SINGLE-CHANNEL SPEECH SEPARATION USING A SPARSE PERIODIC DECOMPOSITION

Makoto NAKASHIZUKA, Hiroyuki OKUMURA and Youji IIGUNI

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
email: nkszk@sys.es.osaka-u.ac.jp

ABSTRACT

In this paper, we propose a single-channel speech separation method by using a sparse decomposition with a periodic signal model. In our separation method, a mixture of speeches is approximated with periodic signals with time-varying amplitude. The decomposition with the periodic signal model is performed under a sparsity penalty. Due to the sparsity penalty, a segment of the speech mixture is decomposed into periodic signals, each of them is a component of the individual speaker. For speech separation, we introduce the clustering using a K-means algorithm for the set of the periodic signals. After the clustering, each cluster is assigned to its corresponding speaker using codebooks that contain spectral features of the speakers. In experiments, comparison with MaxVQ that performs separation on frequency spectrum domain is demonstrated. The experimental results in terms of signal-to-distortion ratio (SDR) show that our method outperforms MaxVQ with less computational cost for assignment of speech components.

1. INTRODUCTION

In under-determined source separation problems, the sparse representations of sources have been utilized[1]-[3], [8]. Short-time DFT (Discrete Fourier Transform) is one of the sparse representations for speeches and has been applied to multi- and single-channel speech separation problems. Basically, the separation based on the short-time DFT is achieved by masking that eliminates time-frequency components of interferences. The time-frequency masks are composed by two sensors[1] or statistical models[2, 3] of the sources for single-channel speech separation. In other approaches for single channel speech separation, the DFT spectra of the source signals are learnt for the separation. For learning, independent subspace analysis[6], non-negative matrix factorization[9] and non-negative sparse coding[7] have been employed.

For learning the waveforms of the speech to improve the sparsity of the signal representation, the learnt dictionaries that are generated by sparse coding techniques[5]-[8] have been utilized. The sparse coding is a generative model that represents signals in linear combinations of atoms in a dictionary. The dictionary and the coefficients of the linear combination are alternatively updated under a sparsity penalty. The sparse coding yields the dictionaries that mainly consists of oscillatory functions for speech and audio signals[7, 8]. Thus, if the observed signal can be supposed to be a mixture of audio signals, it is expected that the observed signal can be decomposed into the small number of the oscillatory signals.

In order to decompose the signals with a model for oscillatory signals, the sparse periodic decomposition with the sparsity penalty has been proposed[10][11]. In this decomposition method, a signal model that represents periodic signals with time-varying amplitude is employed. The waveforms and the envelopes of hidden periodic signals in the mixture are iteratively estimated under the sparsity penalty. This decomposition can be interpreted as a sparse coding with non-negativity of the amplitude and the periodic structure of signals. By imposing the constraint that are described by the signal model, the sparse periodic decomposition represents a signal with simple computation without preliminarily learnt dictionary.

In this paper, the sparse periodic decomposition is applied to single-channel speech separation. In our approach, the speech mixture is divided into analysis frames. Each frame is decomposed into the periodic signals, each of which is supposed to be a component of the individual source. To estimate the source speeches, the periodic signals are grouped into clusters as many as the expected maximum number of the sources. After the clustering, the clusters are assigned to the sources with spectral features of the sources.

In next section, we explain the definition of the model for the periodic signals that is employed for our speech separation method. Then, the cost function including the sparsity and the algorithm for the decomposition are explained. Next, we introduce the grouping and assignment for the result of the periodic decomposition and realize single-channel speech separation. In this experiment, the separation performance of the proposed decomposition is compared with MaxVQ[2] that performs separation in frequency spectrum domain.

2. SPARSE PERIODIC DECOMPOSITION

2.1 Model for periodic signals with time-varying amplitude

Let us suppose that a sequence $\{f_p(n)\}_{0\le n<N}$ is a finite length periodic signal with length N and an integer period p. It satisfies the periodicity condition with integer period $p \ge 2$ and is represented as

$$f_p(n) = a_p(n) \sum_{k=0}^{K} t_p(n-kp) \tag{1}$$

where $K = \lfloor (N-1)/p \rfloor$ that is the largest integer less than or equal to $(N-1)/p$. The sequence $\{t_p(n)\}_{0\le n<p}$ corresponds to a waveform of the signal in a period and is defined over the interval $[0, p-1]$. $t_p(n) = 0$ for $n \ge p$ and $n < 0$. This sequence is referred to as the p-periodic template. The sequence $\{a(n)\}_{0\le n<N}$ represents the amplitude variation of the periodic signal.

