



# Singer Diarization for Polyphonic Music With Unison Singing

Hitoshi Suda , Daisuke Saito, *Member, IEEE*, Satoru Fukayama, *Member, IEEE*, Tomoyasu Nakano, and Masataka Goto 

**Abstract**—This paper introduces a new framework for singer diarization, which is a technique to reveal who sings when in songs with multiple singers. Although various techniques have been developed to analyze and extract features of singing voices in musical audio signals, most of them assume that a song is sung by a single singer, and singer diarization for multiple singers has not been well studied in the field of singing information processing. To deal with multiple speakers in speech analysis, speaker diarization has been explored to handle overlapped speech voices, but cannot handle singing voices well because of acoustic differences between singing and speech voices. This paper therefore proposes a new diarization framework specialized in singing voices. To achieve high accuracy in overlap detection, this paper proposes a novel acoustic feature named *Cosacorr score*, which is helpful in estimating whether a song is sung by more than one singer. After extracting singing voices from polyphonic music by using a singing voice separation technique, the framework adopts an existing ArcFace technique to extract discriminative singer representations from short segments of the separated singing voices. The framework is evaluated by using a new private dataset of unison singing voices, which is constructed using commercially available compact discs (CDs). The experimental results show that the proposed framework outperformed the baseline method for speaker diarization in terms of diarization error rate (DER).

**Index Terms**—Music information processing, music information retrieval, singer diarization, unison singing.

## I. INTRODUCTION

SINGING is one of the most important elements of music [1], [2] since many people listen to music with a focus on singing [3]. Given its importance, various research activities related to singing have been pursued and attracting attention not only from a scientific viewpoint, but also from the standpoint of commercial applications [4]. Singing information processing [1], [2], [5] is defined as music information processing for singing voices, and covers diverse topics ranging from basic research on the features unique to singing to applied research

Manuscript received August 27, 2021; revised February 6, 2022 and April 4, 2022; accepted April 4, 2022. Date of publication May 3, 2022; date of current version May 6, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefan Bilbao.

Hitoshi Suda and Daisuke Saito are with the Department of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan (e-mail: hitoshi@gavo.t.u-tokyo.ac.jp; dsk\_saito@gavo.t.u-tokyo.ac.jp).

Satoru Fukayama, Tomoyasu Nakano, and Masataka Goto are with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan (e-mail: s.fukayama@aist.go.jp; t.nakano@aist.go.jp; m.goto@aist.go.jp).

Digital Object Identifier 10.1109/TASLP.2022.3166262

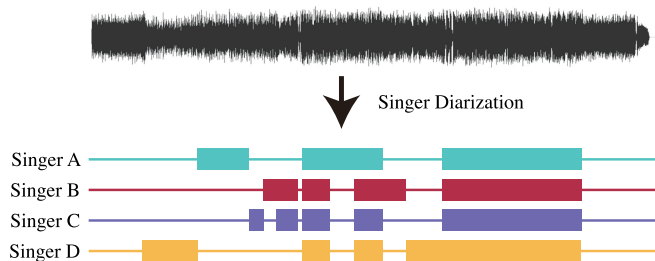


Fig. 1. Conceptual image of singer diarization. An input song is labeled according to singers at each audio frame.

such as that on singing synthesis [6], lyrics transcription and synchronization, vocal timbre analysis, singing skill evaluation, and music information retrieval (MIR) based on singing voices [1]. In the research field of singing information processing, various techniques have been developed to analyze and extract features of singing voices in musical audio signals, such as fundamental frequency ( $F_0$ ), amplitude, phoneme timing, timbre, expression, and skill. Most of such techniques assume that a song is sung by a single singer.

However, there exist a large number of songs sung by multiple singers. In fact, duets, in which two singers alternatively sing, are common in popular music. In various types of Japanese popular music such as idol songs sung by idol groups and anime songs featured in animation films, a song is often sung alternatively by more than two singers. Such a song can be divided into several temporal sections corresponding to different singers. We call such a divided music structure *song division*, and also call a technique to estimate who sings when in songs with song division *singer diarization*. Fig. 1 shows a conceptual image of singer diarization. Since the technique reveals how many singers are singing and when they are singing, singer diarization is a fundamental technique for the analysis of songs sung by multiple singers.

Singer diarization is derived from speaker diarization, which is a technique to reveal who speaks when in conversational speech. In the field of speech analysis, speaker diarization has been studied since the late 1990s and can be applied to automatic annotation of a wide range of conversational speech such as telephone conversations, broadcast news, debates, and meetings [7]. On the other hand, automatic singer diarization has been underexplored in the field of singing information processing. The purpose of this study is to achieve singer diarization to extract

TABLE I

COMPARISON OF THE STATISTICS OF THE ACOUSTIC FEATURES BETWEEN SPEECH AND SINGING VOICES. AS A SPEECH DATASET, THE SPEECH DATA OF JAPANESE PHONETICALLY BALANCED SENTENCES UTTERED BY THE SPEAKER FTK IN THE ATR JAPANESE SPEECH DATASET [14] WAS USED. AS A SINGING VOICE DATASET, TOHOKU KIRITAN'S SINGING-VOICE DATABASE [15] WAS USED

Category	$\log F_0$ [log Hz]	Phoneme duration [ms]
Speech	$5.44 \pm 0.24$	$82 \pm 37$
Singing voices	$5.87 \pm 0.31$	$167 \pm 204$

$\pm$  denotes the standard deviation.

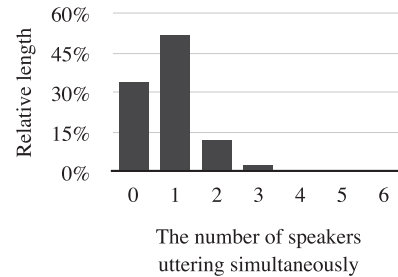
song-division information from songs sung by multiple singers so that the extracted information could be utilized in applications such as analysis and visualization of music structure, extraction of a particular singer's voice, and music information retrieval based on singer diarization.

In the literature, a related task is target-singer tracking (TST) [8]. It estimates whether a target singer is singing or not at each audio frame under the assumption that its singer is known and the acoustic model of the singer can be trained in advance. On the other hand, singer diarization does not have such an assumption and estimates who sings when as shown in Fig. 1.

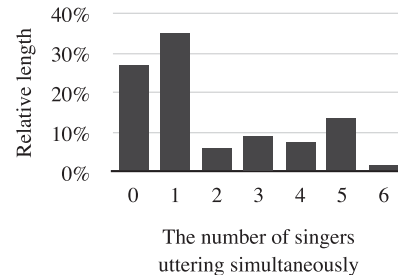
In the field of ethnomusicology, Thlithi *et al.* [9] studied a technique for singer diarization of music signals recorded in sub-Saharan countries. Since it simply used a traditional speaker diarization method, its performance was limited and there were three issues from our viewpoints. Firstly, the technique was not able to handle overlapped singing voices. Although some studies [10]–[12] proposed methods that can perform recognition of overlapped segments in speaker diarization, handling overlapped singing using such techniques could cause a different type of difficulty. This is because the temporal overlap structures are significantly different between singing voice and speech. Secondly, the sounds of background instruments degraded the performance of singer diarization. To address this issue, reduction of background music or sound source separation should be employed. Lastly, acoustic differences between singing and speech voices could deteriorate the diarization performance, though this issue was not discussed in [9].

According to our analysis in comparing speech and singing voices, the range of  $F_0$  is wider and the duration of phonemes is longer in singing voices than that in speech voices as shown in Table I. A comparative study [13] indicates that spectral features can be affected by those differences, which makes it difficult to acquire singer information from short segments. Moreover, the distribution of the number of simultaneous speakers and singers is also different. Fig. 2 shows an example of such distributions. Compared to conversational speech, singing tends to have more simultaneous singers singing at the same time. Furthermore, singing voices tend to be more synchronized than speech voices; multiple singers sometimes sing in almost the same rhythm and even at the same pitch (i.e., unison singing). Therefore, because of these factors, singer diarization has a different type of difficulty from speaker diarization.

