Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 547–563
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Detection of speech and music based on spectral tracking

Toru Taniguchi [a,*], Mikio Tohyama [b], Katsuhiko Shirai [a]

[a] Department of Computer Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
[b] Global Information and Telecommunication Institute, Waseda University, 1011 Okuboyama, Nishi-Tomida, Honjo-shi Saitama 367-0035, Japan

Received 29 July 2006; received in revised form 12 December 2007; accepted 9 March 2008

Abstract

How to deal with sounds that include spectrally and temporally complex signals such as speech and music remains a problem in real-world audio information processing. We have devised (1) a classification method based on sinusoidal trajectories for speech and music and (2) a detection method based on (1) for speech with background music. Sinusoidal trajectories represent the temporal characteristics of each category of sounds such as speech, singing voice and musical instrument. From the trajectories, 20 temporal features are extracted and used to classify sound segments into the categories by using statistical classifiers. The average F_1 measure of the classification of nonmixed sounds was 0.939, which might be sufficiently high to apply to subsequent detection of sound categories in a mixed sound. To handle the temporal overlapping of sounds, we also developed an optimal spectral tracking algorithm with low computational complexity; it is based on dynamic programming (DP) with iterative improvement for the sinusoidal decomposition of signals. The classification and detection of a temporal mixture of speech and music are performed by a statistical integration of the temporal features of their trajectories and the optimization of the combination of their categories. The detection method was experimentally evaluated using 400 samples of mixed sounds, and the average of the narrow-band correlation coefficients and improvement in the segmental signal-to-noise ratio (SNR) were 0.55 and +5.67 dB, respectively, which show effectiveness of the proposed detection method.

Keywords: Speech detection; Speech and music discrimination; Sinusoidal model; Spectral tracking; Sinusoidal trajectory

1. Introduction

1.1. Background

In speech information processing in real-world environments, how to handle acoustic mixtures including speech with nonstationary sound such as background music or competing speech remains a problem, even though enormous progress has been made in the processing of speech including stationary noise. Speech and music are types of sound that have complex spectral structures and time-variant streams composed of many musical notes or phonemes; this makes it difficult to estimate background music signals separated from foreground speech.

For interference-signal estimation of a single-channel sound, the temporal or spectral structures of the interferences are the only available prior information besides the signal itself. Therefore, dealing with nonstationary mixtures such as music or speech in a single-channel sound essentially requires models representing the temporal and spectral structures of the target or the nontarget sounds, while, in the case of using multiple microphones, spatial information regarding the sound sources is useful for interference estimation such as blind source separation (Torkkola, 1999). In contrast to multichannel cases, we focus on the detection problem of speech competing against music in single-channel sounds since there are many audio contents provided in a one or two channel(s) such as TV or podcast in the net.

There have been several studies on sound source analysis based on the sound structures of single-channel sounds. Nawab et al. (1998) proposed an analysis method for speech

* Corresponding author. Tel.: +81 3 5286 3118; fax: +81 3 3200 1399.
E-mail addresses: ttani@ieee.org (T. Taniguchi), m_tohyama@waseda.jp (M. Tohyama), shirai@shirai.cs.waseda.ac.jp (K. Shirai).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.03.007

mixed with environmental sounds based on the sound knowledge derived from ideas of auditory scene analysis (Bregman, 1990). They described sounds as hierarchical structures including sound spectra, peaks, and contours, which are units composing the sounds largely based on the perception of sounds. Melih and Gonzalez (2000), Virtanen and Klapuri (2000), and Virtanen (2003) proposed sound separation methods using sinusoidal modeling. They represented harmonic sounds using sinusoidal trajectories, which are temporal sequences composed of spectral peaks connected on the basis of continuity of frequencies and amplitudes; these trajectories are then clustered into separate sound sources on the basis of the closeness of trajectories defined based on perceptual association cues.

Sound source classification is the other problem in speech detection since it is essential to identify the sound categories of the separated sound units to detect speech in addition to the sound source analysis described above. Several studies have been conducted on speech and music discrimination (SMD). Features of audio signals have been characterized in the frequency-domain by zero-crossing rates (Saunders, 1996), spectral centroid (Scheirer and Slaney, 1997), harmonic coefficients (Chou and Gu, 2001), and mel-frequency cepstral coefficients (MFCC) (Xiong et al., 2003; Kim et al., 2004). These features correspond to the spectral envelope or harmonics for a time segment of sound within 100 ms. In addition, spectral flux (Scheirer and Slaney, 1997), cepstrum flux (Takeuchi et al., 2001), and 4-Hz modulation energy (Scheirer and Slaney, 1997) have also been investigated as combinations of the frequency and time domain characteristics. These features represent relative long-term (about 1 s) fluctuations of spectral structures such as a short-time spectrum or cepstrum. However, these features are useful only for nonoverlapping sounds.

In contrast with these studies for nonoverlapping sounds, Taniguchi et al. (2005) performed sinusoidal-trajectory decomposition so as to classify mixed sounds composed of speech and music. This decomposition is similar to that in the separation methods (Virtanen, 2003) but considers long-term temporal characteristics to classify trajectories into sound categories. An individual trajectory is characterized using a temporal feature vector extracted from the trajectory, and category classification of each time segment is conducted by a statistical integration of the temporal feature vectors of the trajectories composing the time segment. While Melih and Gonzalez (1999) classified sounds by rules based on four predefined temporal patterns of sinusoidal trajectories, Taniguchi et al. (2005) classified sounds by using the statistical models of the temporal features of sinusoidal trajectories; the models would comprehend the rules of the method by Melih and Gonzalez (1999). In addition, this method can be easily extended because it can be combined with other statistical signal processing methods.

In this paper, we describe a classification method in detail first proposed by Taniguchi et al. (2005) and extend it to a detection method of speech competing with music. These classification and detection methods can be applied to various kinds of speech information processing, e.g., content-based audio indexing.

1.2. Classification and detection framework

Fig. 1 illustrates the procedures for category classification and detection of audio signals. First, the spectral analysis is applied to the input audio signal to select spectral peaks from each analysis frame. Next, spectral tracking is performed by connecting the selected peaks between adjacent analysis frames. The connecting operation produces sinusoidal trajectories, each of which is a series of jointed spectral peaks corresponding to the temporal trajectories of the fundamental frequency (F_0) or its harmonic for voices or music instruments.

The temporal features that represent the shape of a sinusoidal trajectory are then extracted. They represent the temporal changes in an audio signal, particularly the temporal changes in F_0s or harmonics. Finally, the sinusoidal trajectories are statistically integrated into time segments and classified into audio categories such as musical instrument, singing voice, or speech; mixed-categories such as instruments and singing are also considered. Concurrently with the classification of time segments, each individual sinusoidal trajectory can be classified into an individual category. The detection of a mixture is based on the classification of each sinusoidal trajectory.

Tracking individual sinusoidal trajectories and extracting temporal features from them are the key techniques of our method. We describe a method for the spectral tracking of individual sinusoidal trajectories (Taniguchi et al., 2006). We introduce optimization processes into conventional frame-wise spectral tracking (McAulay and Quatieri, 1986; Depalle et al., 1993; Marks and Gonzalez, 2005) to deal with the difficulty of tracking trajectories in mixed sounds and to reduce the computational complexity. The proposed temporal features for the classification will be evaluated through statistical analysis and classification experiments on sinusoidal trajectories.

The remainder of this paper is organized as follows. We describe the spectral peak analysis in Section 2 and the spectral tracking method for extracting sinusoidal trajectories in Section 3. In Section 4, we define the temporal features, which we have statistically analyzed with regard to their variance ratios. We also describe classification experiments using trajectory samples to evaluate the temporal features. We describe the sound classification and detection processes in Section 5 and evaluate these methods in Section 6. We summarize our findings in Section 7.

2. Spectral analysis method

The spectral analysis is performed to pick out candidate peaks from audio signals, which will be used for the subsequent spectral tracking step. In this study, we use two methods: one is the usual local peak picking method using short-time Fourier transform (STFT); the second is

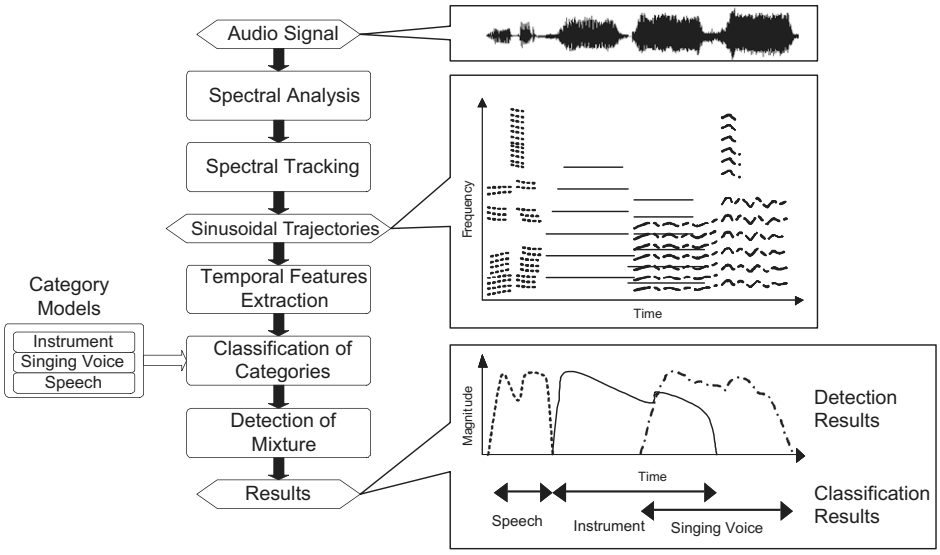


Fig. 1. Overview of the classification and detection system.

clustered line-spectrum modeling (CLSM) (Kazama et al., 2003), which is effective when spectral peaks are located very close together. Other spectral analysis methods may be used instead of these two methods.

The analysis method should be selected considering the trade-off between the computational complexity and the strictness of each analysis method because, generally, more strict analyses are computationally more expensive. However, the strictness of analysis is important since a sinusoidal trajectory that includes a candidate peak cannot be tracked if the peak is not obtained in the spectral analysis step. The trade-off problem will be discussed in Section 3 using an example of a tracked synthesized sound.

2.1. Local peak picking

One analysis method (Taniguchi et al., 2005) is the peak picking method, where local peaks are picked out from a power spectrum of a short-time analysis frame. In this study, a power spectrum is obtained using the STFT with a sampling rate of 16 kHz, Hamming windows with widths of 32 ms (512 points), and frame rates of 16 ms (256 points). To attain fine spectral interpolation, we apply the STFT to the signal of 4096 samples by adding zero samples. The spectral peak is determined as a local peak of the interpolated spectrum. These peaks are candidates for sinusoidal trajectories. Other types of spectral analysis such as instantaneous frequency (IF) amplitude spectrum analysis (Abe et al., 1996; Abe and Honda, 2006) can be used instead of the STFT.