In this section, we discuss the decomposition of mixtures of the periodic signals that can be represented in the form of (1). We assume that the amplitude of the periodic signal varies slowly and can be approximated to be constant within a period. By this simplification, we define an approximate model for the periodic signals with time-varying amplitude as

$$f_p(n) = \sum_{k=0}^{K} a_{p,k} t_p(n-kp). \tag{2}$$

In order to represent periodic components without DC component, the average of $f_p(n)$ over the interval $[0, p-1]$ is zero and the amplitude coefficients $a_{p,k}$ is restricted to non-negative values.

For convenience, a p-periodic signal is represented as an N-dimensional vector $\mathbf{f}_p = \mathbf{A}_p \mathbf{t}_p$. In this form, the amplitude coefficients and the template are represented in an N by p matrix $\mathbf{A}_p$ and a p-dimensional template vector $\mathbf{t}_p$ that is associated with the sequence $t_p(n)$, respectively. $\mathbf{A}_p$ is a union of the matrices as

$$\mathbf{A}_p = \left(\mathbf{D}_{p,1}, \mathbf{D}_{p,2}, \cdots, \mathbf{D}_{p,K+1}\right)^{\mathrm{T}} \tag{3}$$

where superscript T denotes transposition. $\{\mathbf{D}_{p,j}\}_{1\le j\le K}$ are p by p diagonal matrices whose elements correspond to $a_{p,j-1}$. $\mathbf{D}_{p,K+1}$ is a p by $N-pK$ matrix whose non-zero coefficients that correspond to $a_{p,K}$ appear only in (i,i) elements. Since only one element is non-zero in any row of the $\mathbf{A}_p$, $\mathbf{A}_p$ is defined as a matrix that consists of orthogonal columns, l_2 norms of which are normalized. To hold the condition that the average of the periodic signal in (2) over a period is zero, we impose the condition

$$\mathbf{u}_p^{\mathrm{T}}\mathbf{t}_p = 0 \tag{4}$$

on the p-template vector. $\mathbf{u}_p$ is a p-dimensional vector which elements equal to the diagonal elements of $\mathbf{D}_{p,1}$.

Alternatively, p-periodic signals in (2) can be represented as $\mathbf{f}_p = \mathbf{T}_p\mathbf{a}_p$. In this form, the amplitude coefficients and the template are represented in an N by $K+1$ matrix $\mathbf{T}_p$ and $K+1$-dimensional amplitude coefficients vector $\mathbf{a}_p$ whose elements are associated with the amplitude coefficients $\{a_{p,k}\}$, respectively. $\mathbf{T}_p$ consists of the column vectors that correspond to the shifted versions of the p-periodic template. As same as $\mathbf{A}_p$, only one element is non-zero in any row of $\mathbf{T}_p$. We define $\mathbf{T}_p$ as a matrix that consists of columns, l_2 norms of which are normalized.

In this study, we employ an approximate decomposition method that obtains a representation of a given signal $\mathbf{f}$ as a form:

$$\mathbf{f} = \sum_{p\in\mathbb{P}} \mathbf{f}_p + \mathbf{e} \tag{5}$$

for speech separation. $\mathbb{P}$ is a set of periods that are preliminary specified for the decomposition. $\mathbf{e}$ is an approximation error between the model and the signal $\mathbf{f}$. We assume that the observed signal $\mathbf{f}$ is a mixture of signals that consists of few periodic components. Under this assumption, a sparsity penalty is introduced to the periodic decomposition in the form of (5).

2.2 Sparse decomposition with periodic signal model

The sparse representation is a generative model for signals and represents a signal $\mathbf{f}$ into atoms that are column vectors of a dictionary $\mathbf{\Phi}$ as $\mathbf{f} = \mathbf{\Phi}\mathbf{c} + \mathbf{e}$. l_2 norm of each atom in $\mathbf{\Phi}$ is normalized to unity. $\mathbf{c}$ is a coefficient vector. $\mathbf{e}$ is a supposed noise component that appears in the signal. The number of the column vectors of $\mathbf{\Phi}$ is larger than the dimensionality of the signal $\mathbf{f}$. The problem of the sparse representation is to decompose the signal $\mathbf{f}$ while minimizing the number of non-zero coefficients in $\mathbf{c}$. In many studies, the sparsity of the coefficient vector $\mathbf{c}$ is measured by l_1 norm. The sparse coefficients that approximate the signal $\mathbf{f}$ under the noise assumption are obtained by

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \frac{1}{2}||\mathbf{f} - \mathbf{\Phi}\mathbf{c}||_2^2 + \lambda||\mathbf{c}||_1 \tag{6}$$

where $||\cdot||_1$ denotes the l_1 norm of a vector and λ denotes a Lagrange multiplier. This unconstrained minimization problem is referred to as basis pursuit denoising (BPDN) [4]. The dictionary $\mathbf{\Phi}$ is fixed for signal approximate decomposition in the BPDN. In the sparse coding strategies[5, 7, 8], the dictionary $\mathbf{\Phi}$ is adapted to the set of the signals. The dictionary is updated with the most probable one under the estimated coefficients and the set of the signals. From Bayesian point of view, the minimization (6) is the equivalent of MAP estimation of the coefficient vector $\mathbf{c}$ with a Laplacian prior [5].