This paper proposes a new practical framework for singer diarization by addressing the above three technical issues. To



(a) CHiME-6 evaluation data, which is a conversational speech dataset [12, 16]



(b) The evaluation set used in this paper

Fig. 2. Histograms of relative lengths of audio signals according to the number of simultaneously uttering speakers (singers).

reduce diarization errors caused by simultaneous singers, unison singing is particularly difficult. The paper focuses on overlap detection that identifies overlapped segments in which multiple singers sing in unison — i.e., sing at almost the same pitch in the same rhythm, following the same musical notes. To achieve high accuracy in such overlap detection, this paper proposes *Cosacorr score*, a novel acoustic feature based on autocorrelation. A higher *Cosacorr* score indicates that a song is more likely to be sung by more than one singer. Our framework also employs an existing singing voice separation technique to extract singing voices from polyphonic music including background instrumental sounds. The framework then adopts ArcFace[17], an architecture that provides embeddings for face recognition, to extract discriminative singer representations from the separated singing voices. In the experiments, we construct a new private dataset of unison singing voices of Japanese idol songs using commercially available CDs and evaluate the performance using real music samples.

The rest of this paper is organized as follows. Section II describes some advanced speaker diarization techniques that can handle overlapped speech as related works. Section III describes speaker diarization methods on which the proposed method is based. Section IV shows a detailed description of the proposed method for singer diarization. Section V and VI describe experiments to evaluate the proposed method. Section VII discusses the method, and Section VIII concludes the paper.

## II. RELATED WORKS

In the speaker diarization field, several studies have introduced methods that can effectively handle overlapped speech.

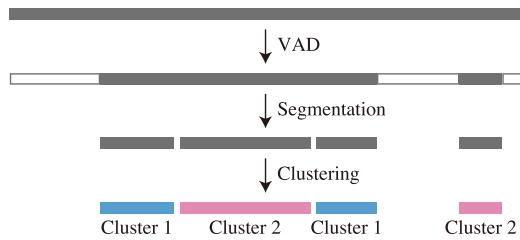


Fig. 3. Overview diagram of the clustering-based method for speaker diarization. The method consists of 3 steps: VAD, speaker segmentation, and speaker clustering.

End-to-end neural diarization (EEND) is a method to perform speaker diarization directly by neural networks [11]. In the study, permutation-invariant training (PIT) loss function is adopted to deal with the permutation problem in end-to-end diarization. To improve the performance of EEND, self-attentive EEND (SA-EEND), which incorporates self-attention with EEND, has been proposed [18]. EEND has a problem that the maximum number of speakers is predetermined by the architecture of the model. To enhance EEND to speaker diarization with the unknown number of singers, an architecture named encoder-decoder based attractor calculation (EDA) has been adopted to EEND [19].

Some studies detect overlapped segments at first, and perform clustering-based diarization for single-speaker segments [20], [21]. The proposed method in this paper adopts a similar approach to the studies.

Diarization methods based on source separation have also been introduced [22]. The study performs speech separation in advance to ensure that each separated source contains only one speaker at most. Such an approach can also be regarded as a method to detect overlap in advance. However, in singer diarization, singing voice separation is more challenging than speech separation. This is because singing voices are often synchronized and harmonized. Moreover, singing voices have almost the same pitch when singers sing in unison. Therefore, this paper does not adopt this approach.

### III. BASELINE METHODS FOR SPEAKER DIARIZATION

This section describes key baseline methods for speaker diarization which underlies the proposed method in this paper. These methods mainly aim at conversational speech such as broadcast news shows and dinner parties.

#### A. Clustering-Based Speaker Diarization

The most fundamental method for speaker diarization is a clustering-based one [7], [23]. The method consists of three steps: voice activity detection (VAD), segmentation, and clustering. Fig. 3 shows a diagram of the method. The method assumes that at most one speaker speaks at the same time, and does not take overlapped speech into account.

VAD is a process to detect whether each segment is speech or non-speech. VAD is an indispensable process for high-quality diarization because non-speech segments can degrade the performance of the acoustic models used in the latter processes.

Moreover, VAD errors directly deteriorate evaluation metrics. VAD based on support vector machines (SVMs) and linear discriminant analysis (LDA) classifies each frame using acoustic features and discriminative models [24]–[26]. Instead of frame-by-frame analysis, time-series modeling is also adopted. Similarly to automatic speech recognition, VAD methods based on hidden Markov models (HMMs) are proposed [27]. Recurrent neural networks (RNNs) and long-short term memory (LSTM) networks are also incorporated for time-series modeling of speech occurrences [28], [29].

In terms of singing voice detection, several approaches are proposed to improve VAD in mixed signals. As an approach to improve acoustic features, for example, fluctograms, which represent vocal fluctuations without extracting pitch, and vocal variances, which detect slow changes of the shape of the vocal tract, are adopted as well as spectral flatness and spectral contraction [30]. Harmonic-percussive sound separation is also adopted as preprocessing of singing voice detection [31]. By adding restriction to parameters in convolutional neural networks, a VAD method that is not affected by loudness is proposed [32].

Speaker segmentation is a step to detect speaker turns. In this step, speech is separated into segments so that each segment is spoken by a single speaker. Speaker segmentation is performed by repeating the operation of splitting a segment at the most reasonable position. The most traditional segmentation method is based on the Bayesian information criterion (BIC) [33]. To determine where to split the segment and whether to stop splitting, a hypothesis with the least BIC value is selected.

In the clustering step, acquired segments are grouped into some clusters by speakers. The step is the most principal process on speaker diarization. Mainly two clustering approaches are adopted: bottom-up one and top-down one. Bottom-up clustering, or agglomerative hierarchical clustering (AHC), starts from an under-clustered state and reduces the number of clusters gradually. As an initial state, the approach often assumes that all segments belong to different speakers. On the other hand, top-down clustering supposes that all segments are clustered into one speaker at first, and then iteratively splits the clusters. The clustering is performed based on a criterion such as BIC or Kullback–Leibler (KL) divergence [33], [34]. Spectral clustering is also adopted as a clustering method in speaker diarization [35]. By utilizing normalized maximum eigengap (NME) values, a method that can effectively perform spectral clustering and estimate the number of speakers without tuning parameters is proposed [36].

Speaker representation, which is an embedding of speaker information, is also utilized for speaker clustering. I-vector [37] is one of the common representations for speaker recognition and is employed in speaker diarization [38], [39]. As a distance metric for i-vectors, scores based on probabilistic linear discriminant analysis (PLDA) are adopted. Alternate segmentation and clustering methods are also introduced by utilizing speaker representations that can be acquired from extremely short audio signals [40]. Since the methods can extract speaker representations from fixed-length short segments, the methods only need to perform uniform segmentation and do not require precise segmentation based on the speaker turns.

## B. Speaker Diarization Based on Target-Speaker VAD

Since the traditional clustering-based method assumes that each segment is uttered by at most one speaker, the method cannot recognize overlapped speech. This paper refers to a speaker diarization method based on target-speaker voice activity detection (TS-VAD) [12], and the proposed method is based on the concept of this TS-VAD-based diarization. The method uses TS-VAD, which can directly perform speaker diarization of audio signals including overlapped segments. The method aims at speaker diarization in dinner-party scenarios where 4 speakers speak with each other. This method consists of the following steps.

