The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Exploiting Harmonic Structures to Improve Separating Simultaneous Speech in Under-Determined Conditions

Yasuharu Hirasawa, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan
{*hirasawa, tall, komatani, ogata, okuno*}*@kuis.kyoto-u.ac.jp*

***Abstract*— In real-world situations, a robot may often encounter "under-determined" situation, where there are more sound sources than microphones. This paper presents a speech separation method using a new constraint on the harmonic structure for a simultaneous speech-recognition system in under-determined conditions. The requirements for a speech separation method in a simultaneous speech-recognition system are (1) ability to handle a large number of talkers, and (2) reduction of distortion in acoustic features. Conventional methods use a maximum likelihood estimation in sound source separation, which fulfills requirement (1). Since it is a general approach, the performance is limited when separating speech. This paper presents a two-stage method to improve the separation. The first stage uses maximum likelihood estimation and extracts the harmonic structure, and the second stage exploits the harmonic structure as a new constraint to achieve requirement (2). We carried out an experiment that simulated three simultaneous utterances using impulse responses recorded by two microphones in an anechoic chamber. The experimental results revealed that our method could improve speech recognition correctness by about four points.**

I. INTRODUCTION

Since people currently have increasing opportunities to see and interact with humanoid robots, *e.g.*, the Honda ASIMO [1], Kawada HRP [2], and KOKORO Actroid [3], verbal communication is critical in accomplishing symbiosis between human and humanoid robots in everyday life. For example, verbal communication is the most effective way of interaction when we ask a robot to do housework, or when a robot informs us about what happened today.

A robot's capabilities are quite unbalanced in verbal communication. Robots can speak very fluently and sometimes in a fully emotional way thanks to up-to-date text-to-speech systems. In contrast, they cannot hear well due to limited automatic speech recognition (ASR). Poor ASR is mainly caused by interfering sounds, *e.g.*, other people speaking at the same time, air-conditioning, the robot's own cooling fans, and the robot's movements. In other words, robots or people usually hear a mixture of sounds and thus the robots' audition should separate speech signals from this mixture of sounds.

Research on robot audition has mainly focused on sound-source localization, sound-source separation, and the recognition of speech or other sound sources. The most common underlying assumption in robot audition is that the number of sound sources should not exceed the number of microphones. Independent Component Analysis (ICA) [4] for sound-source separation assumes that there are equal numbers of sound sources and microphones. Beamforming [5] usually assumes that the number of microphones is larger, *i.e.*, an "over-determined" situation.

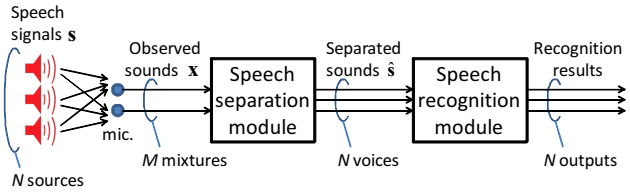


Fig. 1. Simultaneous speech-recognition system

Robots working with humans often encounter an "under-determined" situation, *i.e.*, there are more sound sources than microphones. Since previous systems assume non-under-determined conditions, robots that work in real situations need many microphones. However, deploying many microphones is unfavorable from the viewpoint of space needed to deploy these microphones, the cost in using a multi-channel synchronizer, and the satisfaction of working in real time. Furthermore, even though robots have many microphones, we cannot be sure that there are not more sound sources than microphones.

This paper focuses on the method for under-determined speech separation that can be used for a simultaneous speech-recognition system. Generally speaking, a simultaneous speech-recognition system consists of a speech-separation module and a speech-recognition module, as shown in Fig. 1. When we consider under-determined conditions, it is difficult to develop the speech-separation module because this module has more outputs than inputs.

There have been some methods for speech separation in under-determined conditions. Yılmaz *et al.* [6] used a time-frequency mask that was estimated by using histogram clustering assuming that at most one speech was dominant in each time-frequency region. Nakadai *et al.* [7] also used a time-frequency mask that employed an active direction pass filter assuming sparseness of talker directions. Lee *et al.* [8] estimated a mixing matrix and speech signals concurrently using a maximum a posteriori estimation. Bofill *et al.* [9] and Zhang *et al.* [10] introduced a two-stage approach, which meant the estimation of mixing matrix and that of the speech signals were done in series, assuming that each time-frequency region contains not more dominant sources than the number of microphones.

These methods use the characteristics of the power distri-

bution of voices: when we convert human voices into time-frequency expressions, only a few time-frequency regions have high power and most of the time-frequency regions have low power. When we take into consideration this characteristic, we can assume that each time-frequency region has only a few "dominant sources", which are sound sources that have relatively high power in one time-frequency region. Using this assumption, we can handle under-determined speech separation as if there are fewer talkers than microphones by ignoring non-dominant utterances.