For our periodic decomposition, we also impose the sparsity penalty on the decomposition under the assumption that the mixture contains a small number of periodic signals that can be approximated in the form of (2). Our objective is to achieve signal decomposition to obtain a small number of periodic subsignals rather than the basis vectors. In order to achieve this, the probability distribution of the l_2 norm of each periodic signals is assumed to be the Laplacian, then the distribution of the set of the periodic signals is $P(\{\mathbf{f}_p\}_{p\in\mathbb{P}}) \propto \prod_{p\in\mathbb{P}} \exp\left(-\alpha_p||\mathbf{f}_p||_2\right)$. The noise is assumed to be Gaussian, then the conditional probability of $\mathbf{f}$ is $P\left(\mathbf{f}|\{\mathbf{f}_p\}_{p\in\mathbb{P}}\right) \propto \exp\left(-\frac{1}{2\lambda}||\mathbf{f} - \sum_{p\in\mathbb{P}}\mathbf{f}_p||_2^2\right)$. Along with Bayes' rule, the conditional probability distribution of the set of the periodic signals is

$$P\left(\{\mathbf{f}_p\}_{p\in\mathbb{P}}|\mathbf{f}\right) \propto P\left(\mathbf{f}|\{\mathbf{f}_p\}_{p\in\mathbb{P}}\right) P\left(\{\mathbf{f}_p\}_{p\in\mathbb{P}}\right). \tag{7}$$

Substituting the prior distributions of the periodic signals and the noise into (8), we can derive the likelihood function of the set of periodic signals. From the likelihood function, we define the cost function E for the periodic decomposition as:

$$E\left(\{\mathbf{f}_P\}_{p\in P}\right) = \frac{1}{2} \parallel \mathbf{f} - \sum_{p\in\mathbb{P}} \mathbf{f}_p \parallel_2^2 + \lambda \sum_{p\in\mathbb{P}} \alpha_p \parallel \mathbf{f}_p \parallel_2 . \tag{8}$$

In our periodic decomposition, a signal $\mathbf{f}$ is decomposed into a set of periodic signals while reducing the cost E. To find the set of the periodic subsignals $\{\mathbf{f}_p\}_{p\in P}$, we employ a relaxation algorithm. This relaxation algorithm always updates one chosen periodic subsignal while assuming all the other periodic subsignals to be fixed. The template and amplitude coefficients of the chosen periodic signal are alternatively updated in an iteration. In the algorithm, we suppose that the set of the periods $\mathbb{P}$ consists of M periods which are indexed as $\{p_1, \cdots p_M\}$. The relaxation algorithm for the sparse periodic decomposition is as follows:

1) Set the initial amplitude coefficients for $\{\mathbf{A}_p\}_{p\in\mathbb{P}}$
2) $i = 1$
3) Compute the residual $\mathbf{r} = \mathbf{f} - \sum_{j\neq i} \mathbf{f}_{p_j}$
4) Represent $\mathbf{f}_{p_i}$ as $\mathbf{A}_{p_i}\mathbf{t}_{p_i}$. If $\parallel \mathbf{f}_p \parallel_2 = 0$, then the amplitude coefficients are specified to be constant. Update the template $\mathbf{t}_{p_i}$ with the solution of a subproblem:

$$\min_{\mathbf{t}_{p_i}} \frac{1}{2} \parallel \mathbf{r} - \mathbf{A}_{p_i}\mathbf{t}_{p_i} \parallel_2^2 + \lambda\alpha_{p_i} \parallel \mathbf{t}_{p_i} \parallel_2 \text{ s. t. } \mathbf{u}_{p_i}^{\mathrm{T}}\mathbf{t}_{p_i} = 0 \tag{9}$$

5) Represent $\mathbf{f}_{p_i}$ as $\mathbf{T}_{p_i}\mathbf{a}_{p_i}$. Update the template $\mathbf{a}_{p_i}$ with the solution of a subproblem:

$$\min_{\mathbf{a}_{p_i}} \frac{1}{2} \parallel \mathbf{r} - \mathbf{T}_{p_i}\mathbf{a}_{p_i} \parallel_2^2 + \lambda\alpha_{p_i} \parallel \mathbf{a}_{p_i} \parallel_2 \text{ s. t. } \mathbf{a}_{p_i} \geq \mathbf{0} \tag{10}$$

6) If $i < M$, update $i \leftarrow i+1$ and go to step 3). If $i = M$ and the stopping criterion is not satisfied, go to step 2).