2.2. Clustered line-spectrum modeling for short signal intervals

Local spectral peak picking using STFT is a good approach with regard to the computational and implementation complexity. However, if several sinusoidal components are located closely together, this method is not suitable because it is difficult to distinguish the components in a narrow frequency interval along with their envelopes, which are yielded by the STFT analysis window.

The principle of clustered line-spectrum modeling (CLSM) was developed by Kazama et al. (2003) to estimate the true sinusoidal components, eliminating the effects of analysis windows, from clustered spectral records. If a target signal is composed of a finite number of clustered sinusoidal components, the components can be estimated by obtaining the solution based on least-square-error (LSE) using the overdetermined simultaneous equations in the frequency domain instead of the time region where conventional sinusoidal modeling is performed.

Suppose that a signal $x(n)$ has the record length of N and the interpolated spectrum $X(k)$ is analyzed by taking the M-point fast Fourier transform (FFT) after zero padding so that

$$X(k) = \frac{1}{M}\sum_{n=0}^{M-1} x(n)w(n)\mathrm{e}^{j2\pi kn/M}, \tag{1}$$

where $w(n)$ denotes an N-point analysis window. We assume that the target signal can be expressed in an analytic form $x_a(n)$ as

$$x_a(n) \equiv \sum_{i=1}^{I} A_i \mathrm{e}^{j2\pi f_i n} + \epsilon(n), \tag{2}$$

where I is the number of dominant sinusoidal components that are clustered around the FFT components at $k = k_p$; A_i and f_i denote the ith dominant component's complex magnitude and frequency, respectively; and $\epsilon(n)$ denotes the residual component including the modeling error and external noise, which we look forward to minimizing.

If we attempt to represent the signal by clustered P sinusoidal components between $k = k_p - m$ and $k = k_p - m + P - 1$,

the P parameter sets can be estimated based on the LSE criterion by using a set of linear equations for L observation frequency points between $k = k_p - l$ and $k = k_p - l + L - 1$:

$$\mathbf{x}_0 = W\mathbf{x}_s, \tag{3}$$

$$\hat{\mathbf{x}}_s = (W^{\mathrm{T}}W)^{-1}W^{\mathrm{T}}\mathbf{x}_0. \tag{4}$$

Here, $\mathbf{x}_0$ denotes the observed subband spectrum $(X(k_p - l), \ldots, X(k_p - l + L - 1))^{\mathrm{T}}$, $\mathbf{x}_s$ denotes the target subband spectrum $(X(k_p - m), \cdots, X(k_p - m + P - 1))^{\mathrm{T}}$ to be estimated, $\hat{\mathbf{x}}_s$ denotes the estimated target spectrum, and W is the matrix representing spurious spectra due to the window function $w(n)$. Here we set

$$W \equiv \{w_{i,j}\}, \tag{5}$$

$$w_{i,j} \equiv W_{NM}(-l + i - m + j - 2), \tag{6}$$

$$W_{NM}(q) \equiv \frac{1}{N}\sum_{n=0}^{N-1} w(n)\mathrm{e}^{-j\frac{2\pi kn}{M}}|_{k=q} \tag{7}$$

and $L > P$, $l > m$,

$$m \equiv \begin{cases} \frac{P-1}{2} & P : odd \\ \frac{P}{2} & P : even \end{cases}; \quad l \equiv \begin{cases} \frac{L-1}{2} & L : odd \\ \frac{L}{2} & L : even. \end{cases} \tag{8}$$

3. Spectral tracking method

The spectral tracking problem is how to connect temporally adjacent spectral peaks to organize sound objects as units of a sound representation. In this paper, spectral tracking is defined as a sound information processing that connects temporally adjacent spectral peaks so as to minimize the summation of the distances between the peaks in an interval of the target audio signal, in other words, so as to maximize the "smoothness" of trajectories.

From a viewpoint of auditory scene analysis (Bregman, 1990), this minimal-distance tracking is a reasonable form of information processing inspired by the sound segregation of human beings. One of the general principles of perceptual organization (Moore, 2004) strongly supports this tracking basis of minimal-distance; the principle is "good continuation," which means that the changes in frequency, intensity, etc. tend to be smooth and continuous.

The tracking problem can be formulated as an optimal combination problem of spectral peaks, as described in the remainder of this section, and its optimal solution can be computed using an algorithm of order $\mathrm{O}(M!)$ in the case of a naive approach, which implies that over 10^{157} calculations of the sum of the distances are required when the number of peaks is $M = 100$; this algorithm is not computable in practice. Hence, conventional methods have partly adopted heuristic approaches, which are not mathematically optimal, to reduce the complexity. The complexity and the optimality are important factors for the tracking algorithm. Further, the definition of the distance between spectral peaks is the other problem in the tracking, which can be discussed independently of the algorithm itself.

McAulay and Quatieri (1986) proposed a heuristic tracking algorithm based on the difference of the frequencies of spectral peaks. Depalle et al. (1993) proposed a method that minimizes the summation of the distances of spectral peaks in several analysis frames that are defined using both the frequency and the amplitude of the spectral peaks. They used left-to-right hidden Markov models (HMM)[1] with decoder by employing the Viterbi algorithm as the minimizing algorithm, which mathematically ensures the minimal of the sum of the distances and is fast. This algorithm has the order of $\mathrm{O}(M!N)$ complexity where M and N denote the number of peaks and analysis frames, respectively, which is rather high since it does not solve the complexity problem caused by the huge number of peak combinations between two adjacent frames, while it dose solve the problem for three and more frames. Therefore, some heuristic assumptions were introduced to it with the purpose of reducing the complexity, such as not changing the number of trajectories during tracking. Marks and Gonzalez (2005) also proposed a tracking algorithm using the distance defined by the differences in the Δ frequencies and the Δ amplitudes, which makes it easier to extract crossed trajectories and has lower order of $O(M^3)$ complexity than that of the algorithm proposed by Depalle et al. (1993), when connecting peaks between adjacent frames. However, their algorithm does not mathematically ensure the minimal of the sum of the distances.

In order to overcome these computational complexity and optimality problems, we propose a locally optimal tracking method composed of a dynamic programming (DP) step and an iterative improvement step; this method involves lower computational complexity of order $\mathrm{O}(M^2)$ than the conventional optimal methods described above. Consequently, the proposed method makes it possible to extract sinusoidal trajectories rapidly, even if two trajectories cross each other in the frequency domain.

3.1. Formulation of spectral tracking

Spectral tracking requires that we determine an optimal combination of spectral peaks between two adjacent analysis frames, and the number of the candidate combinations is extremely large, as mentioned above. Sakakibara and Osaka (1998) reported a solution to this problem where the DP algorithm was originally used for the purpose of sound morphing, and it vastly reduced the computational complexity of the combination problem with a limitation that the connected trajectories did not cross on the time–frequency plane. We adopt this solution for our tracking problem. To deal with the limitation of the DP method, we propose an additional optimization step through iterative improvement.

[1] Using HMM in this method does not imply that it is a statistical method like the acoustic modeling of speech recognition.

Let $X^n = \{\mathbf{x}_i^n | i = 1, \ldots, |X^n|\}$ denote a set of peak vectors corresponding to an analysis frame including spectral peaks $\mathbf{x}_i^n$ at time n. A peak vector $\mathbf{x}_i^n$ is defined in the frequency and the amplitude domain. The index integer i of a peak vector is numbered in the frequency order. $|X^n|$ is the number of vectors of X^n. The combination problem is to determine the mapping γ from X^{n-1} to $X^n \cup null$ so that the cumulative cost function $C_{X^{n-1},X^n}(\gamma)$ between two successive frames, X^{n-1} and X^n, is minimized. The cumulative cost function that denotes the summation of the distances between two adjacent frames is defined as

$$C_{X^{n-1},X^n}(\gamma) = \sum_{i=1}^{|X^{n-1}|} d(\mathbf{x}_i^{n-1}, \mathbf{x}_{\gamma(i)}^n), \tag{9}$$

where $d(\mathbf{x}_i^{n-1}, \mathbf{x}_j^n)$ is the distance between peaks $\mathbf{x}_i^{n-1}$ and $\mathbf{x}_j^n$. This distance can be defined using the frequencies and amplitudes of the peaks $\mathbf{x}_i^{n-1}$ and $\mathbf{x}_j^n$. In this paper, we define it using the second derivative of the frequencies and the amplitudes in addition to the first derivatives of the frequencies and the amplitudes, in order to allow the trajectories to cross and so that it is less susceptible to the effects of noise peaks (Depalle et al., 1993; Marks and Gonzalez, 2005). The distance $d(\mathbf{x}_i^{n-1}, \mathbf{x}_j^n)$ is defined as

$$d(\mathbf{x}_i^{n-1}, \mathbf{x}_j^n)^2 = \left(\frac{\Delta f^n(i,j)}{C_{\Delta f}}\right)^2 + \left(\frac{\Delta a^n(i,j)}{C_{\Delta a}}\right)^2 + \left(\frac{\Delta^2 f^n(k,i,j)}{C_{\Delta^2 f}}\right)^2 + \left(\frac{\Delta^2 a^n(k,i,j)}{C_{\Delta^2 a}}\right)^2, \tag{10}$$

where $C_{\Delta f}$, $C_{\Delta a}$, $C_{\Delta^2 f}$, and $C_{\Delta^2 a}$ are experimentally determined constants, and k is the index number of the peak $\mathbf{x}_k^{n-2}$ connected to the peak $\mathbf{x}_j^{n-1}$ at time $n-1$. The first derivatives of frequency $\Delta f^n(i,j)$ and amplitude $\Delta a^n(i,j)$ are defined as

$$\Delta f^n(i,j) = f(\mathbf{x}_j^n) - f(\mathbf{x}_i^{n-1}), \tag{11}$$

$$\Delta a^n(i,j) = a(\mathbf{x}_j^n) - a(\mathbf{x}_i^{n-1}) \tag{12}$$

and the second derivatives of frequency $\Delta^2 f^n(k,i,j)$ and amplitude $\Delta^2 a^n(k,i,j)$ are defined as

$$\Delta^2 f^n(k,i,j) = \Delta f^n(i,j) - \Delta f^{n-1}(k,i), \tag{13}$$

$$\Delta^2 a^n(k,i,j) = \Delta a^n(i,j) - \Delta a^{n-1}(k,i), \tag{14}$$

where $f(\mathbf{x})$ is the frequency of peak x; $a(\mathbf{x})$, the amplitude. In our experiments, the frequency and amplitude were both log scaled. The distance $d(\mathbf{x_i}, null) = d_{null}$ when peak $\mathbf{x_i}$ does not connect to any peak is also defined as a constant, which works as a threshold such that if the distance $d(\mathbf{x}, \mathbf{y})$ is greater than d_{null}, peaks $\mathbf{x}$ and $\mathbf{y}$ are never connected. In the case of $k = null$ in Eqs. (13) and (14), the values of $\Delta^2 f^n(null,i,j)$ and $\Delta^2 a^n(null,i,j)$ are calculated as

$$\Delta^2 f^n(null,i,j) = \Delta f^n(i,j) \tag{15}$$

$$\Delta^2 a^n(null,i,j) = \Delta a^n(i,j), \tag{16}$$

respectively.