When we use this characteristic, it is very important to estimate the dominant sources because one mis-estimation affects all separation results. There are two specific types of effects. The first is the lack of spectra in the mis-estimated source because the power of non-dominant sources is regarded to be zero. The second is the leakage noise of other sources that is derived from the residual power of the mis-estimated source.

We propose a method with constraints using the harmonic structure to improve the accuracy of dominant source estimation. Since most methods for under-determined speech separation do not effectively use the characteristics of human voices, the output sounds with these methods are likely to be too distorted to recognize. We use the L1-norm method, which can handle many sound sources, and add new constraints using the harmonic structure to maintain acoustic features and improve ASR results.

The rest of this paper is organized as follows: Section II focuses on under-determined speech separation, and describes problem setting, requirements, an L1-norm separation method, and problems with that method. Section III proposes an under-determined speech separation method that uses the constraints of the harmonic structure. Section IV presents the results of experiments, and Section V summarizes this paper.

II. UNDER-DETERMINED SIMULTANEOUS SPEECH SEPARATION

This section presents the problem settings, requirements, and an L1-norm method for separating under-determined speech, and we explain some problems with that method in simultaneous speech-recognition systems.

A. Problem Settings for Underdetermined Speech Separation

The problem settings of speech separation used in this paper are as follows. Note that $N > M$ because we have considered under-determined conditions.

Input	M mixtures of N simultaneous utterances.
Output	Each speech signal of N talkers.
Assumption	Mixing matrix $\boldsymbol{H}$ is known.

We add a supplementary explanation to the assumption. Since our method is based on one of the L1-norm methods for separation [9] [11], which needs mixing matrix $\boldsymbol{H}$, our proposed method also needs mixing matrix $\boldsymbol{H}$. To satisfy this requirement, we can measure the transfer function in advance. If this is impossible, we can use the method to estimate the mixing matrix from observed signals [12].

TABLE I
DEFINITION OF VARIABLES

N	Number of talkers
M	Number of microphones ($M < N$)
t	Time frame index
f	Frequency bin index
$s_j(t)$	Speech signal of talker j in time frame t
$\boldsymbol{s}(t)$	$[s_1(t), s_2(t), ..., s_N(t)]^T \in \mathbf{C}^N$
h_{ij}	Transfer function from talker j to microphone i
$\boldsymbol{h}_j$	$[h_{1j}, h_{2j}, ..., h_{Mj}]^T \in \mathbf{C}^N$
$\boldsymbol{H}$	Matrix consists of h_{ij}
$x_i(t)$	Observed signal of microphone i in time frame t
$\boldsymbol{x}(t)$	$[x_1(t), x_2(t), ..., x_M(t)]^T \in \mathbf{C}^N$
K	Set of indices of dominant sources
k_i	Index of i-th dominant source

B. Requirements for Underdetermined Speech Separation

There are two main requirements for under-determined speech separation.

1) Ability to handle a large number of talkers
2) Reduction of distortion in acoustic features

We will now discuss these two requirements.

1) Ability to Handle a Large Number of Talkers: Since our aim is to develop a system that can be used when there is a lot of sound, our method have to be able to handle many sources in each time-frequency region. Many separation methods for under-determined conditions have assumed that the number of dominant sources in one time-frequency is a small constant to simplify the separation. However, when the number of talkers increases, each time-frequency region contains more and more number of dominant sources. This means that a separation method must not have too strict assumption about the number of sources in each time-frequency region.

2) Reduction of Distortion in Acoustic Features: Since we aim at developing a system that can recognize simultaneous utterances, the separation results are used for ASR. Since the speech-recognition module calculates several values, called acoustic feature values, in order to recognize easily, our separation method must keep these acoustic features of the original speech signals. Since we use the Mel-Frequency Cepstrum Coefficient (MFCC) [13] as the feature value, separating the following two areas is especially important.

1) Time-frequency regions that have high power
2) Time-frequency regions that are low in frequency

The former comes from the fact that MFCC is calculated by the sum of the power spectra of each frequency band, and the latter is comes from the fact that MFCC is calculated using mel-scaled frequency.

C. L1-Norm Method of Speech Separation

Let us first describe the speech mixing process. We modelize the sound-transfer function as a linear time-invariant

function and use time-frequency expression to separate mixtures because this enables speech separation to be considered independently in each time-frequency region. Table I defines the basic variables used in this paper. Using the Short-time Fourier Transform (STFT), the speech-mixing process can be written in the time-frequency domain as follows.

$$\boldsymbol{x}(f,t) = \sum_{j=1}^{N} \boldsymbol{h}_j(f) s_j(f,t) \quad (1)$$

$$= \boldsymbol{H}(f)\boldsymbol{s}(f,t) \quad (2)$$

To determine the proper STFT frame length, we use the knowledge by Yılmaz *et al.* [6].

