16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

A SPARSE PERIODIC DECOMPOSITION AND ITS APPLICATION TO SPEECH REPRESENTATION

Makoto Nakashizuka, Hiroyuki Okumura and Youji Iiguni

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
phone: +81-6-6850-6376, fax: +81-6-6850-6376, email: nkszk@sys.es.osaka-u.ac.jp

ABSTRACT

This study proposes a method to decompose a signal into a set of periodic signals with time-varying amplitude. The proposed method imposes a penalty on the resultant periodic subsignals in order to improve the sparsity of the decomposition. This penalty is defined as the sum of the l_2 norms of the resultant periodic subsignals to find the shortest path to the approximation. By this penalty of the sparsity, the proposed decomposition extracts significant periodic components from a mixture and has ability of the source estimation for mixtures of periodic signals. In experiments, we apply the proposed decomposition to speech mixtures and demonstrate that the proposed decomposition can estimate periodic components of each source speech and detects pitch contours of the source. In additionally, the single-channel speech separation using a lazy assignment of the periodic signals is presented to demonstrate the source separation capability of the proposed decomposition.

1. INTRODUCTION

Periodicities are found in speech signals, musical rhythms, biomedical signals and machine vibrations. In many signal processing applications, signals are assumed to be periodic or quasi-periodic. Especially in acoustic signal processing, signal models based on periodicities have been proposed. The sinusoidal modeling [1] has been proposed to transform an acoustic signal to a sum of sinusoids. The frequency of sinusoids for the representation is detected in the short-time Fourier transform (STFT) spectrum by peak-picking. This approach depends on the frequency spectrum of the signal. The signal modeling in the time-domain [2] has been proposed to extract a waveform of an acoustic signal and its parameters of the amplitude and frequency variations. These approaches aim to represent an acoustic signal that have single fundamental frequency.

For detection and estimation of more than one periodic signal hidden in a signal mixture, several signal decomposition techniques capable of decomposing a signal into a set of periodic subsignals have been proposed [3, 4, 5]. In [4], an orthogonal decomposition method based on periodicity has been proposed. This technique achieves the decomposition of a signal into periodic subsignals that are orthogonal to each other. The periodicity transform [3] decomposes a signal by projecting it onto a set of periodic subspaces. In this method, seeking periodic subspaces and rejecting found periodic subsignals from an input signal are performed iteratively. For reduction of the redundancy of the periodic representation, a penalty of sparsity has been introduced to the decomposition in [5]. In these periodic decomposition method, the amplitude of each periodic signal in the mixture is assumed to be constant. So, it is difficult to obtain the significant decomposition results for the mixtures of quasi-periodic signals with time-varying amplitude.

In this paper, we introduce a model for periodic signals with time-varying amplitude into the periodic decomposition. In order to reduce the number of resultant periodic subsignals obtained by the decomposition and represent the mixture with only significant periodic subsignals, we impose a sparsity penalty on the decomposition. This penalty is defined as the sum of l_2 norm of the resultant periodic subsignals to find the shortest path to the approximation of the mixture. The waveforms and the amplitude of the hidden periodic signals are iteratively estimated with the penalty of the sparsity. The proposed decomposition can be interpreted as a sparse coding [10] with non-negativity of the amplitude and the periodic structure of signals. In our approach, the decomposition results are associated with the fundamental periods of the source signals in the mixture. So, the pitches of the source speeches can be detected from the speech mixtures by the proposed decomposition.

First, we explain the definition of the model for the periodic signals. Then, the cost function including the sparsity measure is proposed for the periodic decomposition. A relaxation algorithm for the sparse periodic decomposition is also explained. The source estimation capability of our decomposition is demonstrated by several examples of the decomposition of synthetic periodic signal mixtures. Next, we apply the proposed decomposition to speech mixtures and demonstrate the speech separation. In this experiment, the ideal separation performance of the proposed decomposition is compared with the separation method obtained by an ideal binary masking [11] of a STFT. Finally, we provide the results of the single-channel speech separation with simple assignment technique to demonstrate the possibility of the proposed decomposition.

