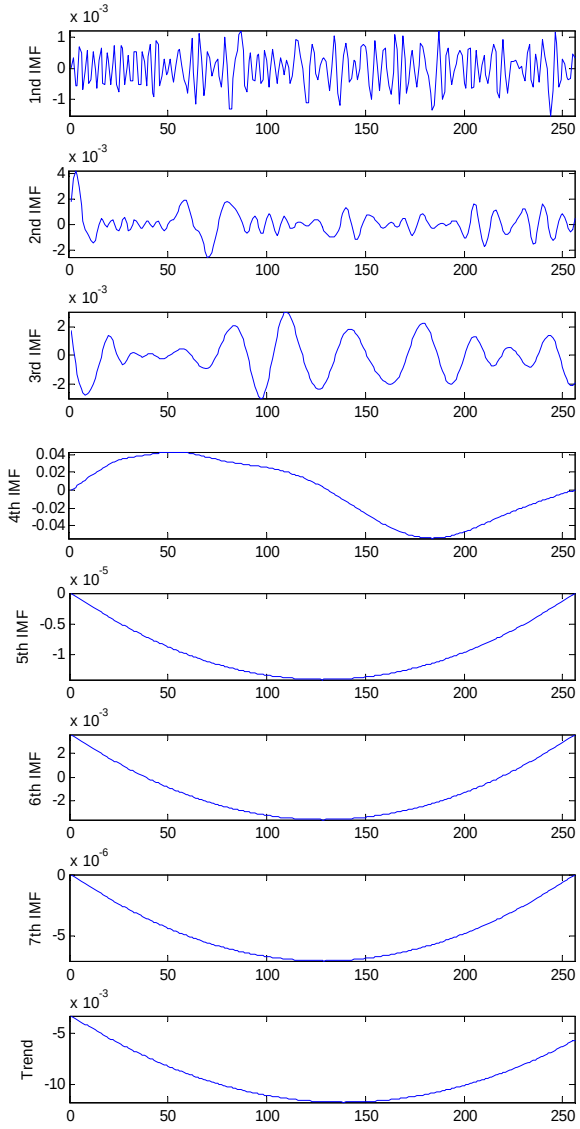

Fig. 2 Decomposition of the original audio signal in Fig. 1

K-means algorithm (also known as the generalized Lloyd algorithm - GLA [8]), is one of the most well-known methods for data clustering. The goal of K-means is to find k points of a dataset that can best represent the dataset in a certain mathematical sense. These k points are also known as cluster centers, prototypes, centroids, or codewords, and so on. The K-means method is aimed to find the best possible dictionary $\Phi \in R^{M\times N}$ to represent the data samples $Y \in R^N$ by nearest neighbor and achieve the representation $X \in R^M$, by solving

$$\min_{C,X}\left\{\left\|Y-\Phi^T X\right\|_F^2\right\} \text{ s.t. } \forall i, x_i = e_k, i=1,...,M \quad \textit{for some } k \quad (4)$$

The algorithm is described below,

Algorithm 2: K-means Clustering Method

1: Initialization: Dictionary $\Phi_0 \in R^{M\times N}$, $j=1$, $\phi_k^0 = 0$

2: While $j < J$ do

3: Sparse Coding Stage:

Partition the indices of Y into k subsets $\left(R_1^{j-1}, R_2^{j-1}, ..., R_k^{j-1}\right)$, each holding the sample indices most similar to the column ϕ_k^{j-1},

$$R_k^{j-1} = \left\{ i \left| \forall l \neq k, \left\| y_i - \phi_k^{j-1} \right\|_2 < \left\| y_i - \phi_l^{j-1} \right\|_2 \right. \right\}$$

4: Dictionary Update Stage:

For each k, update $\phi_k^j = \dfrac{1}{\left|R_k\right|} \sum_{i\in R_k^{j-1}} y_i$

5: $j = j+1$

6: End while

C. EMD Based Dictionary Learning Method

In this section, the proposed EMD based dictionary learning method, which is called K-EMD method, will be introduced in details.

As described above, the K-EMD method is to decompose training set into IMFs and residual components to form a dictionary and to reduce the size of the dictionary using the K-means method. However, directly reducing the dictionary is unadvisable because IMFs of different scales contain different information entropies. Take the first scale of IMF and the residual component for example, as shown in Fig. 1 and Fig. 2, the first-scale-IMF represents an oscillatory mode of the audio signal while the residual component represents the increasing or decreasing trend. It is wise to use the K-means method separately for each scale.