Now, we will explain the L1-norm separation method [11] that is based on the assumption that the power distributions of speech signals are Laplacian distributions. This method also assumes that the number of dominant sources in one time-frequency region is at most M, *i.e.*, the number of microphones. Note that we omit f and t from the formulas because the calculation will be done independently in each time-frequency region.

First, let us explain the separation method when M dominant sources are known. When the set of indices of these M dominant sources are expressed as $K = \{k_1, k_2, ..., k_M\}$, Eq. (1) can be written as follows by ignoring low-power speech signals, which indices are not in K.

$$\boldsymbol{x} = \sum_{u=1}^{M} \boldsymbol{h}_{k_i} s_{k_i} \quad (3)$$

$$= \boldsymbol{H}_K \boldsymbol{s}_K \quad (4)$$

Here, $\boldsymbol{H}_K = [\boldsymbol{h}_{k_1}, \boldsymbol{h}_{k_2}, ..., \boldsymbol{h}_{k_M}]$ is the part of the original mixing matrix, $\boldsymbol{H}$, and $\boldsymbol{s}_K = [s_{k_1}, s_{k_2}, ..., s_{k_M}]^T$ represents the speech signals of dominant sources. Since $\boldsymbol{H}_K$ is an $M \times M$ square matrix, we can separate mixtures as follows.

$$\hat{\boldsymbol{s}}_K = \boldsymbol{H}_K^{-1}\boldsymbol{x}, \quad (5)$$

$$\hat{s}_i = 0 \qquad \forall i \notin K, \quad (6)$$

where $\hat{\boldsymbol{s}}_K = [\hat{s}_{k_1}, \hat{s}_{k_2}, ..., \hat{s}_{k_M}]^T$ and the separation result is represented as $\hat{\boldsymbol{s}}'_K = [\hat{s}_1, \hat{s}_2, ..., \hat{s}_N]^T$.

Next, we will explain how M dominant sources are chosen. This is based on the assumption that the speech-power distributions follow Laplace distributions. When each talker's distribution follows the same Laplace distribution independently, logarithm of joint probability is expressed as follows.

$$\log p(\boldsymbol{s}) = -\lambda \sum_{k=1}^{N} |s_k| + C \quad (7)$$

where $\boldsymbol{s} = [s_1, s_2, ..., s_N]^T$, λ is a positive parameter of the Laplace distribution, and C is a constant value. After an observation, since we can use Eq. (2), logarithm of posterior probability can be written as follows.

$$\log p(\boldsymbol{s}|\boldsymbol{x}) = \begin{cases} -\lambda \sum_{k=1}^{N} |s_k| + C' & (\boldsymbol{x} = \boldsymbol{H}\boldsymbol{s}) \\ -\infty & (otherwise) \end{cases} \quad (8)$$

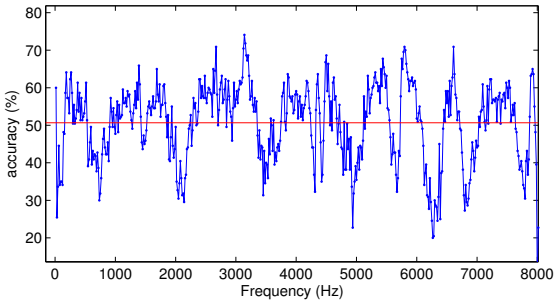


Fig. 2. Accuracy of estimating dominant sources

where C' is an another constant value.

When all elements are real numbers, $\boldsymbol{s}$ that maximizes Eq. (8) will be $\hat{\boldsymbol{s}}'_K$ when we choose K optimally. The most naive method to find optimal K is to calculate $\hat{\boldsymbol{s}}_K$ with all variation of K, and determine K that maximizes Eq. (8) as K_{opt}. This is written as the combinatorial optimization problem as follows.

$$K_{opt} = \underset{K}{\operatorname{argmin}} \sum_{i=1}^{M} |\hat{s}_{k_i}| \quad (9)$$

where

$$\hat{\boldsymbol{s}}_K = \boldsymbol{H}_K^{-1}\boldsymbol{x} \quad (10)$$

$$\boldsymbol{H}_K = [\boldsymbol{h}_{k_1}, \boldsymbol{h}_{k_2}, ..., \boldsymbol{h}_{k_M}] \quad (11)$$

$$K = \{k_1, k_2, ..., k_M\} \quad (12)$$

$$(1 \le k_1 < k_2 < ... < k_M \le N)$$

Since our method uses a time-frequency expression, elements are not always real numbers but they can be complex numbers. When the elements are complex numbers, the above combinatorial optimization does not theoretically maximize Eq. (8). However, Winter *et al.* [14] demonstrates that this combinatorial optimization can be solved much more quickly than a strict solution and the solution is similar to the strict one even when elements are complex numbers. Thus we use the above combinatorial optimization to obtain the solution in this paper.