2. 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 study, we discuss the decomposition of mixtures of the periodic signals that can be represented in the form of (1). To approximate periodic signals, we obtain a model with discretized amplitude variations. 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 a periodic component without DC component, the average of $t_p(n)$ over the interval $[0, p-1]$ is zero and the amplitude coefficients $a_{p,k}$ is restricted to non-negative values.

These p-periodic signals can also be represented in a matrix form as:

$$\mathbf{f}_p = \mathbf{A}_p \mathbf{t}_p. \tag{3}$$

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

$$\mathbf{A}_p = \left(\mathbf{D}_{p,1}, \mathbf{D}_{p,2}, \cdots, \mathbf{D}_{p,K+1}\right)^{\mathrm{T}} \tag{4}$$

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 $\mathbf{A}_p$, we define the $\mathbf{A}_p$ as an orthonormal matrix. Alternatively, p-periodic signals in (2) can be represented as

$$\mathbf{f}_p = \mathbf{T}_p \mathbf{a}_p. \tag{5}$$

In this form, the amplitude coefficients and the template are represented in a 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 $\mathbf{T}_p$. So, we define $\mathbf{T}_p$ as an orthonormal matrix.

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

$$\mathbf{f} = \sum_{p\in P} \mathbf{f}_p + \mathbf{e} \tag{6}$$

where P is a set of periods. $\mathbf{e}$ is an approximation error between the model and the signal $\mathbf{f}$. We suppose that the signal $\mathbf{f}$ is a mixture of some periodic signals that can be approximated by the form of (2), however, the periods of the source signals are unknown. So, we specify the set of periods P as a set of all possible periods of the source signals for the decomposition. If the number of the periods in P is large, the set of the periodic signals $\{\mathbf{f}_p\}_{p\in P}$ that approximate the signal $\mathbf{f}$ with small error is not unique. To achieve the significant decomposition with the periodic signals that are modeled as form of (2), we introduce the penalty of the sparsity into the decomposition.

3. SPARSE APPROXIMATE DECOMPOSITION FOR PERIODIC SIGNAL MIXTURES

3.1 Sparse Periodic Approximate Decomposition

In many sparse decomposition methods, l_1 norm plays vital role as the sparsity measure. The basis pursuit (BP) [8] decomposes a signal $\mathbf{f}$ into a linear combination of basis vectors included in an over-complete dictionary $\mathbf{\Phi}$ as:

$$\min ||\mathbf{c}||_1 \text{ subject to } \mathbf{f} = \mathbf{\Phi}\mathbf{c}, \tag{7}$$

where $||\cdot||_1$ denotes the l_1 norm of a vector. $\mathbf{\Phi}$ and $\mathbf{c}$ are the matrix that contains the normalized basis vectors and the coefficient vector, respectively. Since the l_1 norm is defined as the sum of the absolutes of the elements in the coefficient vector $\mathbf{c}$, BP determines the shortest path to the signal from the origin through the basis vectors. The number of the basis vectors with nonzero coefficients obtained by choosing the shortest path is much smaller than the least square solution obtained by minimizing the l_2 norm [8]. An approximation of the solution of BP is obtained from the penalty problem of (7) as follows:

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \frac{1}{2}||\mathbf{f} - \mathbf{\Phi}\mathbf{c}||_2^2 + \lambda||\mathbf{c}||_1 \tag{8}$$

where λ denotes a Lagrange multiplier. This unconstrained minimization problem is referred to as basis pursuit denoising (BPDN) [8]. The dictionary $\mathbf{\Phi}$ is fixed for signal representation in the BP and BPDN. In a sparse coding strategy [10], 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 [10].

For our periodic decomposition, we also impose the sparsity measure 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 basis vectors. In order to achieve this, we define the sparsity measure as the sum of l_2 norms of the periodic subsignals to find the shortest path to the approximation of the signal as well as BPDN.