- 1) *Initial diarization*: Diarization is performed for the entire speech. Each segment is supposed to be uttered by a single speaker at most in this step, and thus any diarization method can be incorporated. The method executes diarization by clustering fixed-length segments according to speaker representations. The method adopts x-vectors [41] as speaker representations.
- 2) *Extraction of speaker-specific representations*: Based on the results of initial diarization, an i-vector [37] is calculated for each speaker.
- 3) *TS-VAD*: In this step, voice activity at each frame is estimated for each speaker. Since this is equivalent to VAD about a specific speaker, this step is named target-speaker VAD. The TS-VAD model  $f_{\text{TS-VAD}}$  estimates each speaker's activity based on acoustic features and the speaker's representation as follows:

$$\begin{aligned} & [s_{1,[1:T]}, s_{2,[1:T]}, s_{3,[1:T]}, s_{4,[1:T]}] \\ & = f_{\text{TS-VAD}}(\mathbf{x}_{[1:T]}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4), \quad (1) \end{aligned}$$

where  $\mathbf{y}_i$  denotes the speaker representation of the  $i$ -th speaker,  $\mathbf{x}_{[1:T]}$  is a sequence of acoustic features, and  $s_{i,[1:T]}$  denotes a sequence of VAD scores about the  $i$ -th speaker. The maximum number of speakers must be determined in advance. In this case, the maximum number of speakers is set to four.

- 4) *Postprocessing*: To suppress improper results, some filtering is applied to the results of diarization. In the framework, four types of postprocessing are adopted: median filtering, elimination of short segments, score thresholding, and Viterbi decoding.

Since the speaker-specific representations are extracted from the entire speech, the representations are affected by overlapped segments, and the performance of TS-VAD can be degraded. To improve the overall performance, the method iteratively updates the speaker representations by repeating steps 2) and 3). The results of TS-VAD are gradually optimized by recalculating i-vectors based on those of the previous iteration.

In TS-VAD, the number of recognized speakers needs to be predetermined in the original paper that proposes TS-VAD. In this study, the number of speakers was fixed to 4. The authors of [42] enhanced TS-VAD to the unknown number of speakers by inputting random representations when the number of speakers is less than that of predetermined, or ignoring the least frequent speaker when the number of speakers is more than that of predetermined.

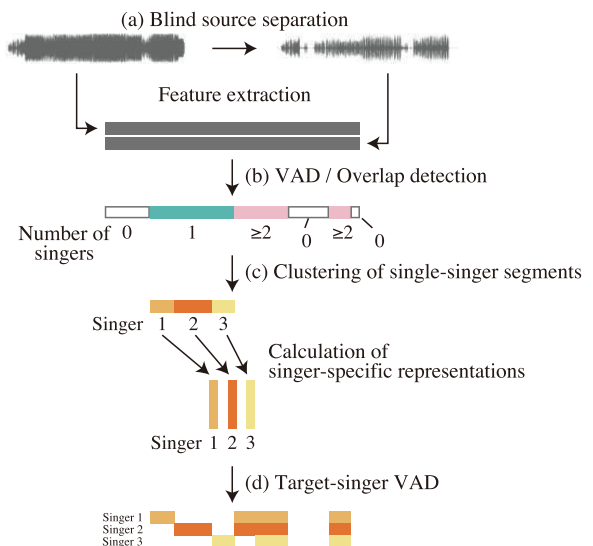


Fig. 4. Overview diagram of the proposed framework. The acoustic features are extracted from both mixed and separated signals.

## IV. PROPOSED FRAMEWORK FOR SINGER DIARIZATION

### A. Overview of the Framework

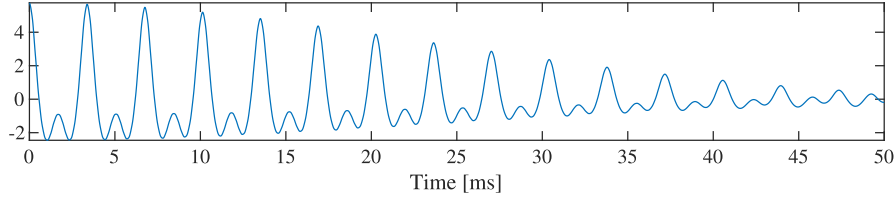
This paper introduces a diarization framework tuned for singer diarization. In contrast to the TS-VAD-based approach, the proposed framework handles overlapped segments before the main diarization process and adopts a singing voice separation technique to reduce background music. Fig. 4 shows the overview of the framework. The remainder of this section describes the modules of the framework.

### B. Preprocessing

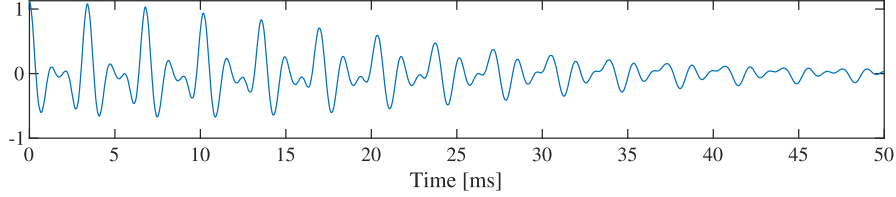
Singing voices are extracted from mixed signals by a singing voice separation technique based on blind source separation (BSS). This step corresponds to (a) in Fig. 4. This paper adopts Spleeter [43], which is an open-source BSS library for music signals. The pretrained model of Spleeter utilizes 12-layer U-Net [44] for estimation of soft masks of spectrograms. Since the separation is not ideal, the separated signals are unnatural and distorted. Therefore, the overall performance is degraded if only the separated signals are used. In the proposed method, acoustic features extracted from both mixed and separated signals are jointly used in the subsequent processes.

### C. Voice Activity and Overlap Detection

Traditional frameworks for speaker diarization often perform diarization with the assumption that all the segments are uttered by at most a single speaker and then handle overlapped segments by postprocessing. However, total durations of overlapped segments in multiple-singer songs are generally longer than those in conversational speech, and thus the acquired singer representations based on the results of the prior diarization can be degraded because of long overlapped segments. Hence, the traditional diarization method may not be able to achieve as high accuracy in the case of singer diarization as in the case of speaker diarization. To suppress the effects of long overlapped segments,



(a) Vocal signals where one singer sings. The summation of the 8th-order Cosacorr scores is 0.0029.



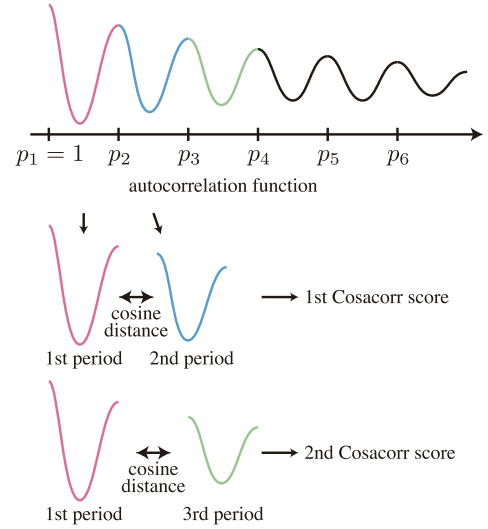
(b) Vocal signals where two singers sing. The summation of the 8th-order Cosacorr scores is 0.3048.

Fig. 5. Examples of autocorrelation functions of vocal signals. Both images show signals of the same song at the same time.

the proposed framework adopts an approach that performs overlap detection before the main diarization process. Some studies have already adopted similar approaches to improve the quality of speaker representations [20]–[22].

In this step, according to the number of singers, each frame is classified into three classes: no singers, one singer, and multiple singers. Therefore, this step is named voice activity and overlap detection. This step corresponds to (b) in Fig. 4. This classification is performed using bidirectional LSTM (BiLSTM) networks. In this paper, acoustic features that are used as the input to the networks consist of mel-frequency cepstral coefficients (MFCC), power, and Cosacorr scores, which are introduced in this paper.