We will now discuss the problem when we use this speech separation method in a simultaneous speech-recognition system. Since this method can handle at most M dominant sources in each time-frequency region, it can deal with speech mixtures that contain many utterances. However, even when the mixing matrix is known, the accuracy of dominant source estimation depends on the power distribution of the original speech signals [12].

In addition, the accuracy of dominant source estimation. depends on the frequency, because the accuracy depends on the mixing matrix, and the mixing matrix depends on the frequency bin. Fig. 2 show the relation between frequency and accuracy, and the horizontal line plots the average accuracy. This experiment was carried out under the same conditions as those in Section IV, *i.e.* we simulated the

situation where three simultaneous utterances were recorded by two microphones in an anechoic chamber, and separate it using Eq. (9) - (12). The results of estimates were considered to be accurate only when all M, *i.e.* 2, dominant sources were correctly estimated. As Fig. 2 indicates, accuracy of dominant source estimation is much different between frequency bands with poor accuracy such as 0-200 Hz and 600-1000 Hz, and frequency bands with good accuracy such as 200-600 Hz and 2500-3200 Hz. Even when we use different impulse responses and different speeches to synthesize speech mixtures, separated results also have frequency bands with poor accuracy and ones with good accuracy.

As we have stated in Subsection II-B, we need a separation method that satisfies (1) ability to handle a large number of talkers and (2) reduction of distortion in acoustic features. This method satisfies requirement (1); however, it does not satisfy requirement (2), because the poor accuracy of dominant source estimation makes spectrum lacks and leakage noise, and this interference greatly distorts the acoustic features. To improve the system's ASR results, it is necessary to reduce this interference.

III. UNDER-DETERMINED SPEECH SEPARATION USING HARMONIC STRUCTURE

A. Our Approach Using Harmonic Structure

We focus on the harmonic structure of speech sounds and separate speech mixtures with a new constraints using harmonic structure. There are four main reasons we focus on the harmonic structure.

1) Separation results with highly accurate estimates can be used to modify separation results with inaccurate estimates since the harmonic structure has an overtone structure.
2) We can extract the harmonic structure relatively easily because it does not overlap frequently since it has a sparse power distribution.
3) The proper separation of the harmonic structure, which has high power, is important since MFCC depends on high-power time-frequency regions.
4) The harmonic structure, most of which is contained in low-frequency areas, is important since MFCC is calculated using mel-scaled frequency.

We present a method that is based on the L1-norm separation explained in Subsubsection II-C and we use this as a baseline. We found that this L1-norm separation method has a disadvantage in that the accuracy of dominant source estimation depends on the frequency, and that causes acoustic feature values to be distorted. We improve this situation using constraints that involve the harmonic structure, and obtain results with less distortion. Note that our method does not need harmonic structure as an additional input, because it can estimate the harmonic structure internally.

B. Outline of Our Method

Fig. 3 outlines our proposed method. First, we use the L1-norm method and obtain a tentatively separated sounds. Second, we extract the harmonic structure from these sounds.

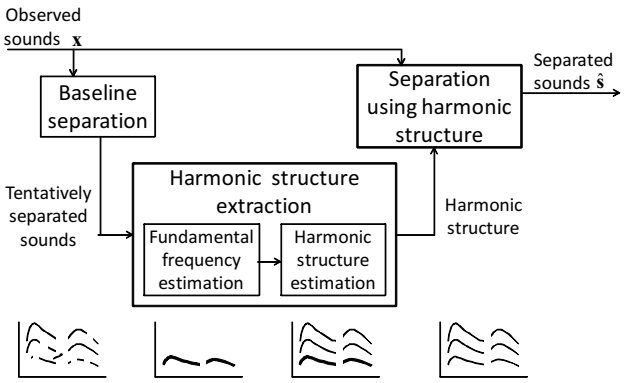


Fig. 3. Outline and flow of our method

Finally, we separate the speech mixtures again, with constraints that involved the harmonic structure. We will refer to these two separations as first and second separations. The first separation is only done to obtain the tentatively separated sounds in order to extract the harmonic structure, and the second separation is done to obtain the output sounds.

Our method divides the extraction of the harmonic structure into two phases: the first is the phase to estimate the fundamental frequency, and the second is the phase to estimate the harmonic structure. The reason we use this two-phase approach is that the tentative sound is the sound separated by the L1-norm method, *i.e.*, sound with spectrum lacks and leakage noise. We improve the robustness for extracting the harmonic structure by using these two phases.

Another arrangement with our method is to make the second separation similar to the first one. Speech separation should be done quickly because our system will be deployed in robots. Since we make the second separation similar to the first, we can reuse the calculated results from the first separation, and reduce the additional cost in time to use the harmonic structure.

Since the essential part of our method is the second separation, we first introduce this separation, and after that, we introduce the method for extracting the harmonic structure robustly.