With the penalty of the sparsity, we define the cost function E for the periodic decomposition as:

$$E\left(\{\mathbf{f}_P\}_{p\in P}\right) = \frac{1}{2} \parallel \mathbf{f} - \sum_{p\in P} \mathbf{f}_p \parallel_2^2 + \lambda \sum_{p\in P} \alpha_p \parallel \mathbf{f}_p \parallel_2 . \tag{9}$$

In our periodic decomposition, a signal $\mathbf{f}$ is decomposed into a set of periodic signals $\mathbf{f}_p$ that is a minimizer of E. In this cost function, the weights $\{\alpha_p\}_{p\in P}$ are included. A periodic signal with period p and a constant amplitude is also a periodic signal with a period that is a multiple of p. So, if the weight α_p is constant for all periods, the period of the p-periodic signal can be represented as any multiple of p while preserving the cost function. To avoid overestimation of the signal period, we impose the following condition on the weights:

$$\alpha_{p_1} < \alpha_{p_2} \text{ for } p_1 < p_2. \tag{10}$$

Under this condition, a periodic signal will be represented at the smallest possible period to reduce the cost function. The overestimation of periods will be avoided; however, few harmonics of the source periodic signal in signal $\mathbf{f}$ may exhibit a tendency to separate from the fundamental period.

3.2 Algorithm for Sparse Periodic Decomposition

When dictionary is a union of orthonormal bases, the solution of the BPDN can be obtained by employing the block coordinate relaxation (BCR) algorithm [9]. In this subsection, we also apply a relaxation algorithm that updates each periodic subsignal to minimize the cost function E as well as BPDN. This relaxation algorithm always updates one chosen periodic subsignal while assuming all the other periodic subsignals to be fixed. In the updating of the chosen periodic signal, the minimization of E is reduced to a scalar minimization problem. The template and the amplitude coefficients of the chosen periodic signal are alternatively updated in an iteration.

In the algorithm, we suppose that the set of periods 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 $\{\mathbf{A}_p\}_{p\in P}$
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 $\mathbf{A}_{p_i}\mathbf{t}_{p_i}$. Update the template $\mathbf{t}_{p_i}$ with the solution of a subproblem:

$$\min_{\mathbf{t}_{p_i}} \frac{1}{2} \parallel \mathbf{r} - \mathbf{A}_{p_i}\mathbf{t}_{p_i} \parallel_2^2 + \lambda\alpha_{p_i} \parallel \mathbf{A}_{p_i}\mathbf{t}_{p_i} \parallel_2 \text{ subject to } \mathbf{u}_{p_i}^{\mathrm{T}}\mathbf{t}_{p_i} = 0 \tag{11}$$

   where $\mathbf{u}_{p_i}$ is a p_i dimensional vector whose elements are unity.
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:

$$\min_{\mathbf{a}_{p_i}} \frac{1}{2} \parallel \mathbf{r} - \mathbf{T}_{p_i}\mathbf{a}_{p_i} \parallel_2^2 + \lambda\alpha_{p_i} \parallel \mathbf{T}_{p_i}\mathbf{a}_{p_i} \parallel_2 \text{ subject to } \mathbf{a}_{p_i} \geq \mathbf{0} \tag{12}$$

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).

The closed form solutions of (11) as follows:

$$\mathbf{t}_{p_i} = \mathbf{v}_{p_i} - \frac{1}{p_i}\left(\mathbf{u}_{p_i}^{\mathrm{T}}\mathbf{v}_{p_i}\right)\mathbf{u}_{p_i} \tag{13}$$

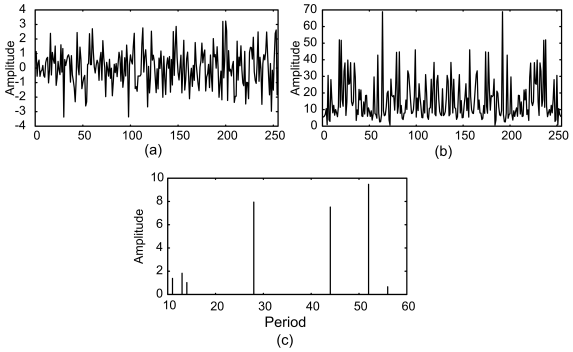

Figure 1: (a) Mixture, (b) DFT spectrum and (c) energy distribution obtained by the proposed method.