The Cosacorr scores proposed in this paper are cosine distance scores of autocorrelation. When multiple singers sing at the same time, the fundamental frequencies are slightly different from each other even if they sing in unison. Because of the difference, autocorrelation of acoustic signals shows a non-periodic (quasi-periodic) trajectory. Fig. 5 shows examples of autocorrelation functions of acoustic signals of singing voices. Fig. 6 shows a conceptual visualization of the calculation process. A Cosacorr score measures the difference between the shapes of the first and the other periods. The feature can evaluate the non-periodicity and quantify the multiplicity of singers. Let  $X = [x_1, x_2, \dots, x_N]$  be an autocorrelation sequence of the acoustic signals. First, local maxima  $x_{p_1}, x_{p_2}, \dots, x_{p_P}$  in  $X$  are detected using a peak detection algorithm. Because of the nature of the autocorrelation function, the position of the first peak  $p_1$  is always 1 (i.e.,  $x_1$  is the highest). As shown in Fig. 6, the 1st Cosacorr score represents the difference between the 1st period and the 2nd period in the autocorrelation of singing voices. The  $n$ -th Cosacorr score represents the difference between the 1st period and the  $n + 1$ -th period. In calculating such differences, the length of the 1st period and the  $n + 1$ -th period can be slightly different from each other; thus, the  $n + 1$ -th period needs to be resampled to match with the length of the 1st period. That is, to evaluate the difference, the  $n + 1$ -th period of the autocorrelation may need to be expanded or contracted so that

Fig. 6. Visualization of the calculation process of Cosacorr scores. The  $n$ -th Cosacorr score is obtained by calculating cosine distance between 1st and  $(n + 1)$ -th period of autocorrelation function.

the number of samples in the  $n + 1$ -th period matches that in the first period. In detail, the  $n$ -th Cosacorr score is calculated by

$$\text{Cosacorr}_n = \frac{P_{n+1}}{P_1} \left( 1 - \frac{\sum_{i=1}^{p_2-1} x_i y_{n,i}}{\sqrt{\sum_{i=1}^{p_2-1} x_i^2} \sqrt{\sum_{i=1}^{p_2-1} y_{n,i}^2}} \right), \quad (2)$$

$$P_m = \frac{1}{p_{m+1} - p_m} \sum_{i=p_m}^{p_{m+1}-1} x_i^2. \quad (3)$$

The sequence  $y_{n,1}, y_{n,2}, \dots, y_{n,p_2-1}$  is generated by linearly resampling the  $n + 1$ -th period  $x_{p_{n+1}}, x_{p_{n+1}+1}, \dots, x_{p_{n+2}-1}$  so that it can be used to calculate cosine distance with the 1st period  $x_1, x_2, \dots, x_{p_2-1}$ . That is, by performing linear interpolation several times, the  $n + 1$ -th period is expanded or contracted, and a sequence with  $p_2 - 1$  samples is generated.  $\frac{P_{n+1}}{P_1}$  is a scale

factor for weighting scores by power. A higher Cosacorr score indicates that the segment is more likely to be sung by multiple singers. In this paper, a sequence of scores obtained from  $n$  periods is called  $n$ -th-order Cosacorr scores. For example, the 3rd-order Cosacorr scores are a sequence of  $\text{Cosacorr}_1$  (the 1st Cosacorr score),  $\text{Cosacorr}_2$  (the 2nd Cosacorr score), and  $\text{Cosacorr}_3$  (the 3rd Cosacorr score).

If the peak detection algorithm fails or the detected periods are not appropriate, the Cosacorr score could not be appropriately calculated and is set to zero. Although the Cosacorr score is designed to be calculated from singing voices after singing voice separation, this paper dares to calculate the Cosacorr score from singing voices with accompaniments (i.e., music signals before singing voice separation) in addition to the original Cosacorr score from the separated singing voices without accompaniments. This is because both types of scores could be helpful as acoustic features in the overlap detection.

#### D. Extraction of Singer Representations

In this step, singer representations are extracted from fixed-length segments. In traditional frameworks, i-vectors [37] and x-vectors [41] are often adopted as speaker representations. However, singing voices have a wider range of  $F_0$  and longer phoneme duration than speech. Therefore, it can be difficult to learn discriminative representations from singing voices of short duration using the traditional techniques. To extract more discriminative singer representations, this paper adopts ArcFace-based singer representation.

ArcFace, or Additive Angular Margin Loss, is a network architecture that can obtain highly discriminative embeddings [17]. The architecture is originated in deep face recognition. Conventional techniques have utilized bottleneck features of classifiers that are trained based on softmax loss. However, these methods do not acquire proper embeddings for clustering, where intra-class features are aggregated and inter-class features are diverse. Some approaches achieve more discriminant embeddings by adding margin penalties to softmax loss [45], [46]. ArcFace is one of those techniques that replaces softmax loss for

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^n e^{s \cos \theta_j}}, \quad (4)$$

where  $N$  and  $n$  denote batch size and the number of classes, respectively,  $m$  is a margin penalty,  $y_i$  denotes a class index where the  $i$ -th feature  $x_i$  belongs, and  $s$  is a scaling parameter.  $\theta_j$  is defined by

$$\theta_j = \arccos \frac{W_j^\top x_i}{\|W_j\| \|x_i\|}, \quad (5)$$

where  $W_j$  is the  $j$ -th column of a weight matrix of the last fully connected layer of the neural networks. Compared to conventional approaches, ArcFace is easy to implement and achieves effective performance on face recognition. Because of the advantages, the proposed framework utilizes ArcFace as an extractor for singer representation.

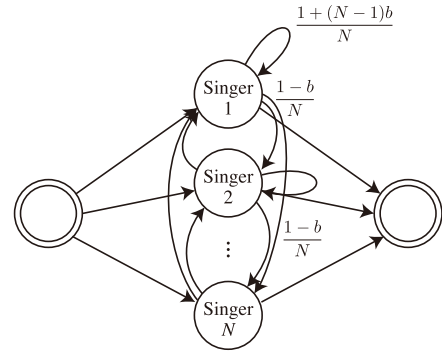


Fig. 7. HMM for postprocessing of singer clustering.  $N$  denotes the number of singers, and  $b$  is a hyperparameter where  $0 < b < 1$ . Each state except initial and final states corresponds to each singer.

#### E. Clustering of Single-Singer Segments

This step performs clustering of single-singer segments based on the extracted singer representations. This step is equivalent to segmentation and clustering steps in the clustering-based diarization, which is described in Section III-A. This step corresponds to (c) in Fig. 4. This paper adopts a spectral clustering algorithm based on NME analysis that can automatically tune parameters of the algorithm [36].

This clustering process does not consider sequential information. Therefore, the results can include frequent singer turns and too short segments. To eliminate such improper results, Viterbi decoding is applied for postprocessing. The framework utilizes an HMM where each state corresponds to each singer. Fig. 7 shows a conceptual image of the HMM.

At the end of the step, a singer representation is calculated for each singer.

#### F. Target-Singer VAD

This step reveals whether each singer sings at each segment where multiple singers sing. This step is shown as (d) in Fig. 4, and corresponds to the TS-VAD step in the TS-VAD-based baseline method. Instead of TS-VAD, this paper constructs networks in the same way as Personal VAD [47]. In contrast to TS-VAD, target-singer VAD estimates each singer's singing state separately. For each target singer, all segments are classified into three classes: 1) no singers are singing, 2) the target singer is singing, 3) someone else than the target singer is singing. The target singer can be identified in the case of the second class.

Fig. 8 depicts the architecture of target-singer VAD. The target-singer VAD model is based on acoustic features, target singer's representation, and cosine similarity between the target singer's representation and the representation at that segment. The architecture adopted in this paper is referred to as *score and embedding conditional training (SET)* defined in [47].