3.2. Optimal matching step by dynamic programming

If we assume that mapping γ has a strictly monotonically increasing limitation, $i_1 < i_2 \Rightarrow \gamma(i_1) < \gamma(i_2)$, we can find the mapping γ that minimizes the cumulative cost function Eq. (9) by calculating the DP matrix $C[i][j]$ $(i = 1, \ldots, |X^{n-1}|; j = 1, \ldots, |X^n| + 1)$ as follows:

$$\min_\gamma C_{X^{n-1},X^n}(\gamma) = \min_i C[i][|X^n| + 1] \quad (i = 1, \ldots, |X^{n-1}|). \tag{17}$$

In addition, the matrix $C[i][j]$ can be calculated using the following equations:

if $i = 1$ **then**

$$C[1][j] = \begin{cases} \min(d(\mathbf{x}_1^{n-1}, \mathbf{x}_1^n), d_{null}) & (\text{if } j = 1) \\ \min(d(\mathbf{x}_1^{n-1}, \mathbf{x}_1^n),\ C[1][j-1]) & \\ \quad (\text{if } 1 < j \leqslant |X^n|) & \\ \min(d_{null},\ C[1][|X^n|]) + (|X^{n-1}| - i) \cdot d_{null} & \\ \quad (\text{if } j = |X^n| + 1), & \end{cases} \tag{18}$$

else if $1 < i \leqslant |X^{n-1}|$ **then**

$$C[i][j] = \begin{cases} (i-1) \cdot d_{null} + d(\mathbf{x}_i^{n-1}, \mathbf{x}_1^n) & (\text{if } j = 1) \\ \min(C[i][j-1],\ d(\mathbf{x}_i^{n-1}, \mathbf{x}_j^n) + c_{i,j}) & \\ \quad (\text{if } 1 < j \leqslant |X^n|) & \\ \min(C[i][|X^n|], d_{null} + C[i-1][|X^n|]) & \\ +(|X^{n-1}| - i) \cdot d_{null} & \\ \quad (\text{if } j = |X^n| + 1), & \end{cases} \tag{19}$$

where $c_{i,j}$ is defined as

$$c_{i,j} = \min_{1 \leqslant k < i} (C[k][j-1] + (i - k - 1) \cdot d_{null}), \tag{20}$$

end if.

The range of index j includes $|X^n| + 1$ to account for the case in which there are peaks $\mathbf{x}_i^{n-1}$ that do not connect to any peak in X^n.

3.3. Optimal matching step by iterative improvement

The optimal result described in Section 3.2 has the limitation that two trajectories are not allowed to cross each other in the frequency domain. However, the combination of peaks, so as to allow trajectories to cross, can be gradually improved by changing the combination one by one from the DP result and selecting the change that reduces the cost function (Eq. (9)). The result of the following procedures must be optimal or at least locally optimal.

The initial condition is $C = C_{X^{n-1},X^n}(\gamma)$, which is calculated in Section 3.2. All the other values and functions follow from the result of Section 3.2. The following procedures are repeated while C is changing:

for all $i = 1, \ldots, |X^{n-1}|$ **do**

if $\gamma(i) = null$ **then**

- Find the least iu satisfying $f(\mathbf{x}_i^{n-1}) \neq f(\mathbf{x}_{iu}^{n-1}) \wedge i < iu$, and the greatest id satisfying $f(\mathbf{x}_i^{n-1}) \neq f(\mathbf{x}_{id}^{n-1}) \wedge i > id$.
- Calculate the candidate distance $\hat{d} = \min_{i' \in \{iu,id\}}\{d(i, \gamma(i'))\}$, and select the candidate index $\hat{j} = \gamma[\arg\min_{i' \in \{iu,id\}}\{d(i, \gamma(i'))\}]$.

else

- When $j = \gamma(i)$, find the least ju satisfying $f(\mathbf{x}_{\gamma^{-1}(ju)}^{n-1}) \neq f(\mathbf{x}_i^{n-1}) \wedge j < ju$, and the greatest jd satisfying $f(\mathbf{x}_{\gamma^{-1}(jd)}^{n-1}) \neq f(\mathbf{x}_i^{n-1}) \wedge j > jd$, where γ^{-1} is the inverse mapping of γ.
- Calculate the candidate distance $\hat{d} = \min_{j' \in \{ju,jd\}}\{d(i, j')\}$, and select the candidate index $\hat{j} = \arg\min_{j' \in \{ju,jd\}}\{d(i, j')\}$.

end if

- Calculate the candidate cost $\widehat{C} = C - d(i, j) + d(i, \hat{j})$.
- If $C > \widehat{C}$, renew C by $\widehat{C}$ and $\gamma(i)$ by $\hat{j}$.

end for.

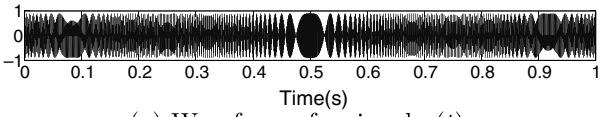
3.4. Spectral tracking experiments

First, we evaluated the proposed tracking method by tracking a synthesized sound $y(t)$, shown in Fig. 2a, which includes crossing spectral partials on the time–frequency plane. Fig. 2b–e shows the tracking results with white lines on the spectrogram of the sound. As shown in Fig. 2e, the crossing sinusoidal trajectories are correctly extracted when iterative improvement tracking was performed using CLSM, but not when STFT was used, even with the iterative improvement tracking. This is because the STFT local peak picking method cannot find two peaks close together, as in the 12th frame in Fig. 2b and c.

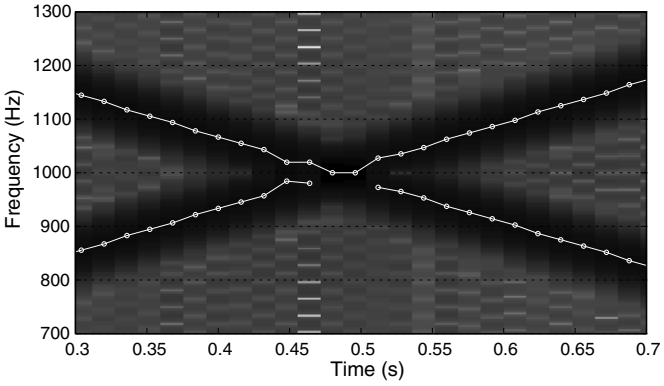
The experimental conditions for spectral tracing were $C_{\Delta f} = C_{\Delta^2 f} = 300.0$ (cent/frame), $C_{\Delta a} = C_{\Delta^2 a} = 20.0$ (dB/frame), the length of the analysis frame interval is 16 ms, and $M = 100$. Here, the cent frequency f_{cent} is defined as

$$f_{\text{cent}} = 1200 \cdot \log_2 \frac{f_{\text{Hz}}}{440 \cdot 2^{\frac{3}{12}-5}} \tag{21}$$

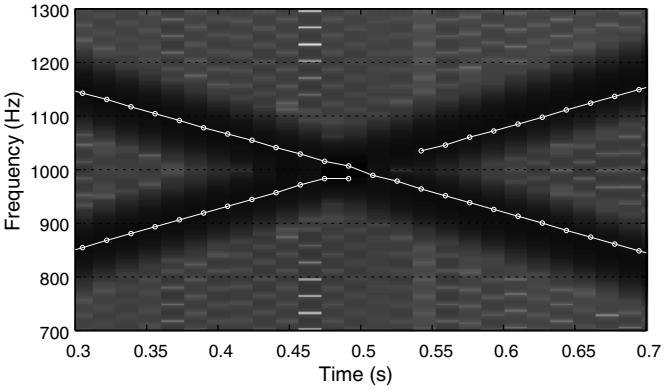
where f_{Hz} is the frequency in Hertz. These parameters are determined based on the results of preliminary investigations using the Keele pitch database (Plante et al., 1995),



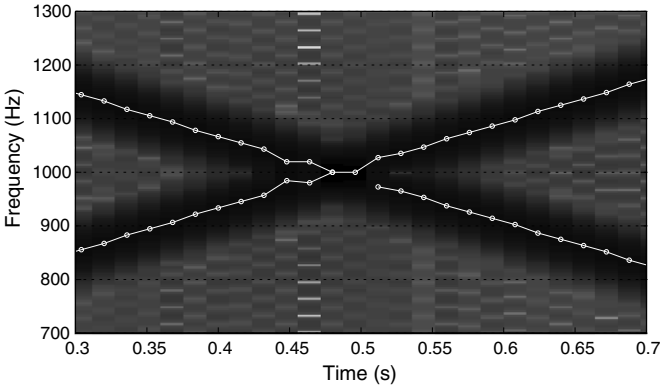
(a) Waveform of a signal $y(t)$



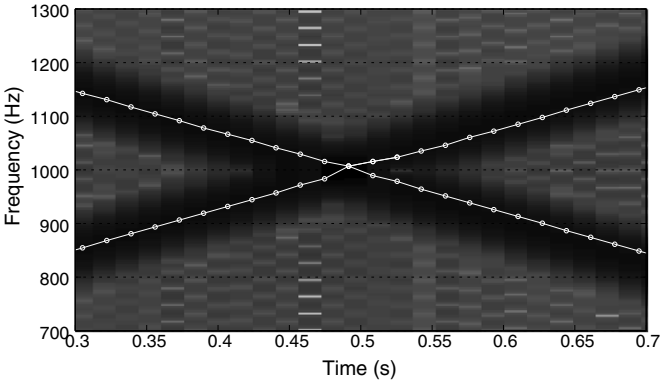
(b) DP tracking with STFT



(d) DP tracking with CLSM



(c) DP + iterative improvement tracking with STFT



(e) DP + iterative improvement tracking with CLSM

Fig. 2. Examples of spectral tracking for the signal: $y(t) = \sin[2\pi\{1000t + 400(t^2 - t)\}] + \sin[2\pi\{1000t - 400(t^2 - t)\}]$.

which includes the F_0 trajectories of speech signals examined by hand, so that sinusoidal trajectories can trace the F_0 trajectories of normal speech.

Second, Figs. 3–5 show examples of realistic tracking results of a speech signal, a singing-voice signal, and a musical–instrument signal, respectively. The parameters of the tracking algorithm are the same as in the previous experiments. The extracted F_0 trajectories and spectrograms of these signals are also shown simultaneously with them. As can be seen in these figures, the trajectories are extracted to correspond to the F_0 trajectories and its harmonics; the extracted trajectories may represent the characteristics of the sound of each category: e.g., the changes in the frequencies of trajectories are roughly greater in the speech and singing-voice signals than in the musical instrument signal, and most trajectories durations in the singing-voice signal are longer than those in the speech signal. These characteristics are utilized for the temporal features described in Section 4.