Table 1: SNR improvements(dB) obtained by the sparse periodic decomposition for mixtures of three periodic signals.

Tested set	Ave.	Max	Min
28, 44, 52	12.1, 17.0, 15.0	19.2, 22.5, 21.3	6.8, 11.8, 9.2
30, 31, 32	16.8, 19.2, 19.2	25.2. 25.7, 25.2	10.1, 12.3,12.9
50, 51, 52	12.6, 14.5, 14.1	18.9, 20.3, 19.5	8.8, 8.7, 10.0

where

$$\mathbf{v}_{p_i} = \begin{cases} \frac{||\mathbf{A}_{p_i}^{\mathrm{T}}\mathbf{r}||_2 - \lambda\alpha_{p_i}}{||\mathbf{A}_{p_i}^{\mathrm{T}}\mathbf{r}||_2}\mathbf{A}_{p_i}^{\mathrm{T}}\mathbf{r} & \text{for} \quad ||\mathbf{A}_{p_i}^{\mathrm{T}}\mathbf{r}||_2 > \lambda\alpha_{p_i} \\ 0 & \text{for} \quad ||\mathbf{A}_{p_i}^{\mathrm{T}}\mathbf{r}||_2 \le \lambda\alpha_{p_i}. \end{cases} \tag{14}$$

The j-th column of the closed form solution of (12) is

$$a_{j,p_i} = \begin{cases} w_{j,p_i} & \text{for} \quad w_{j,p_i} \ge 0 \\ 0 & \text{for} \quad w_{j,p_i} < 0 \end{cases} \tag{15}$$

where w_{j,p_i} is the j-th element of $\mathbf{w}_{p_i}$ that is computed as:

$$\mathbf{w}_{p_i} = \begin{cases} \frac{||\mathbf{T}_{p_i}^{\mathrm{T}}\mathbf{r}||_2 - \lambda\alpha_{p_i}}{||\mathbf{T}_{p_i}^{\mathrm{T}}\mathbf{r}||_2}\mathbf{T}_{p_i}^{\mathrm{T}}\mathbf{r} & \text{for} \quad ||\mathbf{T}_{p_i}^{\mathrm{T}}\mathbf{r}||_2 > \lambda\alpha_{p_i} \\ 0 & \text{for} \quad ||\mathbf{T}_{p_i}^{\mathrm{T}}\mathbf{r}||_2 \le \lambda\alpha_{p_i}. \end{cases} \tag{16}$$

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 minimization is achieved by the iteration of a simple shrinkage for the norm of the projected signal on the subspaces that are spanned by the template and the amplitude matrices. The projections can be implemented simply since both of the matrices are orthonormal.

3.3 Decomposition Results for Synthetic Signal Mixtures

In this subsection, we provide several examples of the sparse periodic approximate decomposition. The examples demonstrate the decomposition of synthetic signals generated by adding three periodic signals. The waveform of each source signal within a period is generated from Gaussian random variables. The length of the mixture N is 256. The envelopes of the three periodic signals are specified as a constant, a decreasing Gaussian $\exp(-(2n/N)^2)$ and an increasing Gaussian $\exp(-(2(n-N)/N)^2$ function, respectively. The squared norm and the average of each source periodic signal are normalized to 10.0 and 0. Since the three source periodic signals can be assumed to be independent, the SNR of each source signal in the mixture is about -3.0 dB. The sets of three periods for mixtures are shown in the first column of Table. 1. The periods for approximation are specified as integers in the range $[10, 59]$. For decomposition, the set of the weights α_p is experimentally specified as $(p/N)^{1/4}$. λ is set at 10% of the l_2 norm of the mixture. The initial amplitude coefficients $a_{p,k}$ are constant for all periods. The iteration of the relaxation method explained in Sect. 3.2 is stopped when the l_∞ norm of the difference between the periodic subsignal before and after updating is lower than a threshold value. This value is specified to be $0.001 \times \lambda$. In order to evaluate the decomposition, we compute the improvement in SNR. The improvement in SNR is computed as the difference of the SNRs of the mixture and decomposition results for each source period.