C. Separation with Constraints

This subsection introduces the separation method when the harmonic structure is already known. Since the harmonic structure has high power, we add constraints where the set of dominant sources, K, must include sources that have a harmonic structure in the time-frequency region. In other words, when we define P as the set of sources that has a harmonic structure in the time-frequency region, $P \subseteq K$ must be true. When more than M harmonic structures exist in one time-frequency region, ($|P| > M$), we cannot define K by using the above constraints. In this case, we can use constraints where sources without a harmonic structure in the time-frequency region are not included in K, *i.e.*, K must be included in P ($P \supset K$).

Using the above constraints, the combinatorial optimization problem can be written as follows.

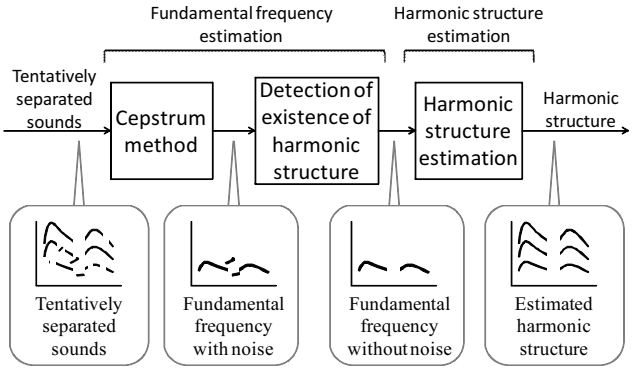


Fig. 4. Dataflow for extraction of the harmonic structure

$$K_{opt} = \underset{K}{\operatorname{argmin}} \sum_{i=1}^{M} |\hat{s}_{k_i}| \quad (13)$$

where

$$\hat{\boldsymbol{s}}_K = \boldsymbol{H}_K^{-1}\boldsymbol{x} \quad (14)$$

$$\boldsymbol{H}_K = [\boldsymbol{h}_{k_1}, \boldsymbol{h}_{k_2}, ..., \boldsymbol{h}_{k_M}] \quad (15)$$

$$K = \{k_1, k_2, ..., k_M\} \quad (16)$$

$$(1 \leq k_1 < k_2 < ... < k_M \leq N)$$

$$P \subseteq K \quad (|P| \leq M) \quad (17)$$

$$P \supset K \quad (|P| > M) \quad (18)$$

The difference between this combinatorial optimization problem and the L1-norm method discussed in Subsection II-C is existence of Eq. (17) and (18). When we take into consideration the time-frequency region which does not have harmonic structure ($P = \phi$), these constraints do not take effect; thus, we can reuse the separation results in the first separation. In addition, even when we consider time-frequency $P \neq \phi$, we can omit the matrix operations in Eq. (14) by memorizing and reusing the calculation results in the first separation for each K.

D. Extraction of Harmonic Structure Using Fundamental Frequency

Fig. 4 shows the flow of the harmonic structure extraction. In this paper, we first estimate the fundamental frequency using the cepstrum and detect the existence of the harmonic structure. After that, we estimate the shape of the harmonic structure using the estimated fundamental frequency.

1) Fundamental Frequency Estimation Using Cepstrum: Here, we introduce fundamental frequency estimation using the cepstrum. In this method, the fundamental frequencies in each time frame are estimated independently. By applying a Discrete Fourier Transform (DFT) to the logarithm power spectrum of one time frame, we obtain a quefrency expression. Since we have to handle human speech signals, we choose a quefrency whose values are maximized in the quefrency region corresponding to 80-350 Hz, which contains almost all the fundamental frequency. Since the quefrency value has a positive correlation to the harmonic structure, a large quefrency value means the existence of a harmonic structure.

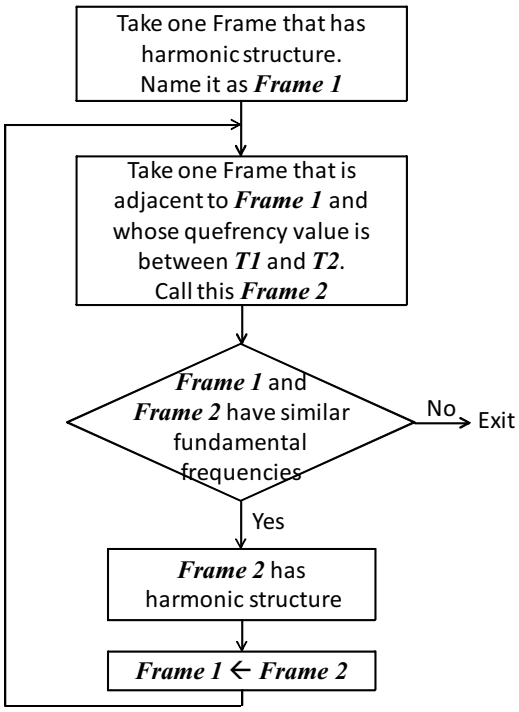


Fig. 5. Flowchart for detection of harmonic structure

2) Two-Threshold Method to Detect Existence of Harmonic Structure: Here, we introduce a method for detecting the existence of a harmonic structure. Since we extract the harmonic structure from sound that is separated by the L1-norm method, *i.e.*, sound with spectrum lacks and leakage noise, it is difficult to detect the existence of a harmonic structure. If we have clean speech signals, we can use a simple threshold for the quefrency value because the quefrency value of time frame without a harmonic structure is quite small. In our case, however, we only have speech signals with interference. This means that the quefrency value of the time frame without a harmonic structure might be high because of leakage noise, and the quefrency value of the time frame with the harmonic structure might be low because of the lack of spectra.