Then, Fig. 6 shows the computing time of each tracking methods when mixed sound of speech and musical instruments were tracked. Length of the sound is 3.5 s. These results are computed on a GNU/Linux OS using a Xeon™ 3.0 GHz CPU and 4 GB memory. As can be seen, our proposed methods (DP method and DP + iterative improvement (II) method) are faster than the Marks and Gonzalez (2005) (MG) method. In the case $M = 100$, which is the same as the tracking conditions of above experiments, the computing speed of our methods is about two times faster than that of the MG method. These computing results agree with the complexity order of the algorithms, as mentioned in the beginning of this section.

The effectiveness of the tracking cannot be investigated using only the tracking results themselves since it depends on the effectiveness in the application. Therefore, the evaluations of the tracking methods will be conducted using the results of speech detection, which are illustrated in Section 6.

4. Temporal features from sinusoidal trajectories

In this section, first, the extraction of the proposed temporal features from sinusoidal trajectories is illustrated. Second, the statistical characteristics of these features are described. Then, a category classification experiment is performed in order to evaluate these temporal features.

4.1. Extraction of temporal features

In this step, 20 temporal features listed below are extracted from each sinusoidal trajectory, which were determined based on the signal signatures extracted with spectral tracking. These signatures are found in experiments of realistic signal tracking, as in Section 3.4. These features represent the temporal patterns of the log magnitude and log frequency of the trajectories. The proposed features are independent of frequency, because the frequency is log-scaled so that two trajectories derived from the same harmonic structure will have similar temporal frequency patterns. In this paper, the frequency in cent defined in Eq. (21) is used as the scale of frequencies. The proposed 20 features are as follows:

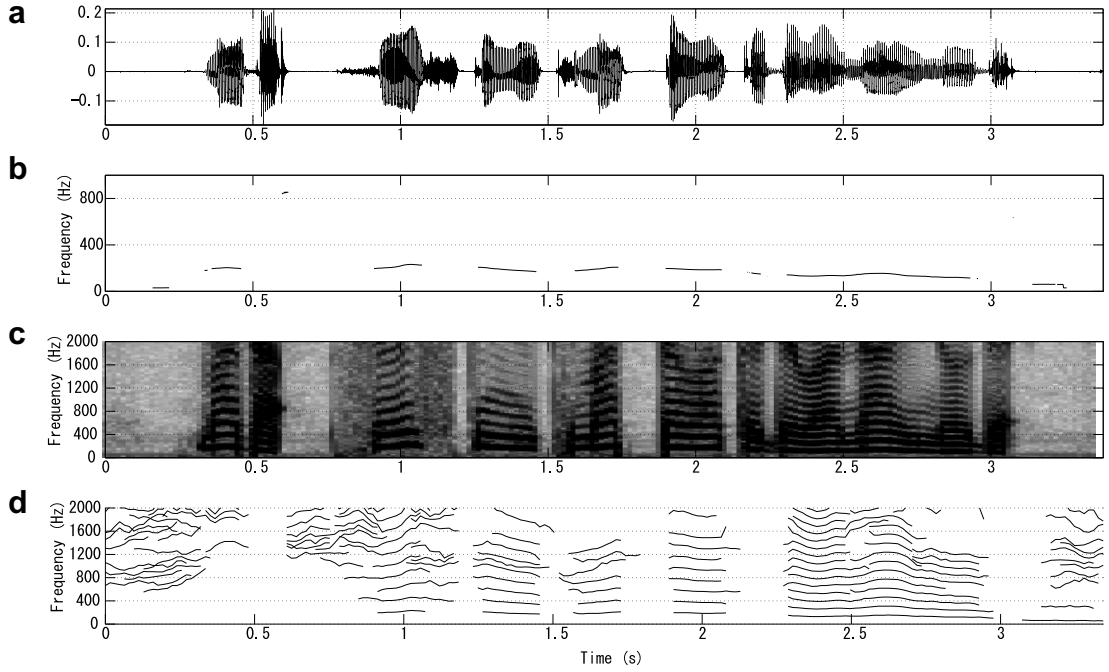


Fig. 3. Various representations of a female-speech signal: (a) time-domain waveform of a speech signal, (b) extracted F_0 trajectories from (a), (c) a spectrogram of (a), (d) sinusoidal trajectories extracted from (a).

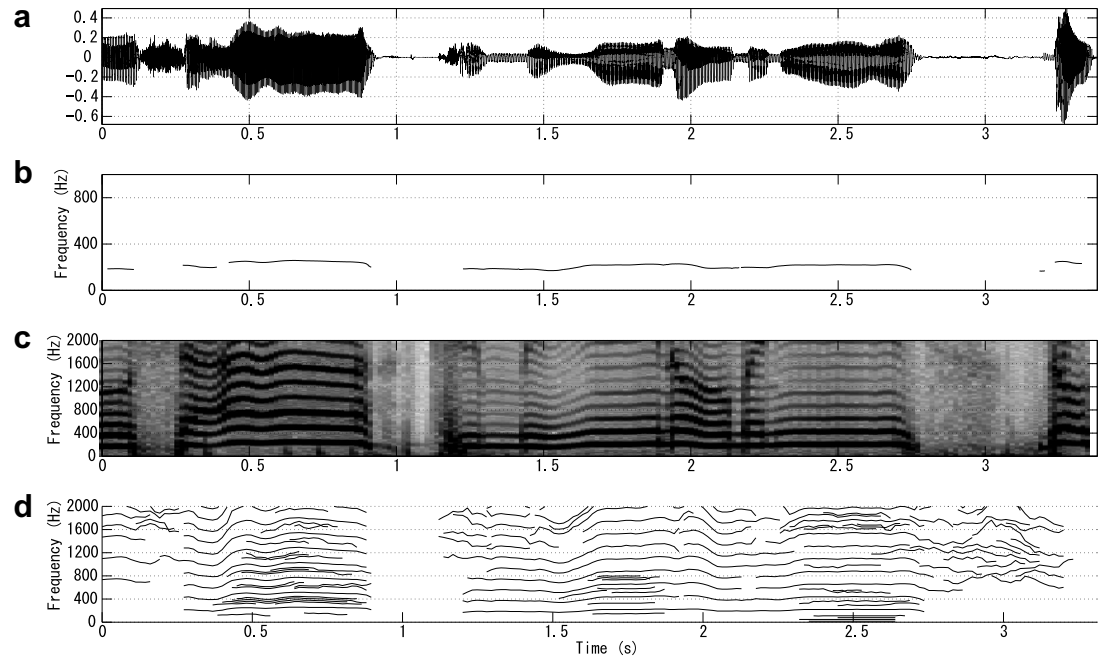


Fig. 4. Various representations of a male singing-voice signal: (a)–(d) are similar to Fig. 3.

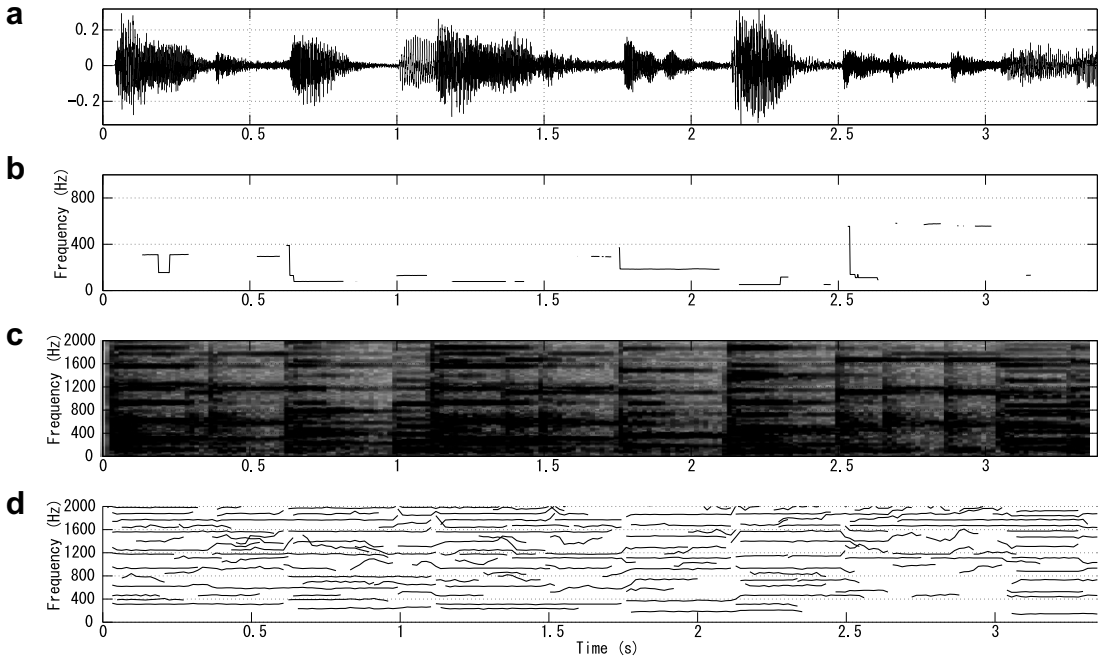


Fig. 5. Various representations of a musical–instrument signal of piano: (a)–(d) are similar to Fig. 3.

- Duration of a sinusoidal trajectory (#1).
- Standard deviation (STD) of frequencies in cent (#2), log powers (#3), and their proportions in an analysis frame (#4).
- The mean (#5) and standard deviation (#6) of differences in frequency in cent between the frequency of a sinusoidal component and the nearest note in an equal temperament: The difference can be calculated as

$$\mathrm{diff} = (f + 50) \pmod{100} - 50, \tag{22}$$

where f is the frequency in cent of the spectral component. (The interval between neighboring notes is always 100 cent in the equal temperament because of the definition of frequency in cent.)

- The mean and the standard deviation of the time derivatives (Δ) and the second derivatives ($\Delta\Delta$) of frequency in cent (#7 (mean of Δ), #8 (STD of Δ), #9 (mean of $\Delta\Delta$), #10 (STD of $\Delta\Delta$)), of log power (#11 (mean of Δ), #12 (STD of Δ), #13 (mean of $\Delta\Delta$), #14

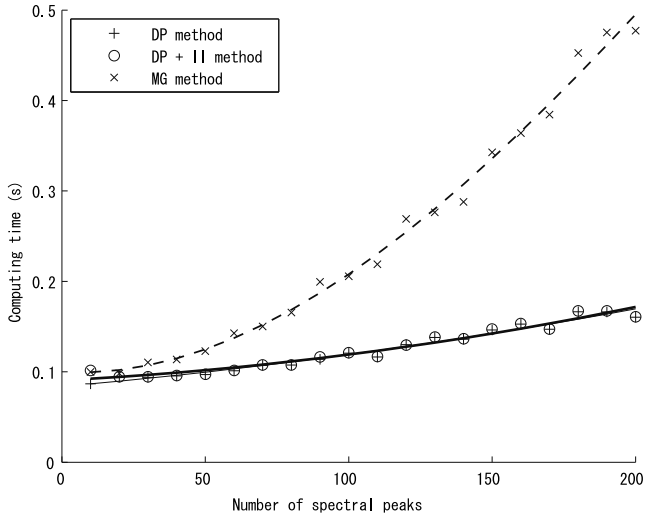


Fig. 6. An example of reducing computing time for spectral tracking. Solid line: second-order polynomial representation for DP method and DP + iterative improvement (II) method, broken line: third-order polynomial representation for Marks and Gonzalez (2005) (MG) method.