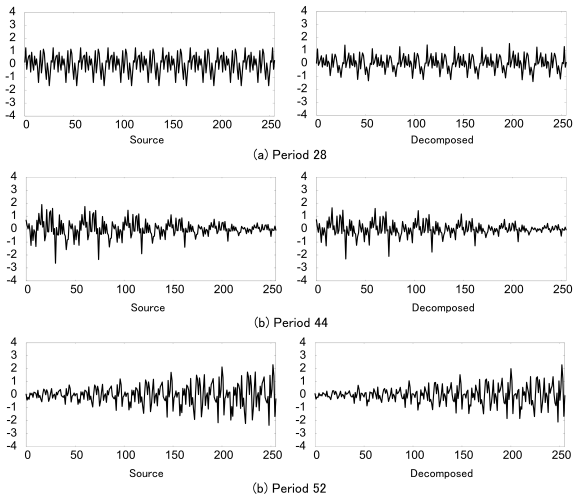

Figure 2: Source and decomposed periodic signals of the mixture shown in Fig. 1(a).

We generate 1,000 mixtures to test the decomposition algorithm for each set of periods. Table 1 shows the average, maximum and minimum of the SNR improvements of the decomposed periodic signals for 1,000 tests. The SNRs of the decomposition results are larger than 6.8dB and the average improvements exceed 12.0dB. By these results, we see that the proposed decomposition can obtain significant decomposition results and separate three sources into its periods. In Fig. 1 and 2, an example of the mixture and its decomposition result are shown. Fig. 1(a) shows a mixture consists of three source periodic signals shown in Fig. 2. The discrete Fourier transform (DFT) spectrum of the mixture is shown in Fig. 1(b). The energy distribution of the resultant periodic signals of the mixture is shown in Fig. 1(c). As we see in Fig. 1(c), three periodic signals with large amplitude appear at the source periods. Small harmonics components are separated from the source periods due to the weighting of the sparsity measure, however, the almost energy of the mixture is decomposed into the three source periods. In Fig. 2, the periodic subsignals that appear in the decomposition result are also shown. In this set of the periods, the harmonics with periods 1, 2, and 4 which are the common divisors of the source periods, cannot be separated accurately. However, the other harmonics are well collected to three fundamental periods.

4. REPRESENTATION AND SEPARATION OF SPEECH SIGNAL MIXTURES

In this section, we apply the proposed sparse decomposition to speech mixtures. The speech signals for the experiments were selected 3 Japanese male and 3 female continuous speeches of about 8 s taken from ATR-SLDB (Spoken Language Database). The sampling rate of each speech signal is converted to 8 kHz. 15 speech mixtures that consist of two different speeches that are normalized to same power are generated. For periodic decomposition, each

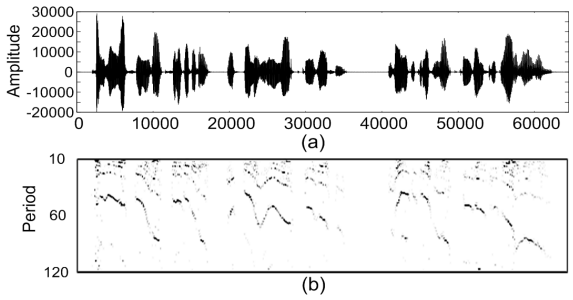

Figure 3: (a) Speech signal (male, duration: 8.1 s, sampling freq. : 8 kHz) and (b) time-period energy distribution of (a)