Although all segments including no-singer, single-singer, and multiple-singer segments are thus classified into the three classes by this architecture, only the identified target singer of each multiple-singer segment is used as the final VAD results. Since each no-singer segment cannot have any singer and the singer of each single-singer segment has already been known before the

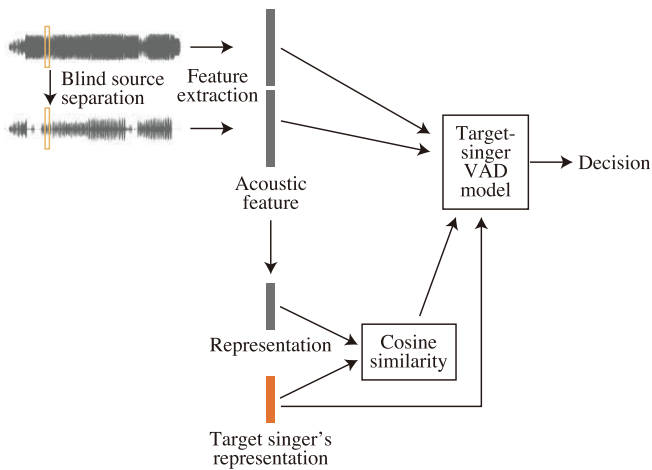


Fig. 8. Architecture of target-singer VAD. In the figure, for simplicity, the acoustic features used for target-singer VAD and singer representation extraction are the same. In the experiment, the acoustic features extracted under different conditions were used.

target-singer VAD, the classified results for those segments do not need to be used.

## V. EFFECTIVENESS OF COSACORR SCORE

This section describes preliminary experiments to check the effectiveness of Cosacorr scores. In this toy experiment with a small dataset, the systems recognized whether given audio signals were sung by one or two singers. In addition to the Cosacorr-based system, i-vector-based one was also constructed. In this experiment, i-vectors were utilized for overlap detection in the same way as speaker recognition [37]. That is, two-class classification was performed based on i-vectors.

### A. Experimental Setups

As the small dataset, twelve songs were prepared. The songs were taken from commercially available CDs that contain an unusual set of singing voices of Japanese idol songs. In these CDs, each song is repeatedly covered by a different singer. Given five singers, for example, the first audio track is a recording of the song  $m_1$  sung by the first singer  $s_1$ , the second track is a recording of the same song  $m_1$  sung by the second singer  $s_2$ , the third track is a recording of the same song  $m_1$  sung by the third singer  $s_3$ , and so on. It thus results in five different CD tracks corresponding to the five singers. The number of singers varies from four to twelve, depending on the song. Since those singers sing along the same accompaniment (backing track), their singing voices are temporally synchronized. The backing tracks (karaoke tracks) of all the songs are also provided in those CDs. We, therefore, took advantage of this special set of CDs. The total number of recordings (audio tracks) used in this dataset is 125, and the number of unique singers is 18. The duration of the songs is about 2 minutes and the sampling frequency is 44100 Hz. The detailed information about the dataset is available at <https://www.gavo.t.u-tokyo.ac.jp/%7ehitoshi/diarization/>.

Each recording taken from the CD track thus contains audio signals of a single singer with accompaniments and does not have

simultaneous singers. Since this paper focuses on unison singing voices, we randomly mixed two recordings by two different singers for each of the songs in order to generate the same number of unison recordings (i.e., 125 unison recordings). We extracted (separated) solo singing voices (125 solo recordings) from polyphonic music recordings by subtracting the corresponding backing track by using an existing tool Utagoe-Rip<sup>1</sup>, and then randomly mixed them to obtain unison recordings. Since all recordings are in stereo, they were converted to monaural signals by calculating the mean over channels.

Ten out of the twelve songs were used for training, and the remaining two songs, song A and song B, were used for evaluation. The songs A and B were sung by 12 and 4 singers, respectively, and 12 solo recordings for the song A and 4 solo recordings for the song B were obtained. The same number of unison recordings (12 unison recordings for the song A and 4 unison recordings for the song B) were also prepared and used for evaluation. To prepare two different conditions, the singers of the song A were included in the training set, and the singers of the song B were not. That is, the song A is in the closed-singer condition, and the song B is in the open-singer condition.

The frame period was 10 ms. From each frame, the summation of the 8th-order Cosacorr scores was calculated. As a feature, the mean and variance of the summation of the consecutive 100 frames were used. Consequently, 2-dimensional features (mean and variance) were extracted from 1 s (100 frames) of audio signals.

To extract i-vectors, MSR Identity Toolbox [48] was used. As acoustic features, 16th-order MFCC and their  $\Delta$  and  $\Delta^2$  features were extracted. The universal background model (UBM) was a Gaussian mixture model (GMM) with 2048 mixtures. The number of dimensions of i-vectors was fixed to 100. In the same way as the system based on Cosacorr scores, recognition was performed using each 1-second audio signal.

A fusion system is also evaluated in this experiment. In the system, i-vectors were used as features in addition to the mean and variance of Cosacorr scores. That is, the input features were 102-dimensional.

As classification models, SVMs with the Gaussian kernel were adopted. SVMs were trained with 5-fold cross-validation.

### B. Results

Table II shows the results. The system based on Cosacorr scores achieved overlap detection at more than 70% on average. The results suggest that the proposed feature helps detection of unison signals. The results also indicate that Cosacorr scores were effective even in the open-singer condition. In this experiment, i-vectors were also helpful in overlap detection and effective in both closed-singer and open-singer conditions. Moreover, the fusion system using both Cosacorr scores and i-vectors outperforms the other independent systems. These preliminary results indicate that not only MFCC-based i-vector features but also Cosacorr scores can be effectively exploited in unison detection.

<sup>1</sup>[Online]. Available: <https://www.vector.co.jp/soft/win95/art/se127635.html>

TABLE II  
ACCURACY OF OVERLAP DETECTION. THE LABELS SOLO AND UNISON  
DENOTE THE CORRECT CLASS

System	Song	Solo	Unison	Average
Cosacorr score	A (closed-singer)	78.4%	63.5%	70.9%
	B (open-singer)	66.6%	87.9%	77.3%
i-vector	A (closed-singer)	81.8%	82.9%	82.4%
	B (open-singer)	81.9%	85.1%	83.5%
Fusion	A (closed-singer)	82.6%	82.9%	82.8%
	B (open-singer)	81.7%	95.3%	88.5%

## VI. EVALUATION OF DIARIZATION SYSTEM

This section describes experiments for evaluation of the proposed singer diarization system. In this experiment, real audio samples recorded in CDs were utilized for evaluation.

### A. Data Preparation

To evaluate the proposed singer diarization method, a new larger dataset was prepared using the same set of unusual CDs described in Section V-A. All songs are sung by female singers in the dataset, and each song is sung by each of multiple singers at almost the same pitch in the same rhythm. Since those singing voices are thus synchronized, they can be mixed to generate unison singings for training purposes. In addition, for the evaluation purpose, we used other CDs to prepare a set of songs sung by multiple singers. The detailed information about the datasets is available at <https://www.gavo.t.u-tokyo.ac.jp/%7ehitoshi/diarization/>.

1) *Training Set*: To prepare a training set and a development set, fifty-three songs were prepared. Each song was sung by about 9 singers. The total number of recordings (audio tracks) used in this dataset was 500. The total duration of the dataset was about 32 hours, and the total duration of the voiced segments in the dataset was about 24 hours. The number of unique singers was 22.

We first extracted (separated) solo singing voices (500 recordings) from polyphonic music recordings by subtracting the corresponding backing tracks<sup>2</sup> as described in Section V-A. In randomly mixing those solo singing voices, unlike the small dataset, we here generated 526 multiple-singer recordings with song division, such as those shown in Fig. 1. That is, each recording of a song contains sections with a solo singing voice, sections with two unison singing voices, and sections with three unison singing voices by changing a way of mixing and concatenating singing voices along the song. All the 526 recordings were generated so that the number of contained singers is 3.