The algorithm of the K-EMD method is illustrated as follows:

Algorithm 3: K-EMD Method

1: Initialization:
Obtain frame length N
Obtain the number of components, i.e. $l = floor(\log_2(N))$
Obtain $K = \{k_1, ..., k_l\}$, a frame of raw audio data $x_0 \in R^N$, Initialize frame No. $i = 1$, empty optimized components KD
2: While (not EOF of training set) do
3: Obtain N-points training data and denote it by x_0^i
4: Apply EMD to x_0^i and get l components $c_1^i, ..., c_l^i$
5: $i = i + 1$
6: End while
7: Obtain the number of frames, i.e. $i = i - 1$
8: Form component sets, i.e. $D_p = \{d_p^q \mid p = 1, ..., l \quad q = 1, ..., i\}$
9: While $j \le l$
10: Apply K-means Method to D_j with K= k_j and get compressed sets KD_j
11: End while
12: Form the dictionary, i.e. $\Phi = [KD_1, ..., KD_l]$

III. Experiment

As we can see from Fig. 2, the higher scales of the IMFs are similar with each other (the higher scales means the 4th to 7th IMFs of the decomposition), thus the higher-scale components can be merged into a single component, as shown in Fig. 3. From Fig. 3, we can also find that the lower scale of decomposed component has higher frequencies than higher scale ones. To our knowledge, lower frequencies contain important information and higher frequencies merely describe the details of the signal. (That's partially the reason why some of the lossy compression methods even abandon the higher frequencies.) For this reason, we choose larger k to clustering the higher scale of the IMFs and smaller k to lower-scale ones.

To find the ration of the k, we use SNR to evaluate the significance of the components. SNR is calculated by,

$$SNR = 20\log_{10}\left(\frac{\|s\|_2}{\|n\|_2}\right) \quad (5)$$

where s indicates the signal and n denotes the difference between the component and the original signal.

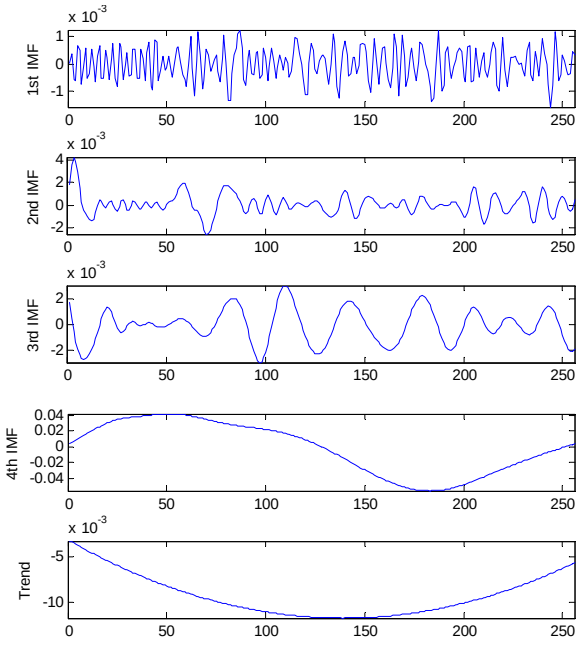

Fig. 3 Decomposition of the signal in Fig. 1 when higher-scale IMFs are merged into one IMF

The SNR curves of the components are shown in Fig. 4, where every value of the SNR curve is the SNR of a frame of one component. We can see from it that the higher-scale IMFs have higher SNRs and hence are more important than lower ones. So we choose the following ratio of k:

$$k_1 : k_2 : k_3 : k_4 : k_5 = 1:2:3:3:4 \quad (6)$$

where k_1 to k_5 is used in K-means method for the 4 IMFs (from scale 1 to scale 4) and the trend, respectively.

Once the values of $k_1 \sim k_5$ are found, the dictionary can be trained by K-means method from the decomposition of the training signal. To get the sparse representation, the Iterative Hard Threshold (IHT) algorithm [9] is applied. In this paper, the training set consists of an audio file with the length of time is 1 minutes and 6 seconds including string music, wind music, and percussion music. The sampling rate is 48000Hz. The training file contains 3146615 sample points and we add 137 points of zero so that we can get 12292 frames each contains 256 sample points. After K-means clustering, 832 atoms are retained to form the dictionary, which means the dictionary is in the size of 832×256 and each atom has the size of 1×256. In this paper, we retain all the coefficients of the sparse representation result and resort them into descending order, as given in Fig. 5. The original signal and the recovered signal are shown in Fig. 6.

It is apparent that the sparse representation coefficients decrease exponentially, which means that the audio signal can be represented by a few coefficients.

To indicate that the proposed algorithm is efficient, the coefficients of the sparse representation of K-SVD and K-EMD are given in Fig. 7, where the coefficients are sorted in

descending order.