This paper proposes a method that uses two thresholds and the relation of the fundamental frequency between two consecutive time frames. The two thresholds are called *T1* and *T2* . *T1* is large enough and *T2* is small enough to detect the existence of the harmonic structure: if the quefrency value is more than *T1* , we consider that the time frame has a harmonic structure, and if the quefrency value is less than *T2* , we consider that the time frame does not have a harmonic structure. If the quefrency value is between *T1* and *T2* , we use the fundamental frequency to detect the existence of a harmonic structure. Details on this algorithm are below.

Fig. 5 has a the flowchart for this method. First, we take a time frame whose quefrency value is more than *T1* , and call this time frame *Frame 1* . Next, we take an adjacent

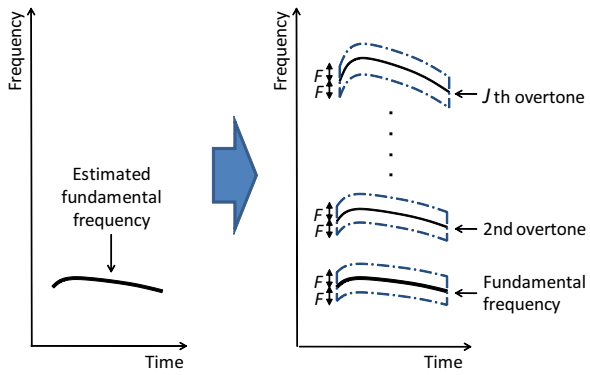


Fig. 6. Estimation of shape of harmonic structure

time frame whose quefrency value is between *T1* and *T2* , and call this time frame *Frame 2* . Now, there are two possibilities for *Frame 2* : the first is that while *Frame 2* has a harmonic structure, the lack of spectra decreases the quefrency value of this frame. The second is that while *Frame 2* does not have a harmonic structure, leakage noise increases the quefrency value of this frame. Since we cannot decide which is true by only using the quefrency value, we use the estimated fundamental frequencies of *Frame 1* and *Frame 2* . Since we know *Frame 1* has a harmonic structure, if the fundamental frequencies of *Frame 1* and *Frame 2* are similar, *Frame 1* and *Frame 2* have a connected harmonic structure. On the other hand, if the fundamental frequencies of these two frames are different, *Frame 2* does not a have harmonic structure and it contains leakage noise. This is how we detect the existence of a harmonic structure in *Frame 2* . In addition, when this method finds *Frame 2* has a harmonic structure, we overwrite variable *Frame 1* with variable *Frame 2* and apply this method recursively. This enables us to detect a harmonic structure when there is a large amount of interference.

3) Estimating Shape of Harmonic Structure: While a harmonic structure has an overtone structure, we can estimate the shape of the harmonic structure from the estimated fundamental frequency. We think about up to the Jth overtone of the fundamental frequency, and regard F frequency bins from center of each overtone are in the harmonic structure. Fig. 6 shows this estimation process visually. The black bold line means the estimated fundamental frequency, and the black thin lines indicate the overtones. We consider the areas surrounded by the blue lines to be time-frequency regions that have a harmonic structure. After estimating the harmonic structure, we separate the speech mixtures using the method we proposed in Subsection III-C.

IV. EXPERIMENTS

To verify improvements obtained with our proposed method, we carried out two experiments using synthesized sounds. Table II lists the experimental conditions and Fig. 7 outlines the arrangement for the microphones and loud speakers.

TABLE II
EXPERIMENTAL CONDITIONS

N, M	3 talkers, 2 microphones
Sampling frequency	16 kHz
Impulse response	Recorded in anechoic chamber
Sound sources	JNAS 200 sentences (males and females)
Talkers' loudness	Same loudness
STFT frame length	1024 points (64 ms)
STFT shift width	256 points (16 ms)
Other parameters	*T1* =0.03, *T2* =0.01, J=6, and F=2
Speech recognizer	Julius 3.5 fast
Acoustic model	PTM triphone, 3 state HMM
Language model	Statistical model, 20k words
Acoustic feature	MFCC 12 + ΔMFCC 12 + ΔPow
Analysis window size	400 points (25 ms)
window shift size	160 points (10 ms)