(STD of ΔΔ)) and of its proportion in a frame (#15 (mean of Δ), #16 (STD of Δ), #17 (mean of ΔΔ), and #18 (STD of ΔΔ)).

- Symbolic representation of a sinusoidal trajectory in frequency in cent (#19) and in log power (#20): Each mean is calculated for five frames of the head, middle, and tail of a sinusoidal trajectory in frequency and in log power. If the mean of the five frames is more than 10.0 cent higher than that of the former five frames, the symbol "H" is used; if lower than 10.0 cent, "L" is used. Otherwise, "M" is used. For example, "HM," "HL," and "MM" are the representations of trajectories. In the case of log power, 2.0 dB is the threshold value.

4.2. Statistical characteristics of temporal features

In order to inspect the statistical characteristics of our proposed temporal features, we performed statistical analyses of the features on a speech and music corpus consisting of three categories: instrument, singing voice, and speech. The structure of the corpus is shown in Table 1. The samples constituting the corpus were obtained from the databases listed below.

- RWC Music Database of Popular, Classical, and Jazz Music (Goto et al., 2002): This contains music pieces whose quality is as high as that of commercially distributed music. Instrumental pieces were selected from it.
- Corpus of Spontaneous Japanese (Maekawa et al., 2000): This contains the samples of spontaneous speech in Japanese. Speech pieces were selected from it.
- Originally recorded music data: The above databases contain a few singing voice pieces that are not overlapped with instrumental sounds. Therefore, we recorded original music consisting of only singing voices.

Table 1
The structure of the experimental dataset composed of three category sounds

Samples	Number of sinusoidal trajectories	Duration (s)
Instrument	10,570	460
Singing voice	8,785	600
Speech	10,074	230

Fig. 7 shows the results of one-way analysis of variance (ANOVA) (Hogg and Ledolter, 1987) for each feature. A feature having a higher variance ratio will be more effective for classification than those with lower ratios. As shown in Fig. 7, four features—the mean value of Δ power (#11), the mean of ΔΔ power (#13), the mean of Δ relative power (#15), and the mean of ΔΔ relative power (#17)—had a very low variance ratio. On the other hand, the mean of durations (#1), the STD of frequencies (#2), the STD of relative powers (#4), the STD of differences between frequency and that of the nearest musical note (#6), the STD of Δ frequencies (#8), the STD of ΔΔ frequencies (#10), the STD of Δ power (#12), the STD of ΔΔ powers (#14), the STD of Δ relative powers (#16), and the STD of ΔΔ relative powers (#18) all had higher variance ratios. These values imply that the means of changes in power might contribute relatively little to the category classification, whereas the standard deviations of frequency and power and the durations of trajectories strongly characterize the trajectories of each category and might be highly effective for category classification.

Fig. 8 shows the accumulated variance proportion obtained using principal components analysis (PCA) (Jackson, 1991) of the feature vectors. The dataset of sinusoidal trajectories used in the PCA analysis is shown in Table 1. This analysis was performed to evaluate the number of statistically independent dimensions of our proposed feature vector, which might be effective in the classification. Fig. 8 indicates that the number of independent dimensions

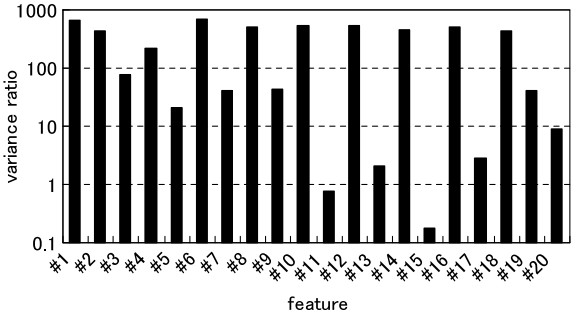


Fig. 7. Variance ratio of each temporal feature obtained using one-way ANOVA.

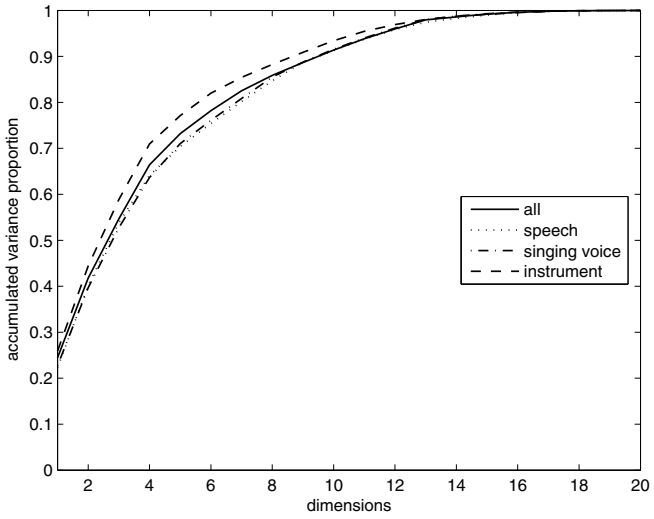


Fig. 8. Accumulated variance proportion obtained using PCA for each category of temporal feature vectors.

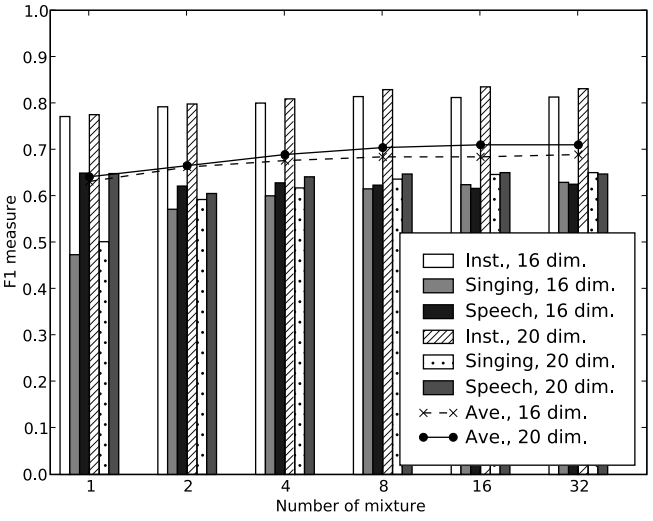


Fig. 9. Classification results of sinusoidal trajectories of three categories indicated by F_1 measure using 16 or 20 features. The results are shown for each number of GMM mixtures from 1 to 32. The solid and broken lines show the averages for F_1 measures in three categories classified using the 20 features and the 16 features, respectively.

was not less than 16 and these 20 feature sets might be useful to classify sound categories.

4.3. Evaluation of temporal features through category classification experiments

We performed experiments on the classification of sinusoidal trajectories into three categories—speech, singing voice, and musical instrument—using the temporal features described in Section 4.1 as feature vectors for classifiers.

The Gaussian mixture model (GMM) classifier was adopted to classify the trajectories. The GMMs were trained with labeled training data using the EM algorithm (Dempster et al., 1977) and the classification was conducted with the maximum likelihood selection of these GMMs. The classification experiments were examined in a 10-fold cross-validation test by varying the number of mixtures of GMMs from 1 to 32 and in each experiment, using the dataset described in Section 4.2. In a cross-validation test, all the labeled samples of each category were randomly divided into 10 sets, and 10 tests were carried out. One set selected in each test was used as the test data, and the other sets were used to train the classifiers. The rates of classification were calculated for each test and aggregated. All the tests were conducted in both the cases of all 20 temporal features and 16 features that excluded the four features of means of changes in power (#11, #13, #15, #17), which showed relatively low variance ratios.

Fig. 9 shows the classification results of the sinusoidal trajectories. It shows the F_1 measures of the GMM classification with 1–32 mixtures in each category and the average for them, comparing the results obtained using 16 features and the 20 features. F_1 measure is defined as

$$F_1 = \frac{2 \cdot (precision \cdot recall)}{precision + recall}, \tag{23}$$

where

$$precision = \frac{|\{\text{correctly classified samples}\}|}{|\{\text{classified samples}\}|}, \tag{24}$$

$$recall = \frac{|\{\text{correctly classified samples}\}|}{|\{\text{relevant samples}\}|}. \tag{25}$$

Comparing the results obtained using the 16 and the 20 features, the F_1 measures of the 20 features were slightly better than those of the 16 features. This shows that using the entire 20 features is not less effective than using the selected 16 features, although the 4 excluded features exhibited low variance ratios. Therefore, we use the entire 20 features in the subsequent experiments of this study.

Increasing the number of mixtures of the GMM to 16 monotonically improved the performance in the cases of 20 features; however, increasing the number from 16 to 32 slightly decreased F_1, possibly because of overestimation. These results suggest that the distribution of the features of sinusoidal trajectories is as complicated as that of a 16-mixture GMM in the case of the samples of this experiment. The best average F_1 was 0.711 when using the 16-mixture GMMs and the 20 features.

5. Classification and detection

In this section, we describe the category classification and speech detection algorithm using the temporal features described in Section 4.

5.1. Classification algorithm for time segments

A time segment of an audio signal is composed of multiple sinusoidal trajectories. We devised a method to classify

the time segment by integrating the sinusoidal trajectories statistically.

The category classification problem is defined as finding a category $C_{\mathrm{mix}} = \widehat{C}_{\mathrm{mix}}$ to maximize the likelihood $p(\mathbf{x}(n)|C_{\mathrm{mix}})$ for the given audio segment $\mathbf{x}(n)$. Since we assume that $\mathbf{x}(n)$ is a mixed-category audio signal, a category C_{mix} may be a combination of multiple categories, for example, speech and instrument. A time segment $\mathbf{x}(n)$ comprises the sinusoidal trajectories derived from the multiple categories. Thus, we can write the likelihood $p(\mathbf{x}(n)|C_{\mathrm{mix}})$ of mixed category C_{mix} as follows:

$$\begin{aligned} p(\mathbf{x}(n)|C_{\mathrm{mix}}) &= p(s_1(n), s_2(n), \ldots, s_M(n)|C_{\mathrm{mix}}) \\ &= p(s_1(n), s_2(n), \ldots, s_M(n)|c_1(n), c_2(n), \ldots, c_M(n)), \end{aligned} \tag{26}$$

where $s_m(n)$ is a sinusoidal trajectory, which is represented using a temporal feature vector as described in Section 4.1, c_m is a sound category of the trajectory $s_m(n)$, and M is the number of trajectories at time n. The maximal likelihood of Eq. (26) is calculated by finding a combination of the categories $c_m = \hat{c}_m$ to maximize the likelihood of each trajectory. Eq. (26) can be decomposed as

$$\begin{aligned} &p(s_1(n), \ldots, s_M(n)|c_1(n), \ldots, c_M(n)) \\ &\quad = \prod_{m=1}^{M} p(s_m(n)|c_m). \end{aligned} \tag{27}$$

In this equation, we assume that the sinusoidal trajectories are statistically independent of each other for ease of calculation, though some sinusoidal trajectories are indeed related to each other since they are included in the same harmonic structure of a sound.