For stable computation, the update stage of the amplitude coefficient in Step 5) is omitted when the l_2 norm of the template $\mathbf{t}_{p_i}$ becomes zero after Step 4). The closed form solution of (9) is

$$\hat{\mathbf{t}}_{p_i} = \begin{cases} \frac{\|\mathbf{v}\|_2 - \lambda\alpha_{p_i}}{\|\mathbf{v}\|_2}\mathbf{v} & \text{for } \parallel \mathbf{v} \parallel_2 > \lambda\alpha_{p_i} \\ \mathbf{0} & \text{for } \parallel \mathbf{v} \parallel_2 \leq \lambda\alpha_{p_i} \end{cases} \tag{11}$$

where

$$\mathbf{v} = \mathbf{A}_{p_i}^{\mathrm{T}}\mathbf{r}_{p_i} - \frac{\mathbf{u}_{p_i}^{\mathrm{T}}\mathbf{A}_{p_i}^{\mathrm{T}}\mathbf{r}_{p_i}}{\parallel \mathbf{u}_{p_i} \parallel_2^2}\mathbf{u}_{p_i}. \tag{12}$$

The solution of (10) is

$$\hat{\mathbf{a}}_{p_i} = \begin{cases} \frac{\|\mathbf{w}\|_2 - \lambda\alpha_{p_i}}{\|\mathbf{w}\|_2}\mathbf{w} & \text{for } \parallel \mathbf{w} \parallel_2 > \lambda\alpha_{p_i} \\ \mathbf{0} & \text{for } \parallel \mathbf{w} \parallel_2 \leq \lambda\alpha_{p_i} \end{cases} \tag{13}$$

where

$$\mathbf{w} = \left(\mathbf{T}_{p_i}^{\mathrm{T}}\mathbf{r}_{p_i}\right)_+ . \tag{14}$$

$(\cdot)_+$ denotes replacing the negative elements with zero. Both solutions of the subproblem guarantee decrement of the cost E. Thus, the cost E decreases until convergence. However, the set of the resultant periodic signals does not always obtain a minimum of the cost function exactly. If any periodic signal becomes zero in an

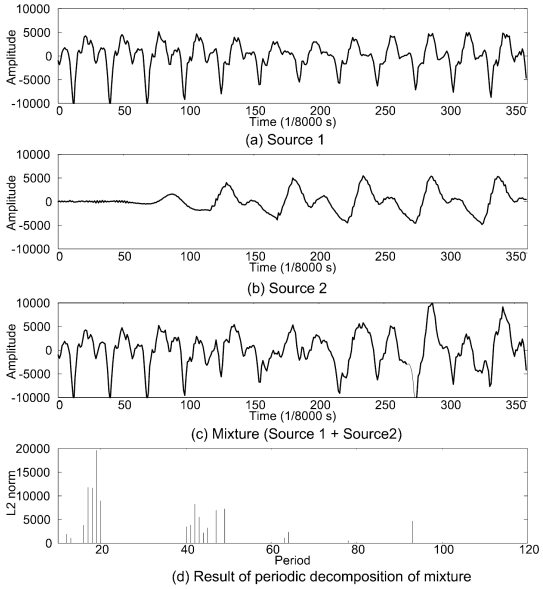

Fig. 1. Example of sparse periodic decomposition. (a), (b) Source speech segment, (c) mixture and (d) distribution of l_2 norm of the decomposed periodic signals.

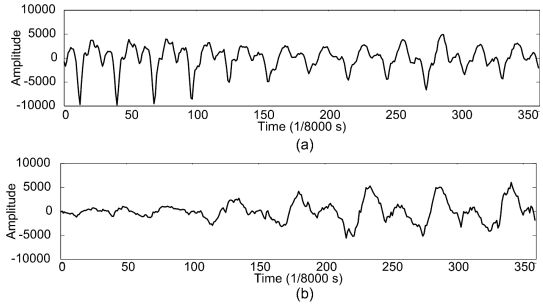

Fig. 2. Estimated sources from mixture in Fig. 1(c) with K-means clustering of the decomposed periodic signals.