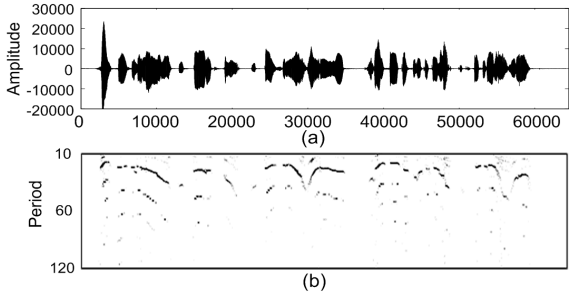

Figure 4: (a) Speech signal (female, duration: 8.1 s, sampling freq. : 8 kHz) and (b) time-period energy distribution of (a)

mixture is divided into segments that contain 360 samples with $3/4$ overlap. In each segment, the periods for decomposition are specified to be integers in the range $[10, 120]$ which correponds to the range of the fundamental frequencies of most men and women. The stopping rule of the iteration of the relaxation method and the parameters are specified as the same rules that are mentioned in Sect. 3. 3.

The examples of the male and female utterances and its time-period energy distributions are shown in Fig. 3 and Fig. 4. respectively. In Fig. 3(b) and 4(b), the brightness indicates the power of the resultant periodic signals for each segment and period. Darker pixels indicate higher powers of the resultant periodic subsignals. Our method decomposes a signal into the periodic signals with only integer periods. Under this limitation, the speech components with non-integer periods and the frequency variations that occurs in a segment are represented as the sum of some periodic signals. So, we see that the pitch contours are represented by some neighboring periods in these time-period distribution. Moreover, small periodic components with periods that are multiples and divisors of the fundamental periods appear. These periodic components appear due to the non-integer periodic components of the speeches and the weighting of the sparsity measure in (9). However, the most of the signal energy is concentrated around the fundamental pitch periods of the speeches.

We also show the time-period energy distributions of the mixture of two speeches. Fig. 5(a) and (b) show the mixture of the source speech signals shown in Fig. 3(a) and Fig. 4(a) and its time-period energy distributions, repspectively. We see that the time-period energy distribution of the mixture in Fig. 5 is almost equal to the sum of the two distributions of the source speeches shown in Fig. 3(b) and Fig. 4(b). The both of the pitch contours of the two source speeches are preserved in the distribution of the mixture. The proposed decomposition method can approximate the mixture while concentrating the energy of each speech to its pitch periods and provides sparse representation of the mixture. It is expected that the pitch periods of both the speech signals will be tracked in

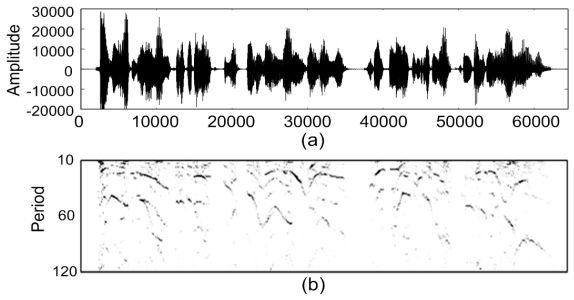

Figure 5: (a) Mixture of female and male speeches and (b) time-period energy distribution of (a)

Table 2: Average, minimum and maximum SNRs (dB) of approximated speech segments and average numbers of periodic signals obtained by the sparse decomposition

Speakers	Ave. SNR	Min. SNR	Max. SNR	Ave. num. of periods
(M, M)	20.1	10.4	28.9	16.1
(F, F)	20.2	10.3	27.6	11.2
(F, M)	20.2	10.2	28.9	14.0

this time-period energy distribution. Moreover, speech separation will be achieved by assigning the resultant periodic signals to the sources. In order to evaluate the approximate decomposition, we compute the SNR and the number of the non-zero resultant periodic signals for each segment where the l_2 norm is greater than the noise level. The average, maximum and minimum SNRs over all voice active segments of mixtures are shown in Table 2. In this table, F an M denote female and male source speeches, respectively. The average numbers of periods for approximation of a segment are also shown. We see that the average approximation precision of the proposed decomposition is about 20 dB in the segmental SNR. The average number of the periods yield by the decomposition is about 14 for speech mixtures consist of two speeches. Next, we demonstrate the speech separation from a mixture with the sparse periodic decomposition. In this experiment, the speech separation is performed by allocating of the resultant periodic signals to the sources in each segment. First, we use the clean source signals for assignment of the resultant periodic signals. The separtion is carried out by the following steps in each segment:

1. The segment of the mixture is decomposed into the set of the periodic signals $\{\mathbf{f}_p\}_{p\in P}$.
2. The normalized correlations between the resultant periodic signals and the clean source segments $\{\mathbf{s}_i\}_{i=1,2}$ are computed.
3. Each resultant periodic signal $\mathbf{f}_p$ are added to the separated output that is associated with the i-th source $\mathbf{s}_i$ that obtains larger correlation.

For recovering source signals, we apply the Hamming window to the resultant periodic subsignals in each segment. This assignment method does not obtain optimum separated results in terms of the SNR exactly. However, this experiment gives the rough ideal performance of the source separation by using the proposed sparse decomposition.

For comparison, the ideal separation results that are obtained by a STFT that is widely utilized for the sparse representation of speech signals are demonstrated. In the separation with the STFT, the ideal binary masks [11] are computed from the clean source speeches. The mixture and the source signals are segmented by 512 points Hamming window with $3/4$ overlap. In each segment, the DFT spectrum of the mixture and the source signals are computed. Each frequency bin of the DFT is assigned to the source whose amplitude

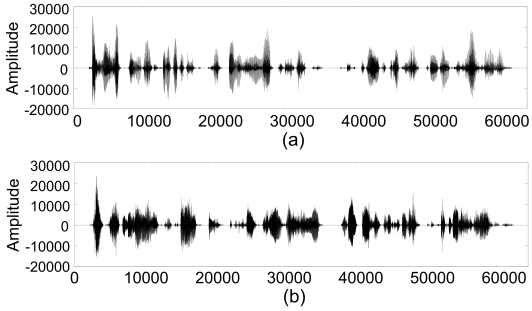

Figure 6: Separated speech signals obtained by sparse periodic decomposition with reference speeches, (a) separated male speech (SNR: 7.2dB) and (b) female speech (SNR: 7.1dB) from the mixture shown in Fig. 5(a)

Table 3: Average SNRs (dB) of separated signals

Speakers	Proposed (with source)	DFT (with source)	Proposed (with ref. sig.)
(M, M)	9.9 ± 0.6	13.5 ± 0.5	3.9 ± 1.0
(F, F)	9.5 ± 0.3	13.5 ± 0.5	3.2 ± 0.9
(F, M)	F: 10.1 ± 1.5 M: 9.8 ± 1.0	F: 14.4 ± 1.0 M: 14.3 ± 1.0	F: 6.5 ± 2.5 M: 6.7 ± 2.5

of the frequency bin is larger than the other. The separation results obtained by the proposed decomposition and the DFT are shown in Table 3. In this table, the SNRs of the separated speech signals are shown. We see that the SNRs of the separated speeches obtained by the proposed method are lower than the DFT by about 3.5 to 4.6dB. In the separation obtained by the the proposed method, the approximation errors caused during the decomposition are involved in the separated output. Since the frequency resolution of the periodic decomposition is lower than the DFT at high frequency bands, the interferences between two speeches occur at high-frequencies.

However, the proposed representation is sparser than the DFT spectrum. In this experiment, the DFT yields 257 frequency bins for each segment. So, the DFT based separation is the problem of the assignment of the 257 frequency bins. In contrast, the average number of the periodic signals yield by the proposed method is about 14 for a segment. Comparing the proposed decomposition with the DFT, the separation problem can be reduced to relatively small size of a combination problem by the proposed decomposition.