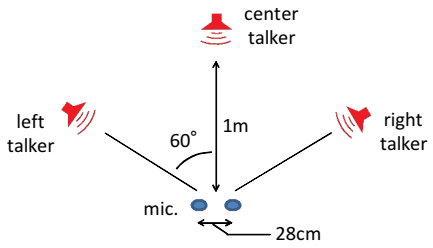


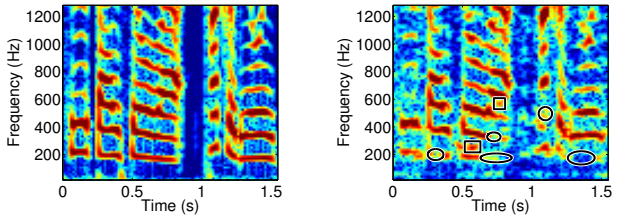
Fig. 7. Arrangement of microphones and talkers

Our evaluations use two kinds of measurements: the first is the signal-to-noise ratio to establish whether our method can accurately separate mixed speech signals, and the second is ASR correctness to establish whether output sounds are suitable for ASR.

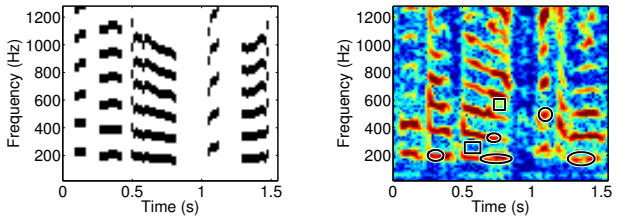
A. Evaluation Using Signal-to-Noise Ratio

Fig. 8 shows the original signal, the separation results with the L1-norm method, estimated harmonic structure, and the separation results with our proposed method. It only shows the low-frequency region because most of the estimated harmonic structure is in the low-frequency region. The black in the lower left figure means the estimated harmonic structure, and blue means low power and red means high power in the other three figures. The results from the L1-norm method, 8(b), indicate that since the accuracy of estimating the dominant source was poor in the 0-200 Hz region, there are some spectrum lacking in the areas surrounded by the black circles. Additionally, there is some leakage noise in the areas surrounded by the black rectangles. However, in the results obtained with our proposed method, 8(d), the spectra in the black circles are recovered, and leakage noise in the black rectangles is reduced. This means that our method improves the accuracy of dominant source estimation and reduces the interference from other talkers.

Fig. 9 has a histogram for the signal-to-noise ratio of the center talker. We can see that there is a peak for the



(a) Original speech (b) Results with L1-norm method



(c) Estimated harmonic structure (d) Results with proposed method

Fig. 8. Spectrograms and estimated harmonic structure

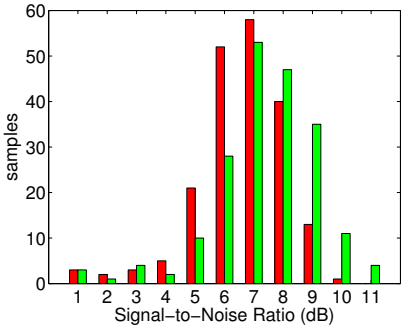


Fig. 9. Distribution of signal-to-noise ratio (Red:baseline Green:proposed)

L1-norm method near 6-7 dB, and there is a peak for our proposed method near 7-8 dB; thus, our method seems to improve the signal-to-noise ratio. However, when we take into consideration the 1-3 dB region, there are very few improvements. This is because our method cannot estimate the harmonic structure accurately when the output of L1-norm method is severely broken.

Table III lists the average signal-to-noise ratio for each talker. When we take into consideration the center talker, the improvement with our method is about 0.8 dB, which is similar to the movement of the peak in Fig. 9. However, when we take into account the left and right talkers, the improvements in the signal-to-noise ratio are only about 0.2 dB.

B. Evaluation Using Continuous Speech Recognition

Table IV lists the average ASR correctness. Note that our evaluation did not use isolated word recognition, but continuous speech recognition. In this table, "(c) optimum harmonic" means the correctness when a harmonic structure was given in all time frames. "(d) optimum all TF" means the correctness when dominant sources were given in all time-frequency regions; thus, this is the upper bound for

TABLE III

AVERAGE SIGNAL-TO-NOISE RATIO (DECIBEL)

Speaker	Left	Center	Right
(a) L1-norm method	8.2	6.6	8.6
(b) Proposed method	**8.4**	**7.4**	**8.8**
Improvements	0.2	0.8	0.2

TABLE IV

AVERAGE ASR CORRECTNESS (%)

Talker	Left	Center	Right
(a) L1-norm method	64.9	59.7	69.6
(b) Proposed method	**69.0**	**63.6**	**71.5**
(c) Optimum harmonic	74.4	68.7	77.5
(d) Optimum all TF	82.3	82.5	85.0

ASR correctness under our experimental condition. The difference between the "(a) L1-norm method" and the "(d) optimum all TF", about 15-23 points, means there is room for improvement with the L1-norm method. The improvement with our proposed method can be seen in the difference between the "(a) L1-norm method" and the "(b) proposed method": when we take into consideration the left and center talkers, this is about four points, and when we take into account the right talker, it is about two points.