<sup>2</sup>Although backing tracks of 16 songs are provided in CDs, backing tracks of 37 songs are not. When backing tracks are not available, we estimated them by leveraging multiple recordings. Given a song, its multiple recordings with different singers include the same backing track. We first compute the amplitude spectrogram of each recording of the same song. Then, since the pitch and timing of their singing voices deviate from each other, we assume that the smallest amplitude of each time-frequency bin in all the spectrograms could correspond to the backing track. We thus estimate the amplitude spectrogram of the backing track and perform phase reconstruction to estimate audio signals of the backing track.

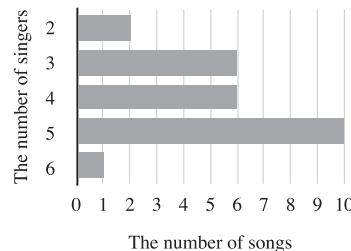


Fig. 9. Distribution of the number of singers in the evaluation set.

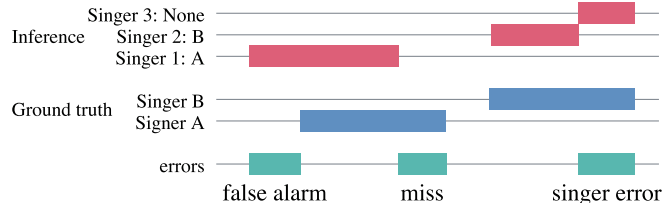


Fig. 10. Conceptual image of three types of errors calculated in DER. In the figure, the singer 1 corresponds to the reference singer A, and the singer 2 corresponds to the reference singer B. In this case, there is no singer to which the singer 3 corresponds.

Fifty-one out of the fifty-three songs were used for the training set. The remaining two songs were used for the development set in which the number of multiple-singer recordings was 19.

2) *Training Set for Voice Activity and Overlap Detection*: To train the voice activity and overlap detection system, twenty-five multiple-singer songs with song division were directly taken from different regular CDs and were used without modification. The duration of each song was about 5 minutes, and the total duration of the dataset was about 113 minutes. They were manually labeled (annotated) according to the number of singers. The songs were divided into three groups: training set, closed-singer development set, and open-singer development set. The numbers of songs were 19, 2, and 4, respectively. The singers in the training set included all the singers in the closed-singer development set, but did not include any singers in the open-singer development set.

3) *Evaluation Set*: For the evaluation purpose, we further prepared twenty-five multiple-singer songs with song division. Those songs were directly taken from different CDs and were used without modification. The duration of each song was about 5 minutes, and the total duration of the dataset was about 112 minutes. The number of singers varied from two to six, depending on the song. Fig. 9 shows the distribution of the number of singers. All 57 singers in this evaluation set were not included in both of the training sets described in the previous sections.

All the songs were manually labeled with the ground truth annotation as shown in the blue bars (Ref) in Fig. 13. In general, each song contains sections with a solo singing voice and sections with multiple unison singing voices. Although two out of the twenty-five songs contain additional sections with multiple non-unison singing voices (with different pitches) for about 32 seconds in total, they did not affect the evaluation of the performance since they are so short.



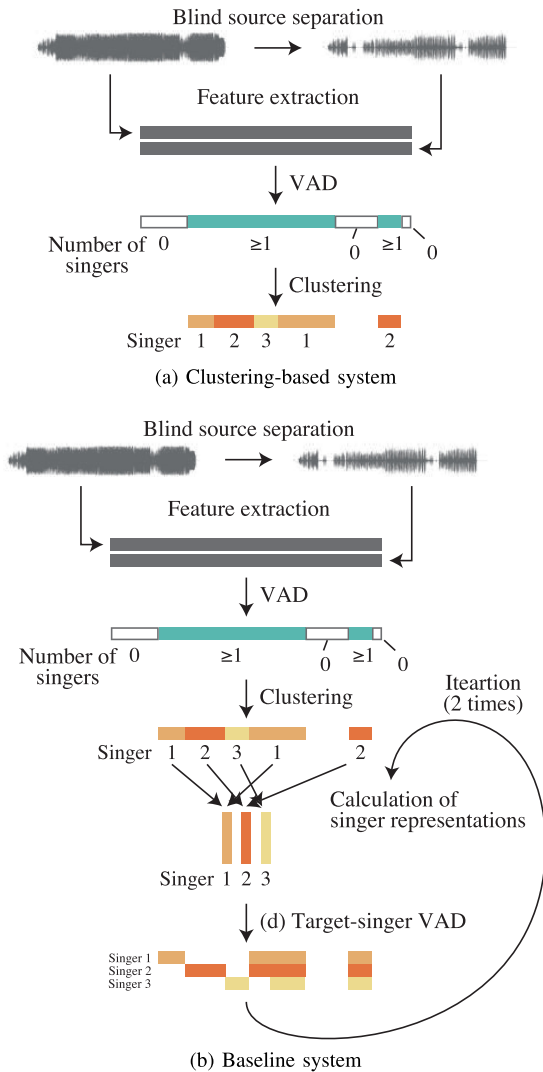


Fig. 11. Diagrams of two evaluated systems in the experiment (cf. the proposed system shown in Fig. 4).

### B. Experimental Setups

Spleeter [43] was used to extract singing voices from polyphonic music. Since the pretrained model *2stems-16 kHz* was used, the separated signals are band-limited by 16000 Hz. Note that acoustic features such as MFCC and Cosacorr scores were extracted from both of the separated singing voices and the original recordings (mixed signals).

For the voice activity and overlap detection, BiLSTM networks were utilized. Table III shows the architecture of the network. The input features were 24th-order MFCC with 0th coefficients, power, and the 8th-order Cosacorr scores. The frame period was 100 ms, and the input features were jointed over 21 frames. The total number of dimensions of input features was  $(25 + 1 + 8) \times 2 \times 21 = 1428$ . The network performed classification into three classes: no singer, one singer, and multiple singers. The network was trained using the Adam optimizer, and the learning rate was 0.001. The batch size was 19, and the number of epochs was 200. In the experiments, the development set was used as a validation set.

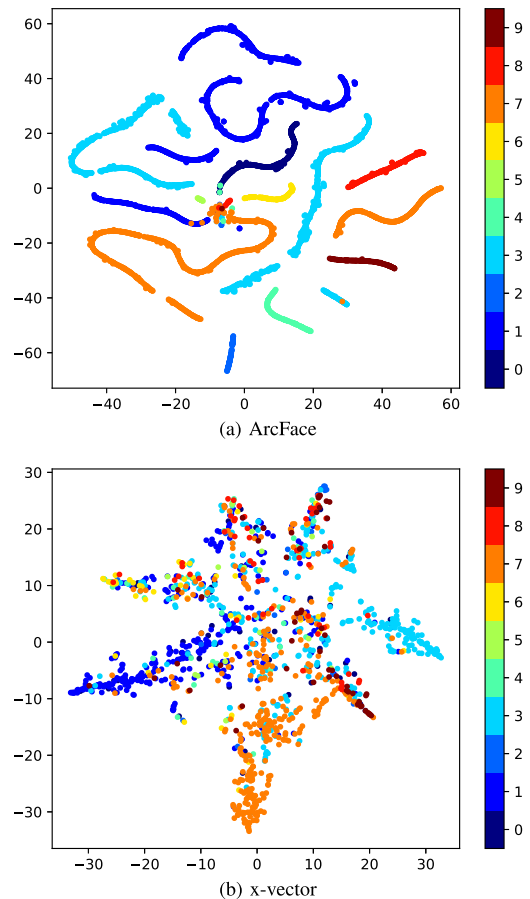


Fig. 12. Visualization of singer representations by *t*-SNE. The colors denote the singers of the samples.