Consequently, the maximization problem of the likelihood of $p(\mathbf{x}(n)|C_{\mathrm{mix}})$ translates to the selection problem of the category of each trajectory that maximizes the likelihood $p(s_m(t)|c_m)$ separately. The likelihood $p(s_m(t)|c_m)$ is represented using GMMs that are trained by the category-labeled sinusoidal trajectories.

5.2. Speech detection algorithm in mixture sound

As described in Section 5.1, each sinusoidal trajectory $s_m(n)$ is classified into a category $\hat{c}_m$ at the same time as the classification of time segments $\mathbf{x}(n)$. As a type of speech detection, the speech magnitude in each time segment can be estimated by simply summing up the magnitudes of the classified trajectories as speech category. As compared to the estimation of speech magnitude, a calculation of an estimated speech waveform is rather complicated, as described below.

An estimated signal $\hat{x}_c(n)$ of a category c can be constructed by the overlap addition method (OLA) (Rabiner and Schafer, 1978) using the amplitude values of the spectral peaks derived from the classified sinusoidal trajectories of each category, as shown in the following equation:

$$\hat{x}_c(n) = \frac{R}{C_W} \cdot \sum_{n=-\infty}^{\infty} \left[\frac{1}{M_c(rR)} \sum_{m=1}^{M_c(rR)} \widehat{A}_m(rR) \mathrm{e}^{j\hat{\phi}_m(rR)} \right], \tag{28}$$

where $M_c(n)$ is the number of spectral peaks of category c at time n, R is the period in the time domain with which the spectral analysis is sampled, and C_w is a constant determined by the analysis window used in the spectral analysis step. Here, $\widehat{A}_m(n)$ and $\hat{\phi}_m(n)$ are the amplitude and the phase of the mth spectral peak at time n respectively, which belongs to the classified sinusoidal trajectory $s_m(n)$. In this paper, the phase computed from the original mixed signal is used as the estimated phase $\hat{\phi}_m(n)$.

6. Speech classification and detection experiments

We performed experiments on the classification and detection of a speech signal and a mixture of the speech with a musical instrument signal, which were provided through a single channel at a 16-kHz sampling rate. The mixed ratio of speech to instrument was 0 dB. In classification, three categories were used: speech, singing voice, and instrument. The GMMs used in the classification/detection were trained using the dataset mentioned in Section 4.2. In this section, first, the evaluation measures are described; then, practical examples of the classification and detection are mentioned. Subsequently, statistical experiments are performed in order to evaluate the performance of the proposed method: classification tests of time segments of three categories, which will show the basic performance to the detection, and detection tests using a 400-sample dataset of speech mixed with instrumental sound.

6.1. Evaluation measures

For objective measurements of detections in the following experiments, we use the average over the cross-correlation coefficients between narrow-band envelopes of a clean and an estimated signals (NBC) to evaluate the intelligibility of speech sounds (Drullman, 1995). The NBC $\mathrm{Corr}(\mathbf{x}, \hat{\mathbf{x}})$ is defined as the following equation:

$$\mathrm{Corr}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{M} \sum_{m=1}^{M} r(\mathbf{e}_m, \hat{\mathbf{e}}_m), \tag{29}$$

where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are, respectively, the clean and the estimated signals; and $\mathbf{e}_m$ and $\hat{\mathbf{e}}_m$ are the mth narrow-band envelope signals of the clean and the estimated ones, respectively; and $r(\mathbf{e}_m, \hat{\mathbf{e}}_m)$ are the cross-correlation coefficients of $\mathbf{e}_m$ and $\hat{\mathbf{e}}_m$. Here, M is the number of narrow-bands. In this study, a 1/4 octave filter bank is used to compute each narrow-band signal. The filter bank is composed of band-pass filters whose bandwidths are 1/4 octave, and they are arranged closely in the frequency-domain. The frequency range of the filter-bank is from 130 to 2000 Hz, and M is 17. The values of the envelope $\mathbf{e}_m = \{e_m(r)\}$ of the narrow-band signal $x_m(n)$ are computed by windowing and summing up the mth narrow-band signal $x_m(n)$ as follows:

$$e_m(r) = \log_{10} \sum_{n=rR-L/2}^{rR+L/2} \mathrm{Half}(x_m(n)), \tag{30}$$

where we have set $R = 50$ and $L = 100$. Here, $\mathrm{Half}(x)$ is the half-wave-rectified function defined as follows:

$$\mathrm{Half}(x) = \begin{cases} x & (x \geqslant 0), \\ 0 & (x < 0). \end{cases} \tag{31}$$

NBC ranges from -1.0 to 1.0; the more it is closer to 1.0, the more similar the estimated signal is to the clean signal.

The segmental SNR were also used to evaluate the distortion of the detected signals. It is defined as follows:

$$\mathrm{SegSNR} = \frac{1}{N} \sum_{r=0}^{N-1} \left[10\log_{10} \frac{\sum_{n=rR-L/2}^{rR+L/2} x^2(n)}{\sum_{n=rR-L/2}^{rR+L/2} \{\hat{x}^2(n) - x^2(n)\}} \right], \tag{32}$$

where $x(n)$ and $\hat{x}(n)$ are, respectively, the clean and the estimated signals, and N is the number of windowed segments of $x(n)$. We set to $R = 80$ and $L = 160$ in the following experiments.

6.2. Practical examples of the classification and the detection

Fig. 10 shows an example of the tracked (b), classification (d), and detection (c) results for nonmixed speech signal (a), which illustrates basic performance of the tracking and detection of nonmixed signals. As can be seen in (b) and (c), there is no severe distortion of the power envelopes from the clean signal (a) even after the tracking or the detection. The NBC of (a) and (b) is 0.95, and that of (a) and (c) is 0.91, which also indicate that there is no significant distortion from the clean speech. The latter NBC might be the baseline of the next experiment on mixed sound since the NBC for the speech signal with musical instrumental interferences cannot be superior to that for the clean speech. Fig. 10d shows category scores of the three categories. The score $S_{C_{\mathrm{mix}}}(n)$ of category C_{mix} at time n is defined as

$$S_{C_{\mathrm{mix}}}(n) = \frac{p(\mathbf{x}(n)|C_{\mathrm{mix}})}{\sum_{C' \subset C} p(\mathbf{x}(n)|C')}. \tag{33}$$

Here, the notations have the same meaning as Eq. (26). If the score of a category at a time is greater than those of the other categories, the time segment is classified into this category. Fig. 10d illustrates that almost all the classification results are *Speech*, which are the correct classifications, except in the silent segments.

Fig. 11 shows an example of the classification (c) and detection (b) results for mixed speech signal (a) that is the same signal as Fig. 10, but with a musical–instrument signal. As can be seen in (a) and (b) in this figure, the power envelope of the detected speech signal was similar to the original one. The NBC for the detected speech increased to 0.57 from 0.33 for the mixed one. In the classification results (c), 84.5% segments were correctly classified into *Speech+Instrument* category. When nonmixed categories are focused on, the *Speech* and the *Instrument* categories made higher scores than ones of the others when these signals appeared, respectively, except some errors.

Fig. 12 shows the spectrograms of the detected speech (c) from the mixed speech (b) and of the clean speech (a). They show that our method can detect speech well, except for some lost components of the speech, and that musical instrumental components were greatly reduced. Fig. 13 shows the narrow-band envelopes of the detected (bold line), the mixed (broken line) and the clean (solid line) speech. We can see the detection works well in high-frequency bands rather than in low-frequency bands. This result can be explained by the fact that the spectral tracking

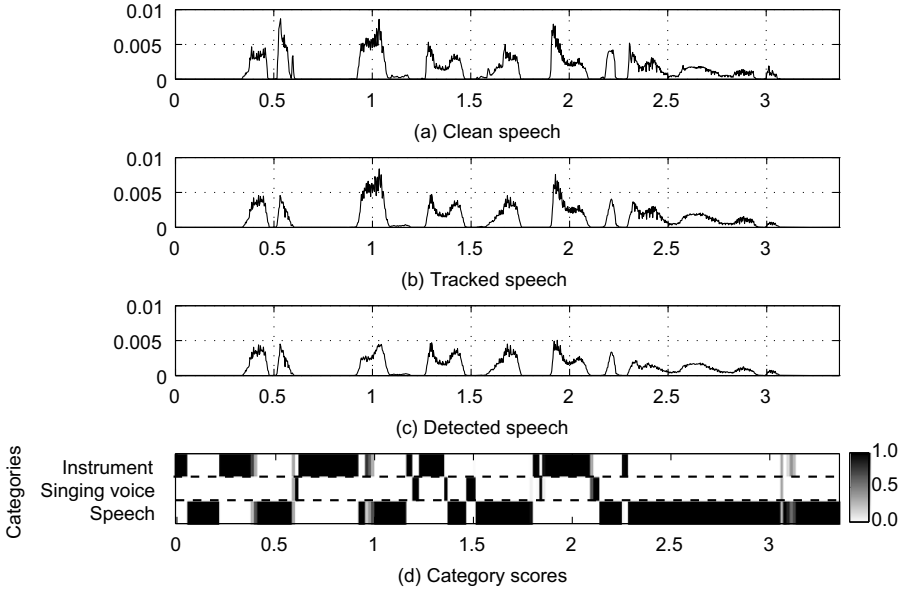


Fig. 10. Sample of power envelopes for tracked (b) and detected (c) signals for a clean speech (a) followed by the category classification scores (d).