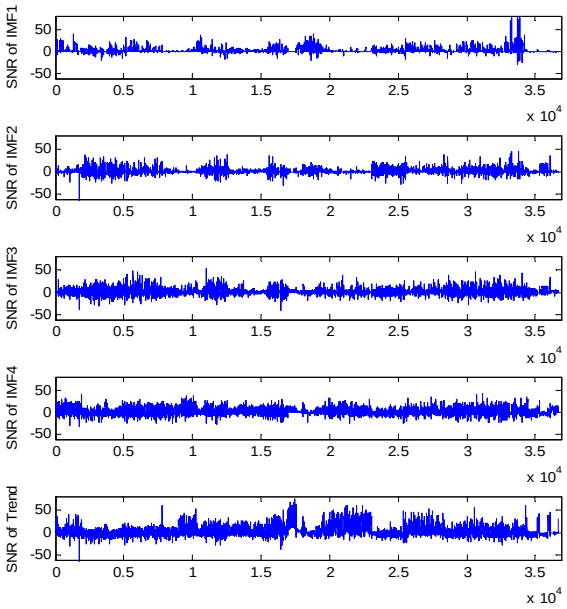

Fig. 4 SNR of the components

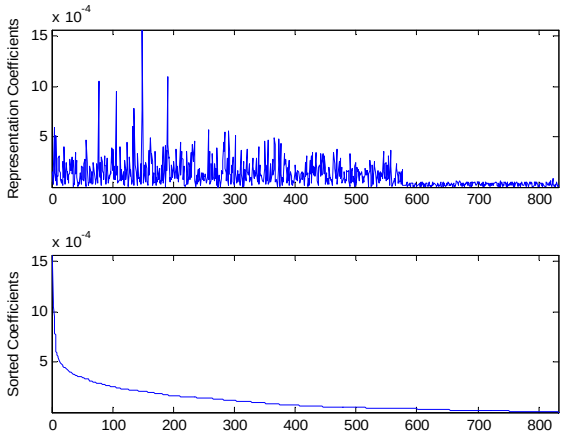

Fig. 5 Sparse representation coefficients and the coefficients sorted in descending order

The sparse coefficients of K-EMD decrease faster than that of K-SVD, and the coefficients are sparser. One of the sparsity measurements is the l_ε^0 method, which is defined as follows:

$$\|\alpha\|_{0,\varepsilon} = \#\{i, |\alpha_i| \geq \varepsilon\} \quad (7)$$

TABLE I
SPARSITY OF THE SPARSE REPRESENTATION USING K-SVD AND K-EMD

Method	l_ε^0 Sparsity			
	20%	30%	40%	50%
K-SVD	66	27	15	8
K-EMD	64	17	6	5

In this paper, ε is defined to be 20%, 30%, 40% and 50% of the largest coefficient, thus the corresponding sparsities of K-EMD and K-SVD are listed in Table I. K-EMD has an average decrease of34.4% in sparsity of the coefficients, which means K-EMD has better sparsity than K-SVD.

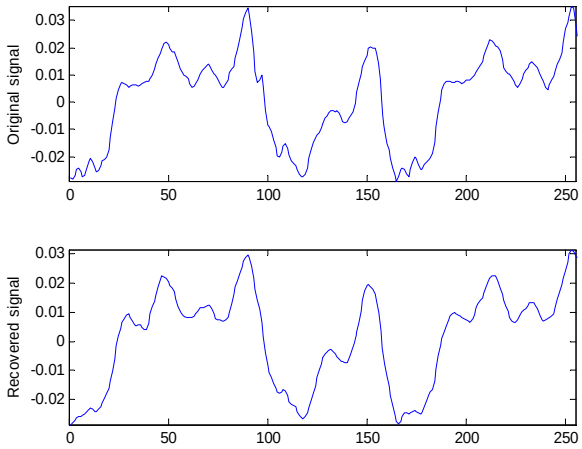

Fig. 6 Original audio and the recovered audio

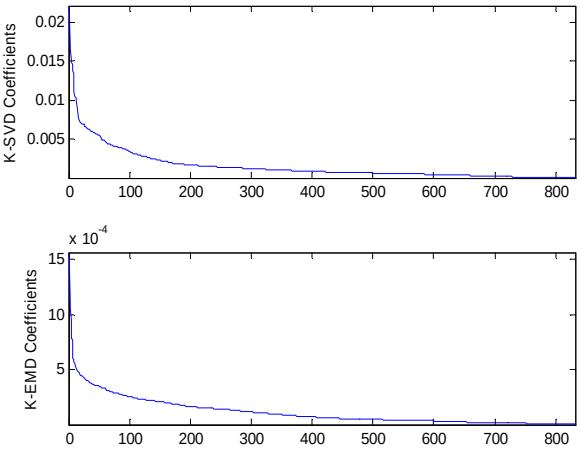

Fig. 7 Sorted sparse representation coefficients of K-SVD and K-EMD

The recovered audio signals of the two dictionary learning methods are shown in Fig. 8, which are recovered using 64 of the coefficients.