iteration, the amplitude coefficients are specified to be constant in step 4) of the next iteration. The proper search direction for $\mathbf{f}_p$ may not be obtained by these amplitude coefficients. However, the l_2 norms of the periodic signals that are eliminated by the shrinkage in (11) and (12) with λ_p are small enough to approximate the signal. Hence, we accept the periodic signals obtained by this algorithm as the sparse decomposition results instead of the proper minimizer of the cost E. An example of the sparse decomposition with the periodic signal models is shown in Fig. 1. In this example, the observed signal $\mathbf{f}$ is the mixture of two speech segments in a rate of 8 kHz. The source frames and mixture are shown in Fig. 1 (a), (b) and (c), respectively. The set of periods for the decomposition is a range $[10,120]$ that corresponds to the region of the pitches of most male and female speeches. The products $\lambda\alpha_{p_i}$ should be specified proportional to the expected l_2 norm of the noise that is approximated with the periodic signal model. In this experiment, the set of the parameters $\lambda\alpha_{p_i}$ is specified as $\sigma\sqrt{p+1+p/N}$ that is the expected l_2 norm of the approximated Gaussian noise with the variance σ^2. σ is specified to 1% of the l_2 norm of the signal. The distribution of the l_2 norms of the resultant periodic subsignals of the mixture in Fig. 1 (c) is shown in Fig. 1 (d). For this mixture, 18 periodic signals appear in the decomposition result. Most of the periods distribute around the fundamental pitch periods of the sources. To recover the sources from the set of the decomposed periodic signals, these periodic signals have to be assigned to the sources. In next section, the clustering and assignment are applied to the periodic signals to achieve single-channel speech separation.

3. SINGLE CHANNEL SPEECH SEPARATION

In this section, we apply the sparse periodic decomposition to the single-channel speech separation. For the speech separation, the observed speech mixture is divide into analysis frames. We suppose that a frame $\mathbf{z}$ of the observed mixture consists of source speech signals $\{\mathbf{x}_i\}_{i=1}^{N_s}$ as

$$\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2 + \cdots \mathbf{x}_{N_s}, \tag{15}$$

where N_s is the maximum number of speakers. Each source $\mathbf{x}_i$ is supposed to be decomposed into a set of periodic signals $\{\mathbf{f}_p^i\}_{p\in\mathbb{P}_i}$ with our periodic signal model shown in (2). $\mathbb{P}_i$ denotes the set of periods that have non-zero periodic signals of the i-th speaker. For the separation of the sources, we assume that the source signals do not share same period as

$$\mathbb{P}_i \bigcap \mathbb{P}_j = \emptyset \tag{16}$$

where $i \neq j$. Under this assumption, the observed mixture $\mathbf{z}$ can be represented with the periodic signals

$$\mathbf{z} = \sum_{p\in\mathbb{P}_1} \mathbf{f}_p^1 + \sum_{p\in\mathbb{P}_2} \mathbf{f}_p^2 \cdots \sum_{p\in\mathbb{P}_{N_s}} \mathbf{f}_p^{N_s}. \tag{17}$$

By using this periodic representation of the mixture, the speech separation problem can be achieved by following three stages at each frame. The first stage is decomposition of the source frame $\mathbf{z}$ into the periodic components shown in (17). We employ the sparse periodic decomposition for this stage. The second stage is grouping of the periodic signals that represent same speaker into same cluster. A K-means clustering is used to group the periodic signals into the clusters as many as the maximum number of speakers N_s. By summing the periodic signals in each cluster, the frame of the mixture is decomposed into N_s signal components. In the third stage, each component is assigned to the proper speaker. For this assignment, we use the preliminary learnt features of the supposed speakers. In following two subsections, we explain the clustering of the decomposed periodic signals and the assignment of the clustered speech components.

3.1 Clustering of periodic signals

For the clustering, the periodic representation of the mixture frame $\mathbf{z}$ in (16) is supposed to be obtained by the sparse periodic decomposition. Due to the sparsity penalty, the periodic subsignal of i-th speaker are highly correlated with i-th source signal $\mathbf{x}_i$. Therefore, we assume that each periodic signal satisfies the relationship

$$\frac{(\mathbf{f}_p^i)^{\mathrm{T}}\mathbf{x}_i}{\|\mathbf{x}_i\|_2} > \frac{(\mathbf{f}_p^i)^{\mathrm{T}}\mathbf{x}_j}{\|\mathbf{x}_j\|_2} \tag{18}$$

where p is any period in $\mathbb{P}_i$ for $i \neq j$. Under this assumption, if the actual source signals are obtained, each periodic subsignal can be classified into the cluster that corresponds to the individual speaker. However, the actual source signals are unknown for the separation. We hence use a K-means clustering algorithm for the grouping. In this stage, the decomposed periodic signals are grouped into the clusters as many as the supposed maximum number of the sources.