TABLE III  
NETWORK ARCHITECTURE FOR THE VOICE ACTIVITY AND OVERLAP DETECTION AND THE TARGET-SINGER VAD

Layer type	Output size
Input	1428 or 849
Batch normalization	
BiLSTM	1024
Fully connected	1024
Batch normalization	
Fully connected	1024
Batch normalization	
BiLSTM	1024
Fully connected	1024
Batch normalization	
Fully connected	1024
Batch normalization	
Fully connected	1024
Batch normalization	
Fully connected	3
Softmax	3

For the target-singer VAD, BiLSTM networks were adopted. Fig. 8 shows the architecture of target-singer VAD, and Table III shows the architecture of the network. The input features were 24th-order MFCC with 0th coefficients, power, the 8th-order Cosacorr scores, cosine similarity between the representation of

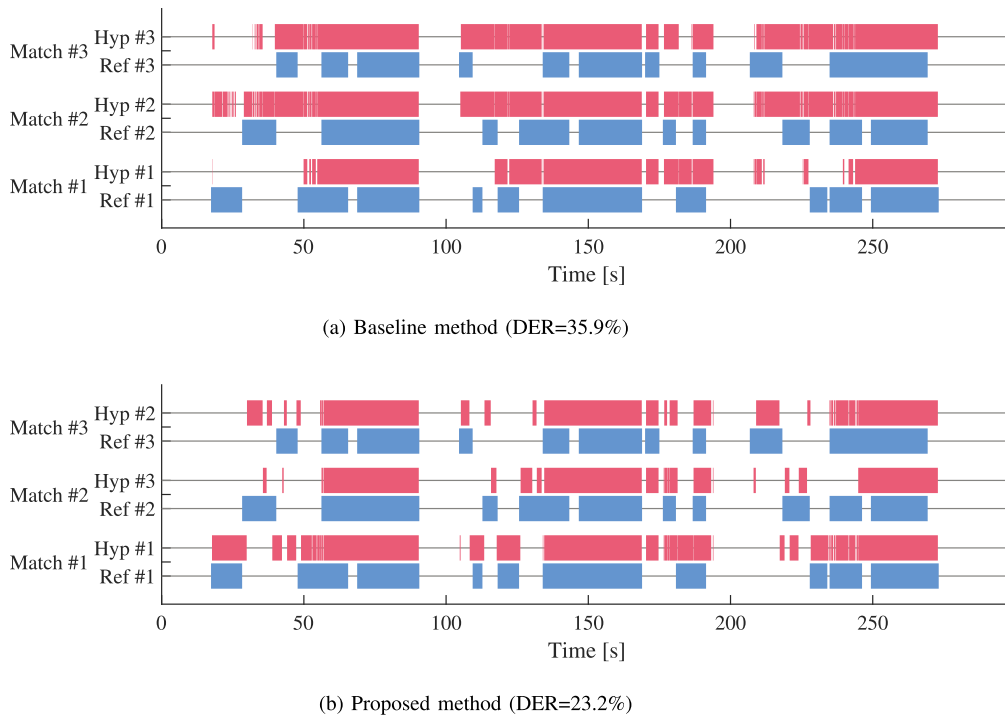


Fig. 13. Diarization results of a song. Blue bars (Ref) denote the ground truth, and red bars (Hyp) denote the estimated results. In the figure, the pairs used in the calculation of DER are shown as “Match”. Both results were obtained in the condition where the number of singers was given.

TABLE IV  
NETWORK ARCHITECTURE FOR THE EXTRACTOR OF SINGER REPRESENTATIONS

Layer type	Output size
Input	4050
Fully connected	2048
Fully connected	2048
Batch normalization	
Fully connected	2048
Batch normalization	
Fully connected	2048
Fully connected	100
Normalization	100
Margin addition (when training)	100
Softmax	100

the target singer and one at the segment, and the representation of the target singer. The frame period was 100 ms, and the acoustic features were jointed over 11 frames. The total number of dimensions of input features was  $(25 + 1 + 8) \times 2 \times 11 + 1 + 100 = 849$ . The network was trained using the Adam optimizer, and the learning rate was 0.001. The batch size was 8, and the number of epochs was 20. In the experiments, the development set was used as a validation set.

The network for extracting singer representations was based on ArcFace [17]. Table IV shows the architecture of the network. The features were jointed over 25 frames with 20 ms frame period, that is, the length of the input features was 500 ms. As acoustic features, 79th-order MFCC with 0th coefficients and power were used. The total number of dimensions of input features was  $(80 + 1) \times 2 \times 25 = 4050$ . The singer representations were 100-dimensional. The margin of ArcFace was fixed

to 0.5. The network was trained using the Adam optimizer, and the learning rate was 0.0001. The batch size was 32768, and the parameters at the 1572nd epoch were utilized. One percent of the training set was used as a validation set.

As a clustering method, auto tuning spectral clustering [36] was adopted, and open-source implementation<sup>3</sup> was utilized. A parameter tuning method based on normalized maximum eigengap, which is named NME-SC in [36], was utilized. After clustering, postprocessing with HMM was performed. HMM is constructed as shown in Fig. 7, and the hyperparameter  $b$  was set to 0.9999, which is fixed based on the development set. The emission probability matrices were also fixed as hyperparameters based on the development set.

As an objective metric for evaluating diarization results, diarization error rate (DER) [49] was adopted. DER is defined by the ratio of the length of three types of errors to the total length of the reference utterances. The three types of errors are

- 1) *Singer error*: The labeled singer is wrong.
- 2) *False alarm*: The number of estimated singers is more than the true number of singers.
- 3) *Miss*: The number of estimated singers is less than the true number of singers.

Fig. 10 shows a conceptual image of the three types of errors. DER can be formulated as

$$\text{DER} = \frac{\sum_{s=1}^S \tau_s \left( \max \left( N_s^{(\text{ref})}, N_s^{(\text{hyp})} \right) - N_s^{(\text{correct})} \right)}{\sum_{s=1}^S \tau_s N_s^{(\text{ref})}}, \quad (6)$$

<sup>3</sup>[Online]. Available: <https://github.com/tango4j/Auto-Tuning-Spectral-Clustering>

TABLE V

RESULTS OF VOICE ACTIVITY AND OVERLAP DETECTION. ACCURACY DENOTES THE CLASSIFICATION ACCURACY FOR DEVELOPMENT SET. THE RESULTS OF BASELINE ARE CALCULATED BASED ON THE RESULTS OF DIARIZATION

Model	Input features	Accuracy	
		Closed-singer	Open-singer
LSTM	MFCC, power	81.1%	68.4%
LSTM	MFCC, power, Cosacorr score	88.1%	76.6%
BiLSTM	MFCC, power	86.6%	75.1%
BiLSTM	MFCC, power, Cosacorr score	<b>89.9%</b>	<b>79.7%</b>
	Baseline	52.5%	60.4%

where  $S$  denotes the number of segments,  $\tau_s$  is the length of the  $s$ -th segment,  $N_s^{(\text{ref})}$  and  $N_s^{(\text{hyp})}$  are the number of singers in reference and estimated label at the  $s$ -th segment, and  $N_s^{(\text{correct})}$  denotes the number of matched singers. Since diarization itself does not perform singer recognition, the technique does not estimate which inferred singer corresponds to a singer in the ground truth. Therefore, to calculate DER, one-to-one mapping is first performed, and then DER is calculated. Because of the definition, DER may exceed 100%. In these experiments, collar tolerance was not used.

### C. Evaluation of Voice Activity and Overlap Detection

To evaluate the performance of voice activity and overlap detection, some systems were compared. For comparison, an LSTM-based model was constructed by replacing BiLSTM layers with LSTM layers. In addition, the performance of a baseline method was evaluated. The baseline method is a modified version of the speaker diarization method based on TS-VAD so that it can be applied to singer diarization. In the baseline method, source separation is performed, and TS-VAD is replaced with target-singer VAD. Fig. 11(b) shows a diagram of the baseline method. In the baseline method, the overlap detection is not performed, and the diarization results are constructed using target-singer VAD for all voiced segments. To calculate the results of the baseline method, the full diarization process was conducted, and then the accuracy in voice activity and overlap detection was calculated. Table V shows the results. By introducing BiLSTM and Cosacorr scores, the total performance achieved 89.9% in the closed-singer set and 79.7% in the open-singer set. The results indicate that Cosacorr scores were effective in voice activity and overlap detection. The results also show that the BiLSTM-based systems outperformed LSTM-based ones. This can be because stronger time-series modeling was effective. In singer diarization, there is no need to perform real-time analysis, and the later information can be utilized. Therefore, the BiLSTM-based model was adopted in the latter experiments.