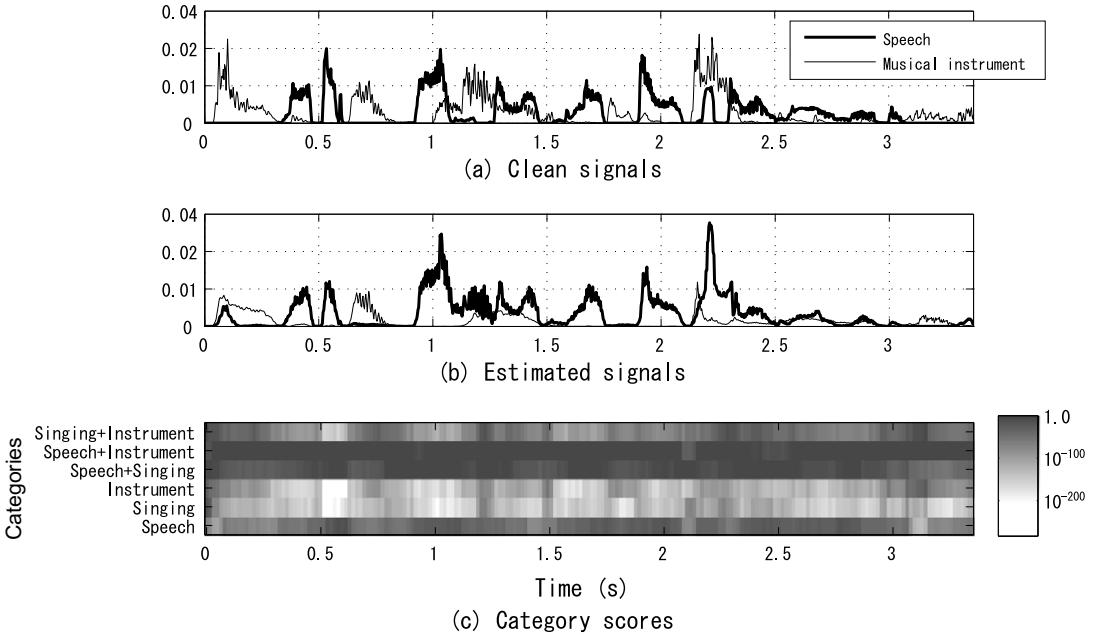


Fig. 11. Detection and classification of signals for a speech signal mixed with a musical–instrument signal. (a) and (b) show clean and estimated power envelopes for speech and musical instrument sound, respectively; (c) shows the category scores of the classification between nonmixed categories and two-mixed categories.

is generally difficult in low-frequency bands, where the multiple fundamental frequencies of harmonic sounds exist closely. The classification is suitably done into the category *Speech+Instrument* in most regions, except the one from 2.3 to 2.5 s.

6.3. Statistical results of the classification and the detection

We also conducted the classification and the detection experiments with multiple samples to statistically evaluate the performance of the method.

First, the classification experiments of time segments in nonmixed sounds were performed by comparing the proposed method to a spectral method using MFCC (Xiong et al., 2003; Kim et al., 2004). The classification performance for a nonmixed signal would provide a basis for detecting a mixed signal since there are many short nonmixed regions even in mixed sounds in practice; further, it would also affect the classification and detection performances in the truly mixed regions.

The MFCC represent the spectral envelope of a sound, while our proposed method uses the temporal characteristics of a sound. The evaluation was performed in a 10-fold cross-validation test with 10,000 one-second-duration samples for each category. The dataset was composed of nonmixed sounds and was obtained from the musical instrument database (Goto et al., 2003) (both instruments and singing voices) and the speech database (Maekawa et al., 2000). Fig. 14 shows the classification accuracy of three methods. In the *MFCC* and the *MFCC_E_D* methods, the GMM classifiers with 16 mixtures and 12th-order MFCC were used; the numbers of the order of MFCC and GMM mixtures were determined considering the results of discussions by Kim et al. (2004) and our preliminary experiments. In the *MFCC_E_D* method, MFCC, logarithmic energy, and their time derivatives were used as the acoustic parameters. In the *ST* method, the classification was done as described in Section 5.1 by considering only three nonmixed categories. The $ST + MFCC$ and $ST + MFCC\_E\_D$ methods employed maximal likelihood classifiers using likelihoods that simply added those of the corresponding two methods. The classification was conducted in one analysis frame in all the three methods.

The precision, recall, and F_1 measure of the *MFCC* and *ST* are shown in Table 2. There were some confusions of the classification between musical instrument and singing voice; thus, the F_1 values of these two categories were lower than those of speech. As shown in Fig. 14, the average of the F_1 measure in the spectral tracking method (*ST*) is 0.939, which is superior to those of *MFCC* (0.918) and *MFCC_E_D* (0.935) in this experiment. Therefore, it can be said that the accuracy of the tracking method is as high as that of conventional spectral classification methods and it might be sufficient to apply to mixed sounds. Moreover, in the combination methods of $ST + MFCC$ and $ST + MFCC\_E\_D$, the F_1 measures were improved as 0.958 and 0.961, respectively, from the methods using the spectral or the temporal characteristics separately. These results imply that temporal and spectral characteristics would contribute to sound classification of category complementarily.

Second, the detection experiments were performed to evaluate the performance of our detection method. Four hundred sentence-utterance-samples mixed with musical

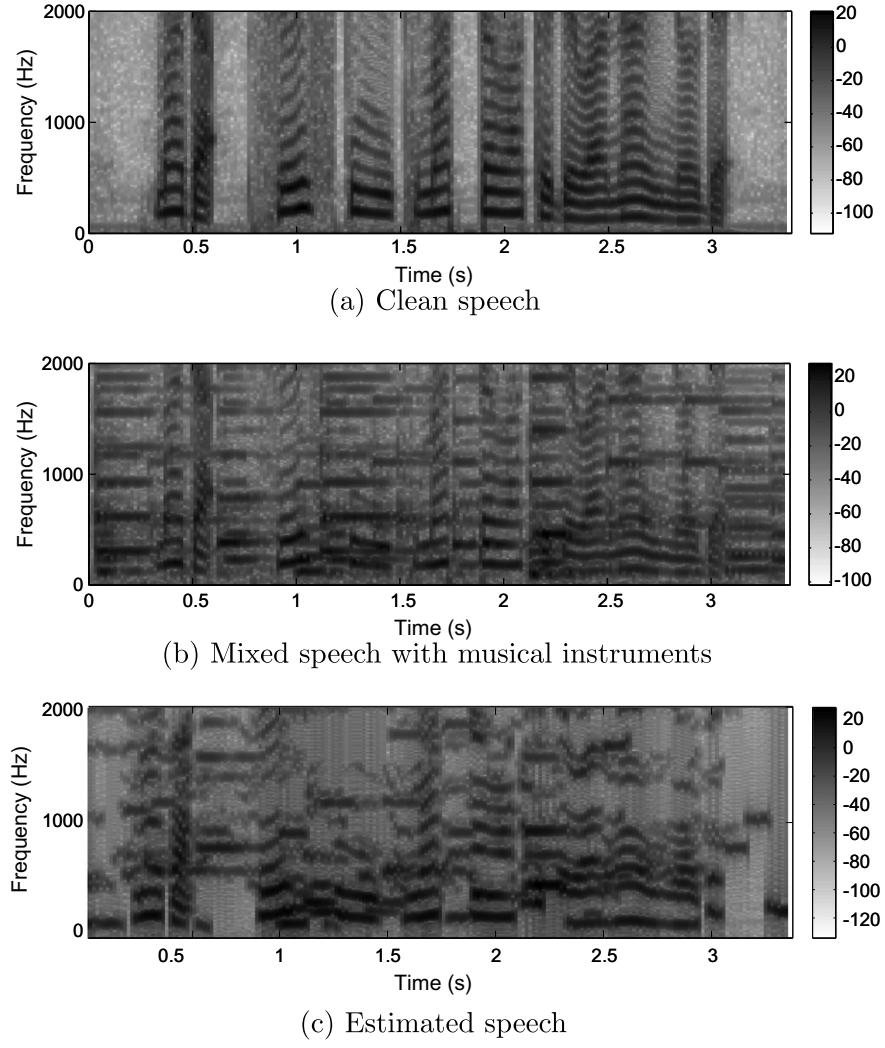


Fig. 12. Spectrograms of the signals same as those in Fig. 11.

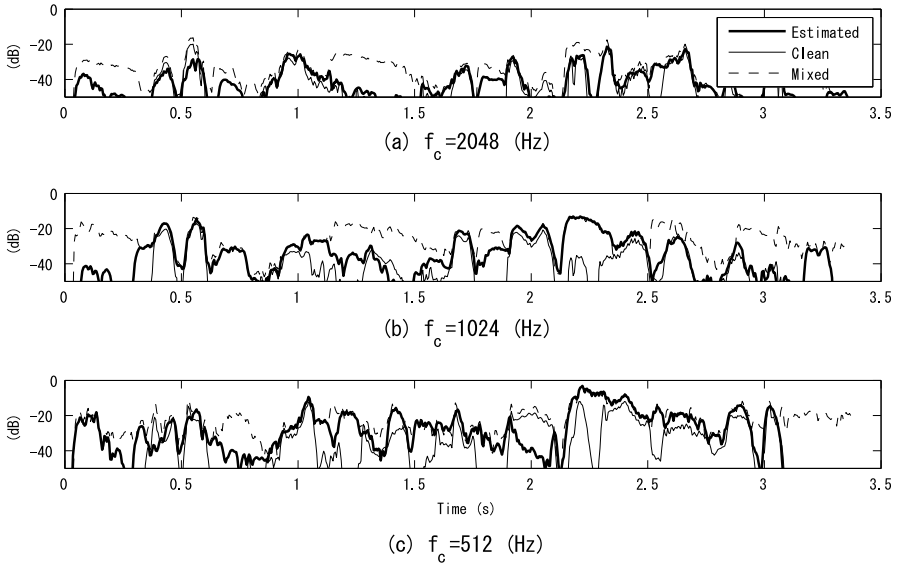


Fig. 13. Envelope of each narrow-band of the clean, mixed, and estimated speech signals shown in Fig. 11.

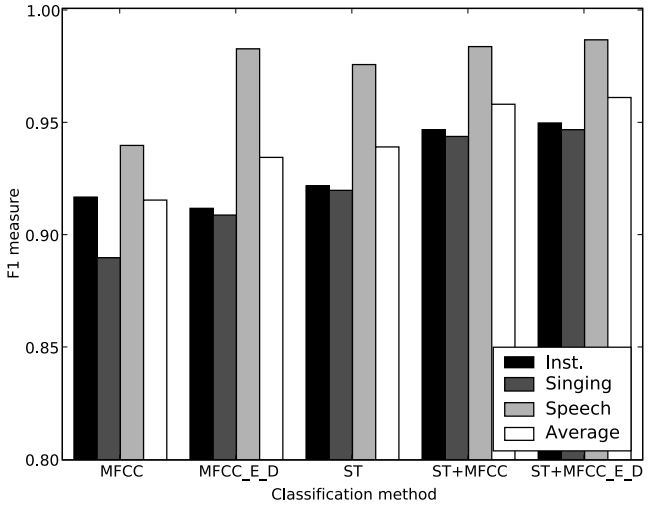


Fig. 14. F_1 measures in the classification of time segments with sinusoidal trajectories, the MFCC and their integrated features: MFCC_E_D implies 12 MFCC coefficients, logarithmic speech energy, and their time derivatives. ST means the sinusoidal trajectory method. ST+MFCC and ST+MFCC_E_D are the integrated methods.

instrumental sounds were prepared. First, 400 utterances from 200 people (100 males and 100 females, 2 utterances each) were extracted from the Japanese Newspaper Article Sentences (JNAS) database (Itou et al., 1999). Second, 400 musical–instrument segments whose durations are individually cut to be the same as that of the prepared 400 utterances of speech were extracted from the tunes of No. 1 to No. 8 of the RWC Jazz Music database (Goto et al., 2002), which are composed of piano or guitar sounds including chords. Then, these sounds were mixed. In order to evaluate the tracking methods, comparisons between our proposed DP + iterative improvement (DP + II) method and the MG method described in Section 3 are also performed.