Commonly, the K-means algorithm minimizes the sum of the Euclidean distances between a center and elements in each cluster. By the assumption shown in (18), we employ the normalized correlation as the metric of the K-means algorithm. In this clustering, the feature vector that is used for the K-means algorithm is defined as the amplitude of the frequency spectrum of each decomposed periodic signal. Our signal model for periodic signals in (2) can

represent the amplitude variations of the periodic signals, but cannot represent frequency variations. The frequency variations that appear within a frame is represented by a sum of the periodic signals, of which periods are neighboring. The frequency spectra of such periodic signals are closely related, but the amplitude variations are not. Therefore, we employ the amplitude spectrum of the periodic signal as the feature of the clustering to ignore amplitude variations and improve the accuracy of the clustering.

After the clustering, the decomposed component is obtained by summing the periodic signals in same cluster. By this clustering, the frame of the mixture is decomposed into the components as many as the maximum number of sources N_s. We suppose that these decomposed components are denoted as $\{\mathbf{d}_i\}_{i=1}^{N_s}$.

The example of the decomposition is shown in Fig. 2. These signals are given by the K-means clustering from the periodic decomposition in Fig. 1(d) of the mixture Fig. 1(c). Each source frame is approximated by the corresponding cluster of the decomposed periodic signals.

3.2 Assignment of clusters to speakers

After the clustering, the frame of the mixture is decomposed into N_s components. The source separation problem is reduced to the problem that is to find the assignment in the possible $N_s^{N_s}$ combinations of the decomposed signals and the sources. To find the assignment, features of the speakers that are preliminary learnt are utilized. We suppose that the clean speech signals of the speakers are available and can be utilized to the assignment. In this work, the codebook that contains the representative vectors that represent the normalized amplitude frequency spectra of the speakers are utilized. The codebook is learnt for each clean speech of the speaker by the shape vector quantization using LBG (Linde-Buzo-Gray) algorithm. For the assignment, the similarities that are defined as the normalized correlations between the representative vectors and candidates of the separated outputs that obtained for all possible combinations are computed. When $K = 2$, the decomposed components for the frame are $\mathbf{d}_1$ and $\mathbf{d}_2$. The possible separated outputs are

$$\hat{\mathbf{x}}_1 = u_1\mathbf{d}_1 + u_2\mathbf{d}_2$$

$$\hat{\mathbf{x}}_2 = (1-u_1)\mathbf{d}_1 + (1-u2)\mathbf{d}_2$$

where $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ are estimates of the 1st and 2nd speaker, respectively. u_1 and u_2 are chosen from $\{0,1\}$ to maximize the similarity between the separated output and any representative vector in the codebooks. Let us suppose that the codebooks of for the amplitude spectra of the 1st and 2nd speaker are obtained as $\{\mathbf{c}_{1,i}\}_{i=0}^{N_c}, \{\mathbf{c}_{2,i}\}_{i=0}^{N_c}$, respectively. Every representative vector is normalized to unity. N_c denotes the number of the representative vectors in each codebook. The similarity is defined as

$$\frac{\max_i(\mathbf{c}_{1,i}^{\mathrm{T}}\tilde{\mathbf{x}}_1) + \max_i(\mathbf{c}_{2,i}^{\mathrm{T}}\tilde{\mathbf{x}}_2)}{\| \tilde{\mathbf{x}}_1 \|_2 \| \tilde{\mathbf{x}}_2 \|_2},$$

where $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ are the amplitude spectra of the candidates of the separated outputs $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$. u_1 and u_2 that maximize this similarity are selected for the assignment. Since the number of the combination of the separated out is $N_s^{N_s}$, the computational cost for this assignment is proportional to $O(N_c N_s^{N_s})$. After the assignment, each estimation of the source is added to the separated output with multiplying the hanning window.

4. SEPARATION RESULTS

For the separation experiments, the speech signals were selected 32 continuous speech signals of about 8 s taken from ATR-SLDB (Spoken Language Database). The speech signals are obtained by 8 Japanese male speakers and 8 female speakers reading sentences. Two speech signals are obtained by each speakers. 480 speech mixtures that consist of two speeches obtained by different speakers are generated. The sampling rate of each speech signal is converted to 8 kHz. For each speaker pair, the mixtures are generated at SNR (Signal to Noise Ration) of the source speeches in the mixtures at 0 dB. Each mixture is divided into frames that contain 360 samples with $3/4$ overlap for the periodic decomposition.