In above experiments, we assume that the source speeches are known. Next, we demonstrate the single-channel speech separation by referencing the clean speech segments. In this scenario of the separation, two speakers in a mixture are known and the clean speeches of the speakers are available, but the contents of the speeches in the mixture are unknown. In order to assign the periodic signals to the sources, a set of the clean speech segments of the i-th speaker is defined as $\{\mathbf{c}_{i,j}\}_{1 \le j \le N_r}$ where N_r is the number of the reference segments. The resultant periodic signal $\mathbf{f}_p$ is assigned to the i-th speaker that gives the maximum of the normalized correlation as:

$$\max_{i,j} \frac{\mathbf{f}_p^{\mathrm{T}} \mathbf{c}_{i,j}}{\| \mathbf{f}_p \|_2 \| \mathbf{c}_{i,j} \|_2}. \tag{17}$$

For this experiment, segments that are generated from a clean speech of 20 s are used for the references $\{\mathbf{c}_{i,j}\}_{1 \le j \le N_r}$ of each speaker. The segments where the voice is not active are rejected from the references. The references do not include the source utterances in the mixtures. The SNRs obtained by the separation with the references are also shown in Table. 3. Obviously, such a simple separation method causes many false assignments. For separation of the mixture consists of the speakers of same gender, the averages of the improvements of SNR are lower than 4dB. However, the averages of SNR close to the ideal results and are about 6.5dB for the speakers of opposite gender. The separated signals from the mixture in Fig. 5(a) are shown in Fig. 6(a) and (b). Audio examples are available at http://www-analab.sys.es.osaka-u.ac.jp/~nkszk/SPD/.

The single channel speech separation methods based on frequency spectrum have been proposed [6, 7]. In these methods, statistical models for the frequency spectra of the speakers are generated. The separation is performed on the frequency spectrum of the mixture by using the statistical models. In our approach, the proposed sparse decomposition yields the small number of the periodic signals which approximate the source signal due to the sparsity measure. So, the separation of the speeches that have less similarity can be performed by such a lazy assignment method.

5. CONCLUSIONS

In this paper, we proposed a sparse decomposition method for periodic signal mixtures. The proposed decomposition is based on the model for the periodic signals with time-varying amplitude and the sparsity of the periods that appear in the result. In decomposition experiments of the synthetic signal and the speech mixtures, we demonstrated that the proposed decomposition has the ability of source separation.

The assignment method that is employed for the single-channel speech separation demonstrated in this paper is too simple to obtain good separation results. In our decomposition results, as seen in the figures in Sect. 4, the speech pitch contours are involved. We can use the continuity of the speech pitches and the similarity of the templates over the consecutive segments for improvement of the accuracy of the assignment. The accurate and robust assignment of the decomposed periodic signals is a topic for future research.

REFERENCES

[1] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation, " *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 8, pp. 744-754, Aug. 1986.

[2] M. Triki and D. T. M. Slock, "Periodic signal extraction with global amplitude and phase modulation for musical signal decomposition, " *Proc. on ICASSP*, vol. 3, pp. 233-236, 2005.

[3] W. A. Sethares and T. W. Staley, "Periodicity transform," *IEEE Trans. on Signal Processing*, vol. 47, no. 11, pp. 2953-2964, Nov. 1999.

[4] D. D. Muresan and T. W. Parks, "Orthogonal, exactly periodic subspace decomposition, " *IEEE Trans. on Signal Processing*, vol. 51, no. 9, pp. 2270-2279, Nov. 2003.

[5] M. Nakashizuka, "A sparse decomposition method for periodic signal mixtures, " *IEICE Trans. on Fundamentals*, Vol.E91-A, No.3, pp. 791-800, March 2008.

[6] S. T. Roweis, "One microphone source separation, " *Adv. Neural Inf. Process. Syst.,* vol. 13, pp. 793-799, 2001.

[7] 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.

[8] 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.

[9] S. Sardy, A. G. Bruce and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising, " *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 361-379, 2000.

[10] M. S. Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes, " *Journal of the Optical Society of America*, vol. 16, no. 7, pp. 1587-1601, July 1999.

[11] 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.