It is apparent that the K-EMD has a better recover accurate rate than the K-SVD because the recovered signal using K-EMD is more similar to the original signal. To find whether the K-EMD has better recovery, we use (5) to measure the SNR of the recovered signal, where n indicates the difference between the original signal and the recovered signal. Thus it is clearer to see the superiority. Table II lists the information of the testing set, including the type of the audio, length of time, total points and number of frames. The sampling rate is 48 kHz and frame length is 256 points. The average SNRs of the recovered signal using K-EMD, K-SVD are given in Table III. The SNR of K-EMD is increased about 20.9% than K-EMD.

TABLE II
INFORMATION OF THE 4 TEST FILES

Type	Time (s)	Points	No. of Frames
String Music	3.96	190000	743
String Music	7.02	336962	1317
Wind Music	11.63	558244	2181
Percussion Music	4.76	228493	893

TABLE III
SNR OF THE RECOVERED SIGNAL USING K-SVD, K-EMD

Method	SNR
K-SVD	34.09dB
K-EMD	41.22dB

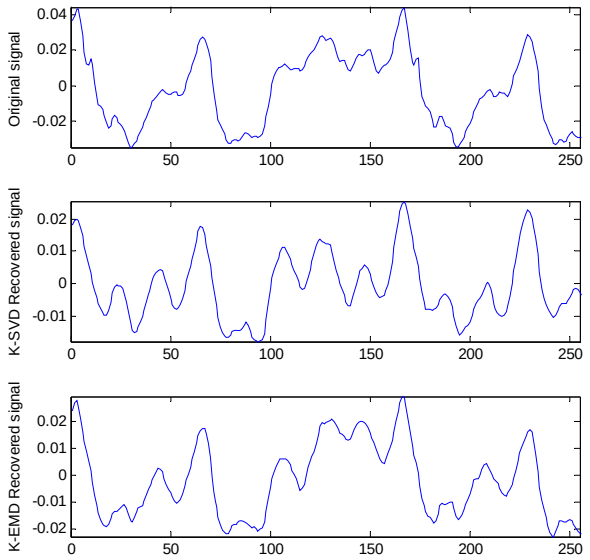

Fig. 8 The original audio and recovered signal using K-EMD and K-SVD

IV. CONCLUSION

In this paper we present a dictionary learning method based on EMD which is called K-EMD method. The dictionary is optimized by K-means clustering method to reduce the size thus can lead to less space and time complexities of the sparse representation algorithm. Experimental results are given to indicate the efficiency of the proposed method. The sparse representation coefficients are 34.4% sparser than that of K-SVD and the SNR of recovered signal of K-EMD is increased by 20.9% compared to K-SVD.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61171171, and the 973 National Basic Research Program of the Ministry of Science and Technology of China under Grant No. 2010CB731904.

REFERENCES

[1] B. Olshausen and D. Field, "Emergence of simple-cell receptive fieldproperties by learning a sparse code for natural images," Nature, vol.381, pp. 607–609, 1996.
[2] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," Neural Comput., vol. 12, pp. 337–365, 2000.
[3] P. Schmid-Saugeon and A. Zakhor, "Dictionary design for matchingpursuit and application to motion-compensated video coding," IEEETrans. Circuits Syst. Video Technol., vol. 14, no. 6, pp. 880–886, 2004.
[4] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEETrans. Signal Process, vol. 54, no. 11, pp. 4311–4322, 2006.
[5] M. Yaghoobi, L. Daudet, and M. Davies, "Parametric dictionary design for sparse coding," in Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS09), 2009
[6] M. G. Jafari and M. D. Plumbley. Fast dictionary learning for sparserepresentations of speech signals. IEEE Journal of Selected Topics in SignalProcessing, 5:1025–1031, Sep. 2011.
[7] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu. H.H. Shih, et al, "The Empirical Mode Decompositionand The Hilbert Spectrum for Nonlinear and Nonstationary Time Series Analysis", Proc. R. Soc. A,1998, 454, pp. 903-995.
[8] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression.Norwell, MA: Kluwer Academic, 1991.
[9] T. Blumensath and M. E. Davies "Iterative thresholding for sparse approximations", J. Fourier Anal. Applicat., vol. 14, no. 5, pp.629 -654 2008