Table 1. Averages and standard deviations of SDR (F and M denote female and male speech, respectively)

	SPD w. oracle	SPD+K-means w. oracle	SPD+K-means w. codebook	Binary Mask w. oracle	MAXVQ w. codebook
F/F+M	11.0/0.9	9.6/0.9	7.2/1.2	14.4/1.2	6.1/1.4
M/F+M	10.6/0.9	9.2/0.9	6.8/1.2	14.0/1.2	5.8/1.5
F/F+F	11.3/1.0	10.0/1.1	2.9/1.9	14.1/1.5	2.5/2.1
M/M+M	9.9/0.9	8.3/0.9	3.1/1.6	12.9/1.0	2.0/1.8

Table 2. Averages and standard deviations of SIR (F and M denote female and male speech, respectively)

	SPD w. oracle	SPD+K-means w. oracle	SPD+K-means w. codebook	Binary Mask w. oracle	MAXVQ w. codebook
F/F+M	16.3/1.4	17.3/1.6	13.2/1.9	35.4/7.0	13.0/2.5
M/F+M	17.4/1.6	18.2/2.1	13.1/2.9	33.8/6.9	17.2/3.3
F/F+F	16.7/1.4	17.2/1.6	8.0/2.5	30.9/3.2	8.9/3.9
M/M+M	15.5/1.4	16.3/1.5	8.3/2.3	29.5/2.8	8.0/3.0

Five separation methods are applied to the mixtures. Three methods are based on the sparse periodic decomposition. The first method is based on the periodic decomposition with oracle. In this method, the decomposed periodic signals are assigned to the speakers by using the actual sources in the mixture. The ideal performance of the source separation by using the periodic decomposition is obtained by this method. Moreover, the assumptions shown in (16) and (18) are verified by this experiment.

The second method performs the separation with the K-means clustering. In the second method, the periodic signals are grouped into two signals by the K-means algorithm at every frame, and each decomposed component that is obtained by summing periodic signals in same cluster is assigned to the speaker by using the actual source in the mixture. By comparing the first with the second, degradations that are caused by the K-means clustering can be evaluated. The third method is the semi-blind separation with the speaker's codebooks that were explained in Sect. 3.

Two DFT-based separation methods are performed for comparison. The first is the ideal binary masking that indicates the ideal performance of the DFT based speech separation. The second is the MaxVQ[2] that also uses codebooks that are generated from log-spectra of clean speeches of the speakers to compose the binary masks. In the MaxVQ, the spectra of the mixture are assumed to be produced with the mixture-maximum model[2]. The MaxVQ seeks all possible pairs of the representative vectors belong to the speakers to find the pair that obtains the closest mixture-maximum to the input mixture[2]. For the DFT based methods, the length of the segments is specified as 512 samples that provides the optimum frequency resolution for speech separation[1]. The number of the representative vectors N_c for the proposed method and the MaxVQ is specified as 256 and are obtained from the clean speeches of 1 min that does not include the sources in the mixtures. For evaluation of the separated results, we employ SDR (Signal to Distortion Ratio), SIR (Signal to Interference Ratio) and SAR (Signal to Artifact Ratio) that are commonly used for the evaluation of signal separation[12].

The average SDRs, SIRs and SARs of the separation results are shown in Tables 1, 2 and 3, respectively. In these table,

Table 3. Averages and standard deviations of SAR (F and M denote female and male speech, respectively)

	SPD w. oracle	SPD+K-means w. oracle	SPD+K-means w. codebook	Binary Mask w. oracle	MAXVQ w. codebook
F/F+M	12.4/0.7	10.6/0.9	8.2/1.1	14.5/1.2	7.4/1.0
M/F+M	12.0/0.7	10.1/0.8	8.2/1.1	14.0/1.2	6.3/1.4
M/M+M	11.4/0.7	9.2/0.9	5.4/1.3	13.0/1.0	4.1/1.3
F/F+F	12.7/1.0	10.9/1.0	5.3/1.3	14.2/1.5	4.6/1.4

"w. oracle" denotes the results of the separation using the actual source signals in the mixture. Audio examples are available at http://sip.sys.es.osaka-u.ac.jp/~nkszk/EUSIPCO2009/.

In Table1, we see that the ideal SDRs obtained by the sparse periodic decomposition is lower than the ideal binary masking of DFT by about 3 dB. Since the frequency resolution of the periodic decomposition is lower than the DFT in the high frequency range, the interferences mainly occur in the high frequency range. These interferences are also reflected in the average SIR in Table 2. Compared with the ideal results of the periodic decomposition, the SDR of the separation results with the K-means clustering is lower than the ideal SDR by about 1.5dB. We see that the K-means clustering decreases the average SARs by about 2 dB in Table 3. However, the decrements of the SIRs due to the K-means clustering do not occur in Table 2. This reason is that the errors of the clustering of the periodic signals degrade both of target speech and interference simultaneously. Hence, the SIRs after the clustering slightly increase. The results of the separation with the K-means clustering and the codebooks of the speakers is about 7 dB for opposite genders and 3 dB for same gender in SDR. As seen in the results, most of the degradations of the separation results occur in the assignment stage using the speakers codebooks.