C. Discussion

These two evaluations demonstrated that our proposed method improves both signal-to-noise ratio and ASR correctness. However, the signal-to-noise ratio improvements with the left and right talkers were negligible. The reason is that our method improves the separation results only for the time-frequency region that have a harmonic structure, while signal-to-noise ratio is calculated from all time-frequency regions with equal weights.

In contrast to signal-to-noise ratio, the improvements to ASR correctness for the left and right talkers were nearly as much as that for the center talker, and these improvements are meaningful. We conclude that our approach, which puts emphasis on the harmonic structure, is a good way to separate mixed speech signals with less distortion of acoustic features.

V. CONCLUSIONS AND FUTURE WORK

We proposed a speech separation method that can be used to achieve a simultaneous speech-recognition system. Since the L1-norm method involves the problems where there are some frequency bands whose estimates of the dominant source are inaccurate, the separation results from the L1-norm method contain spectrum lacking and leakage noise. Such interference greatly distorts acoustic features; thus, the ASR results for these sounds are not good.

We focus on the fact that the harmonic structure has a high-power overtone structure, and we improve the estimation of the dominant source using the harmonic structure.

More concretely, first, we use the L1-norm method and obtain tentatively separated sounds. Second, we extract the harmonic structure from these sounds. Finally, we separate the speech mixtures again, with the constraint that the harmonic structure is always powerful. The experiment revealed that our proposed method improved the correctness of ASR by about four points compared to the baseline method.

In future work, we intend to add new constraints for voiceless consonants, which do not have a harmonic structure. Since voiceless consonants have high power in the high-frequency region, we can expect that new constraints will improve the high-frequency region, which our proposed method cannot do. Another area we need to tackle is the reverberation environment. Even though our experiments were carried out using impulse responses in an anechoic chamber, developing a method that works properly in a standard reverberation room is essential for robots that work in real environments.

In addition, we need to take into consideration the speech-recognition module. We expect that a recognition system that can place more stress on the harmonic structure can improve the ASR results since our proposed method could separate the harmonic structure very well. Another way is to use missing feature theory [15], which would enable us to reduce the effect of interference or enable us to recover feature values that are distorted.

ACKNOWLEDGMENTS: Part of this study was supported by a Grant-in-Aid for Scientific Research (S) and the Global COE Program.

REFERENCES

[1] M. Hirose and K. Ogawa. Honda humanoid robots development. *Philosophical Transactions A*, Vol. 365, No. 1850, p. 11, 2007.

[2] K. Akachi, G. Miyamori, T. Kawasaki, M. Ishizaki, T. Kimura, T. Isozumi, K. Kaneko, F. Kanehiro, H. Hirukawa, and H. Hasunuma. The development of a humanoid robot: HRP-3. *Nippon Robotto Gakkai Gakujutsu Koenkai Yokoshu (CD-ROM)*, Vol. 24, p. 3H15, 2006.

[3] K.F. MacDorman and H. Ishiguro. The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, Vol. 7, No. 3, pp. 297–337, 2006.

[4] A. Hyvärinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, Vol. 13, No. 4-5, pp. 411–430, 2000.

[5] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, Vol. 30, No. 1, pp. 27–34, 1982.

[6] O. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, Vol. 52, No. 7, pp. 1830–1847, 2004.

[7] K. Nakadai, H.G. Okuno, and H. Kitano. Auditory fovea based speech separation and its application to dialog system. In *IEEE/RSJ International Conference on Intelligent Robots and System, 2002*, Vol. 2, 2002.

[8] T.W. Lee, M.S. Lewicki, M. Girolami, T.J. Sejnowski, et al. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, Vol. 6, No. 4, pp. 87–90, 1999.

[9] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal processing*, Vol. 81, No. 11, pp. 2353–2362, 2001.

[10] W. Zhang, J. Liu, J. Sun, and S. Bai. A New Two-Stage Approach to Underdetermined Blind Source Separation using Sparse Representation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, Vol. 3, 2007.

[11] Y. Li, A. Cichocki, and S. Amari. Analysis of sparse representation and blind source separation. *Neural Computation*, Vol. 16, No. 6, pp. 1193–1234, 2004.

[12] Y. Li, S. Amari, A. Cichocki, D.W.C. Ho, and S. Xie. Underdetermined blind source separation based on sparse representation. *IEEE Transactions on Signal Processing*, Vol. 54, No. 2, pp. 423–437, 2006.

[13] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366, 1980.

[14] S. Winter, H. Sawada, and S. Makino. On real and complex valued L1-norm minimization for overcomplete blind source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 86–89, 2005.

[15] B. Raj and R.M. Stern. Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, Vol. 22, No. 5, pp. 101–116, 2005.