Tables 3 and 4 show the performance of our *DP + II* method. In this tracking method, the NBC is 0.55 and the improvement of segmental SNR is +5.67 dB, which show that the detected speech samples maintain intelligibility after effectively suppressing competing musical–instrumental sounds. The performance of our method is quite similar to that of the MG method, though its computing time is much smaller than that of the MG method, as described in Section 3.4. Fig. 15 shows the averaged cross-correlation coefficients of the envelopes of the narrow-bands. As noted in the previous section, in high-frequency bands, the correlations are greater than those in low-frequency bands; further, in 1000–2000 Hz bands the correlations become high. These bands are important for intelligibility; thus, it can be said that our detection method would be effective for speech enhancement in this sense. In contrast, in low-frequency bands, the correlations improve very little because of the existence of the

Table 2
The classification results in precision, recall, and F_1 measure of time segments in each sound category, using the 12th order MFCC and sinusoidal trajectories

Category	Precision	Recall	F_1
(a) MFCC			
Instrument	0.908	0.926	0.917
Singing voice	0.921	0.861	0.890
Speech	0.920	0.961	0.940
Average	0.916	0.916	0.916
(b) Sinusoidal trajectory			
Instrument	0.932	0.913	0.922
Singing voice	0.917	0.924	0.920
Speech	0.970	0.982	0.976
Average	0.940	0.940	0.940

Table 3
Comparison of the cross-correlation coefficients of narrow-band envelopes (NBC) between the detection results of two tracking methods: our proposed DP + iterative improvement (DP + II) method and Marks and Gonzalez (2005) (MG) method averaged over 400 samples each of the clean and the estimated signals

Tracking method	Clean	Mixed
MG	0.88	0.55
DP + II	0.89	0.55

Table 4
Comparison of the improvement of segmental SNR (dB) between the detection results of two tracking methods: our proposed DP + iterative improvement (DP + II) method and Marks and Gonzalez (2005) (MG) method averaged over 400 samples of the estimated signals

Tracking method	Mixed
MG	+5.65
DP + II	+5.67

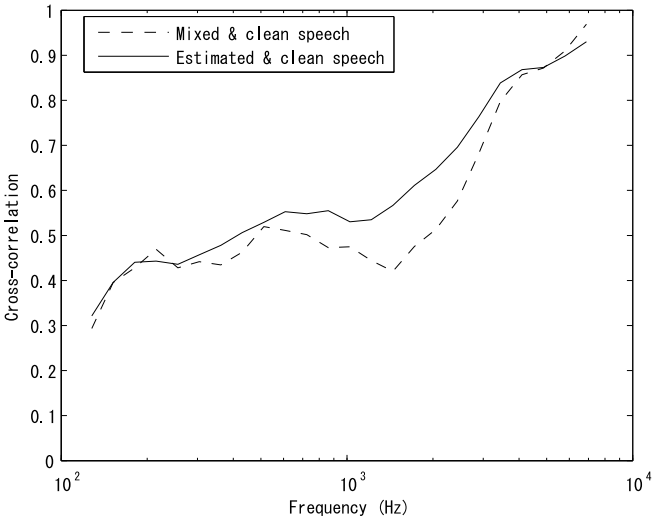


Fig. 15. Narrow-band (1/4 oct. band) cross-correlation coefficients averaged over 400 speech samples. A solid line shows the correlations between an estimated and a clean speech, and a broken line shows the correlation between a mixed and clean speech.

fundamental frequencies of sounds, as mentioned in previous section.

7. Conclusion

We have described (1) a classification method for speech and music and, (2) a detection method based on (1) for speech with background music. To deal with the temporal overlapping of the signal components, we have developed an optimal spectral tracking algorithm based on a DP with iterative improvement for the sinusoidal decomposition of signals. A lower computational complexity of order $O(M^2)$, when the number of spectral peaks in an analysis frame is M, is realized than that of conventional tracking methods. Temporal characteristics of trajectories are utilized as 20 features, and these features were analyzed through PCA and ANOVA to investigate the distribution of feature vectors, and experiments were performed to evaluate their classification performance using GMM into three categories: speech, singing voice, and musical instrument. The average F_1 measure of nonmixed-category time segment classification was 0.939, which might be sufficiently high to apply to subsequent mixed-category detections. The proposed detection has been formulated as determining the optimal combination of trajectory categories by using the maximum likelihood criteria of a time segment. The detection method was also evaluated using a 400-sample dataset of speech with music, and showed that the average of the NBC (narrow-band correlation coefficients) and improvement of the segmental SNR are 0.55 and +5.67 dB, respectively. These results showed that the proposed method performed well for detecting speech segments with background music. It is one of the future problems to apply the proposed method to various speech information processing such as content-based audio indexing or speech-recognition preprocessing. Extension of the detection method for sounds with varying-number mixtures that have more than two categories is also our future problem.

Acknowledgements

This study was partially supported by the Advanced Research Institute for Science and Engineering of Waseda University under the project "Research on Multi-Modal Human Interface Aiming for Spontaneous Communication Systems" and was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research(B), 17300066, 2007.

References

Abe, T., Honda, M., 2006. Sinusoidal model based on instantaneous frequency attractors. IEEE Trans. Audio, Speech Language Process. 14 (4), 1292–1300.

Abe, T., Kobayashi, T., Imai, S., 1996. Robust pitch estimation with harmonics enhancement in noisy based on instantaneous frequency. In: Proc. ICSLP 9, Vol. 2, pp. 1277–1280.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press.

Chou, W., Gu, L., May 2001. Robust singing detection in speech/music discriminator design. In: Proc. ICASSP 2001, Vol. II, pp. 865–868.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39 (1), 1–38.

Depalle, P., Garcia, G., Rodet, X., 1993. Tracking of partials for additive sound synthesis using hidden markov models. In: ICASSP-93, Vol. 1, pp. 225–228.

Drullman, R., 1995. Temporal envelope and fine structure cues for speech intelligibility. J. Acoust. Soc. Amer. 97, 585–592.

Goto, M., Hashiguchi, H., Nishimura, T., Oka, R., 2002. RWC music database: popular, classical, and jazz music databases. In: Proc. 3rd Internat. Conf. on Music Information Retrieval (ISMIR 2002), pp. 287–288.

Goto, M., Hashiguchi, H., Nishimura, T., Oka, R., 2003. RWC music database: music genre database and musical instrument sound database. In: Proc. 4th Internat. Conf. on Music Information Retrieval (ISMIR 2003), pp. 229–230.

Hogg, R., Ledolter, J., 1987. Engineering Statistics. MacMillan.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., Itahashi, S., 1999. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. J. Acoust. Soc. Jpn. (E) 20 (3), 199–206.

Jackson, J., 1991. A User's Guide to Principal Components. John Wiley and Sons, p. 592.

Kazama, M., Yoshida, K., Tohyama, M., 2003. Signal representation including waveform envelope by clustered line-spectrum modeling. J. Audio Eng. Soc. 51 (3), 123–137.

Kim, H., Burred, J., Sikora, T., 2004. How efficient is MPEG-7 for general sound recognition? In: 25th Internat. Audio Engineering Society Conference Metadata For Audio.

Maekawa, K., Koiso, H., Furui, S., Isahara, H., 2000. Spontaneous speech corpus of Japanese. In: Proc. 2nd Internat. Conf. Language Resources and Evaluation (LREC2000), pp. 947–952.

Marks, S., Gonzalez, R., 2005. Techniques for improving the accuracy of sinusoidal tracking. In: Proc. Internet and Multimedia Systems and Applications 2005, pp. 299–304.

McAulay, R., Quatieri, T., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. ASSP ASSP-34 (4), 744–754.

Melih, K., Gonzalez, R., 1999. Audio source type segmentation using a perceptually based representation. In: Proc. 5th Internat. Symposium on Signal Processing and Its Applications, 1999, ISSPA'99, Vol. 1, pp. 51–54.

Melih, K., Gonzalez, R., 2000. Source segmentation for structured audio. In: IEEE Internat. Conf. on Multimedia and Expo, ICME 2000, Vol. 2, pp. 811–814.

Moore, B.C.J., 2004. An Introduction to the Psychology of Hearing, fifth ed. Elsevier Academic Press, pp. 269–298 (Chapter 8).

Nawab, S.H., Espy-Wilson, C.Y., Mani, R., Bitar, N.N., 1998. Computational auditory scene analysis. Lawrence Erlbaum Associates, Knowledge-based analysis of speech mixed with sporadic environmental sounds, pp. 177–194 (Chapter 12).

Plante, F., Meyer, G., Ainsworth, W.A., 1995. A pitch extraction reference database. In: EUROSPEECH'95. pp. 837–840.

Rabiner, L., Schafer, R., 1978. Digtal Processing of Speech Signals. Prentice-Hall, pp. 274–277 (Chapter 6.1.5).

Sakakibara, K.-I., Osaka, N., 1998. On concatenation of musical sounds using a sinusoidal model. In: Technical Report of IEICE, Vol. SP97-108, pp. 1–6 (in Japanese).

Saunders, J., 1996. Real-time discrimination of broadcast speech/music. In: Proc. ICASSP'96, Vol. 2, pp. 993–996.

Scheirer, E., Slaney, M., 1997. Construction and evaluation of a robust multifeature speech/music discriminator. In: Proc. ICASSP'97, Vol. II, pp. 1331–1334.

Takeuchi, S., Yamashita, M., Uchida, T., Sugiyama, M., 2001. Optimization of voice/music detection in sound dat. In: Consistent and Reliable Acoustic Cues for sound analysis (CRAC Workshop).

Taniguchi, T., Adachi, A., Okawa, S., Honda, M., Shirai, K., 2005. Discrimination of speech, musical instruments and singing voices using the temporal patterns of sinusoidal segments in audio signals. In: Proc. Interspeech2005, pp. 589–592.

Taniguchi, T., Tohyama, M., Shirai, K., 2006. Spectral frequency tracking for classifying audio signals. In: IEEE Internat. Symposium on Signal Processing and Information Technology, 2006, pp. 300–303.

Torkkola, K., 1999. Blind separation for audio signals – are we there yet? In: Proc. Internat. Workshop on Independent Component Analysis and Signal Separation (ICA'99).

Virtanen, T., 2003. Sound source separation using sparse coding with temporal continuity objective. In: Proc. ICMC, pp. 231–234.

Virtanen, T., Klapuri, A., 2000. Separation of harmonic sound sources using sinusoidal modeling. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process, ICASSP'00, Vol. 2, pp. 765–768.

Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T., 2003. Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification. In: Proc. Internat. Conf. on Multimedia and Expo, 2003, ICME'03, Vol. 3, pp. 397–400.