Compared with the MaxVQ, the average SDRs obtained by the proposed method with the assignment using the codebooks are significantly larger than the MaxVQ for female-male mixtures. For female-male mixtures, the average SIRs obtained by the MaxVQ are higher than the proposed results. However, the average SARs of the MaxVQ are lower than the proposed separation by about 1 dB for all combinations of genders. Actually, heave bubble noises that are caused by the hard mask that is generated from the codebooks are audible in the results obtained by the MaxVQ. In our results, the interferences and undesired amplitude attenuations of the target speech suddenly occur in the separation results due to the error of the assignment of the decomposed components. If the assignment accuracy of the proposed method is improved, these degradations can be suppressed.

There is room for improvements in the assignment of the decomposed components. The DFT based separation is the problem of the assignment of the 257 frequency bins in this experiment. In contrast, the number of the decomposed signal yield by the proposed method with the K-means clustering is the maximum number of expected speakers N_s. Comparing with the DFT, the separation problem can be reduced to small size of a combination problem in our separation method. The computational cost of the MaxVQ for the assignment of the frequency bins is proportional to $O(N_c^{N_s})$ where N_c is the numbers of the representative vectors in the codebook. The proposed method requires $O(N_c N_s^{N_s})$ operations for the assignment of the decomposed signals where $N_s = 2$ in this work. The computational cost for the assignment of the source components is drastically reduced by the sparse periodic decomposition with the K-means clustering.

5. CONCLUSIONS

In this paper, we proposed a speech separation method using the sparse signal representation that decomposes a signal into the set of the periodic signals with time-varying amplitude. In this decomposition, the signal is decomposed while reducing the cost that includes the sparsity penalty. By this sparsity penalty, each decomposed periodic signal of the speech mixture correlates with the corresponding source and can be grouped to form the source by the simple clustering algorithm. We demonstrate the comparison with the DFT-based separation methods and show that the separation results of the proposed method are comparable to the DFT-based separation methods.

In our separation method, every frame is decomposed into the components as many as the maximum number of the expected speakers. In the clustering stage, if the number of the active speakers is estimated, the separation quality can be improved. The use of an information criterion will facilitate the estimation of the number of the cluster that corresponds to the number of active speakers.

In the separation results in Table 1, we see that most degradations occur in the assignment using the codebooks of the speakers. For the DFT based separation methods, the soft masks that are generated using the GMM (Gaussian Mixture Model) of the speeches have been proposed to improve the quality of the separation results[3]. The assignment stage of the proposed method can be improved by using the advanced speaker models. Moreover, the assignment of this work does not utilize temporal continuity of the speech spectrum and fundamental frequency. We can also improve our method by using the temporal continuity of speech pitches for improvement of the accuracy of the assignment. The accurate and robust assignment of the decomposed signals is a topic for future research.

REFERENCES

[1] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking, " *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830-1847, July 2004.

[2] S. T. Roweis, "Factorial models and re-filtering for speech separation and denoising, " *Proc. on Eurospeech*, vol. 7, no. 6, pp. 1009-1012, Geneva, 2003.

[3] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation, " *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1766-1776, Aug. 2007.

[4] S. S. Chen, D. L. Donoho and M. A. Saunders, "Atomic decomposition by basis pursuit, " *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33-61, 1998.

[5] M. S. Lewicki and B. A. Olshausen, "A probabilistic framework for the adaptation and comparison of image codes, " *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 16, no. 7, pp. 1587-1601, 1999.

[6] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis, " *Proceedings of International Computer Music Conference*, Berlin, August 2000.

[7] M. D. Plumbley, S. A. Abdallah, T. Blumensath and M. E. Davies, "Sparese representation of polyphonic music, " Signal Processing, vol. 86, no. 3, pp. 417-431, March 2006.

[8] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music, " *IEEE Trans, on Audio, Speech and Language Processing*, pp. 50-57, vol. 14, no.1 Jan. 2006.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, " *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066-1074, March 2007.

[10] M. Nakashizuka, "A sparse decomposition method for periodic signal mixtures, " IEICE Trans. on Fundamentals, vol. E91-A, no. 3, pp.791-800, March 2008.

[11] M. Nakashizuka, H. Okumura and Y. Iiguni, "A sparse periodic decomposition and its application to speech representation, " in Proc. on EUSIPCO 2008, Lausanne, Aug. 2008.

[12] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation, " *IEEE Trans. on Audio, Speech and Language processing*, vol. 14, no. 4, pp. 1462-1469, July 2006.