### D. Quality of Singer Representations

In addition to the ArcFace-based system, an x-vector-based system was implemented for comparison. In the x-vector-based system, consecutive 51 frames with 10 ms frame period were used as input.

Firstly, acquired singer representations were visualized. Fig. 12 shows the results. Singer representations acquired by the

TABLE VI

AVERAGE DER FOR SINGLE-SINGER SEGMENTS OF DEVELOPMENT SET WITH DIFFERENT EXTRACTOR OF SINGER REPRESENTATIONS

Extractor	Number of singers	
	Given	Not given
ArcFace	<b>36.7%</b>	<b>48.7%</b>
x-vector	44.8%	59.3%

ArcFace-based system formed clusters according to the singers while the representations were dispersed in x-vector-based one.

Secondly, diarization performance was evaluated. In this experiment, diarization was performed for the single-singer segments in the development set. Table VI shows the results. No matter whether the number of singers was given or not, the ArcFace-based system outperforms the x-vector-based one. Since the singer representations obtained with the ArcFace-based system were more discriminative than that of the x-vector-based system, the performance of the clustering-based diarization has been improved.

### E. Overall Performance Evaluation

To evaluate the performance of the proposed system, three diarization systems were compared: the clustering-based system, the baseline system, and the proposed system. Fig. 4 shows a diagram of the proposed system, and Fig. 11 shows diagrams of the other two systems. The first method performed diarization by only clustering and was not able to handle overlapped singing. In the baseline method, calculation of singer-specific representations and target-singer VAD were iteratively performed three times. The major difference between the baseline system and the proposed system is whether the overlap detection is performed before the clustering step. Both methods adopted ArcFace-based singer representations. The systems were evaluated in two different conditions. The number of singers was given at the clustering step in the first condition and not given in the second condition.

Table VII shows the results of the overall performance in DER. The results show that the proposed system outperformed the other baseline methods. By performing overlap detection before the main diarization step, the performance was improved in both conditions. On the other hand, the results of the clustering-based method show significantly higher DER than the other systems. The clustering-based approach did not handle overlapped segments, and thus a lot of misses raised DER.

The information about the number of singers notably affects the overall performance of singer diarization. The results are reasonable because the errors in the number of singers caused misses or false alarms.

### F. Impact of the Quality of the Modules on the Overall Performance

To evaluate the impact of the quality of the modules on the overall diarization performance, the proposed system is evaluated in two additional conditions. In the first condition (a), the results of the clustering-based diarization of single-singer segments are given based on the ground truth. On the basis

TABLE VII  
AVERAGE OVERALL DER FOR DEVELOPMENT AND EVALUATION SET. DIARIZATION WAS PERFORMED IN TWO DIFFERENT CONDITIONS; THE NUMBER OF SINGERS WAS GIVEN IN THE FIRST CONDITION AND WAS NOT GIVEN IN THE SECOND CONDITION

(a) The number of singers was given									
Method	Development set				Evaluation set				
	DER	singer error	false alarm	miss	DER	singer error	false alarm	miss	
Clustering-based	149.0%	12.6%	1.2%	135.2%	191.2%	31.7%	2.7%	156.7%	
Baseline	77.7%	3.1%	15.9%	58.6%	72.9%	8.0%	20.3%	44.6%	
Proposed	<b>55.9%</b>	7.9%	4.0%	44.0%	<b>52.9%</b>	10.6%	26.5%	15.8%	

(b) The number of singers was not given									
Method	Development set				Evaluation set				
	DER	singer error	false alarm	miss	DER	singer error	false alarm	miss	
Clustering-based	153.5%	17.2%	1.2%	135.2%	191.9%	32.3%	2.7%	156.9%	
Baseline	92.0%	4.6%	22.5%	64.9%	129.5%	12.1%	25.1%	92.3%	
Proposed	<b>69.3%</b>	10.2%	3.3%	55.8%	<b>79.3%</b>	13.2%	24.8%	41.2%	

TABLE VIII  
COMPARISON OF THE PROPOSED SYSTEM IN THREE CONDITIONS IN DER. THE TOTAL NUMBER OF SINGERS IS GIVEN BASED ON THE GROUND TRUTH

Condition	Voice activity and overlap detection	Clustering-based diarization	Development set	Evaluation set
Condition (a)	ground truth	ground truth	24.4%	14.3%
Condition (b)	ground truth	inferred	30.2%	38.1%
Condition (c)	inferred	inferred	55.9%	52.9%

of the oracle clustering results, singer-specific representations were calculated, and then target-singer VAD was performed. The results of voice activity and overlap detection are also generated from the ground truth. In the second condition (b), only the results of voice activity and overlap detection are given from the ground truth. On the basis of the oracle results of voice activity and overlap detection, the clustering-based diarization of single-singer segments is performed. These two conditions are compared with the third condition (c), which is equivalent to the proposed method in Table VII(a) and does not use the ground truth. The total number of singers was given in this experiment.

By comparing these three conditions (a), (b), and (c), the impact of the quality of voice activity and overlap detection and clustering-based diarization of single-singer segments can be evaluated. Since all voiced segments are taken into account in the calculation of DER, the denominator in the definition of DER (6) is identical under all three conditions.

Table VIII shows the results. The results show that the quality of the clustering-based diarization as well as that of voice activity and overlap detection have a large impact on the DER.

## VII. DISCUSSIONS

The experimental results show that the proposed framework outperformed the baseline system. Fig. 13 shows an example of the results of diarization of a song. In the result of the baseline method, a lot of false alarms in the single-singer segments raised overall DER. As shown in Table V, the accuracy in overlap detection was lower in the baseline method, and the result indicates that misses and false alarms degraded the final DER. On the other hand, the proposed framework suppressed these errors by utilizing the results of overlap detection. Therefore, results indicate the effectiveness of this strategy. The experimental results showed the effectiveness of Cosacorr scores in overlap

detection, and thus Cosacorr scores were helpful to lower the final DER.

The proposed method was effective also in overlapped segments. Table IX shows the confusion matrices in the estimated number of singers. In overlapped segments, the results show that the proposed system estimates the singing state of each singer with fewer errors in the number of singers. In addition, the results in Table VII show that the amount of misses is small in the results acquired by the proposed method. This performance difference can be caused by the difference in the quality of the singer representations. The proposed system adopts an approach that performs overlap detection at first and extracts singer representations only from single-singer segments. The improvement in the quality of singer representations seems to lead to the improvement of singer diarization in overlapped segments.

The performance of the proposed method is still limited. One reason is the lack of accuracy in the clustering-based diarization of single-singer segments. Since the proposed system acquires singer-specific representations based on the results of this diarization, the performance greatly affects the final DER. As shown in Table VIII in Section VI-F, when the ground truth was used as the result of the clustering step (condition (a)), the DER was much lower than the other conditions (b) and (c). The improvement of the clustering-based diarization is thus the key point to reduce the DER. Another reason is the lack of quality in the voice activity and overlap detection. As shown in Table VIII, when the ground truth was used as the result of its detection (conditions (a) and (b)), the DER was much lower than the condition (c). This is because errors in this detection deteriorate the results of clustering-based diarization, resulting in worse singer-specific representations. While the accuracy in the voice activity and overlap detection achieves about 80% as shown in Table V in Section VI-C, further study on this step is needed to improve the total diarization performance.

TABLE IX  
CONFUSION MATRICES ABOUT PREDICTED NUMBER OF SINGERS. THE UNITS  
ARE IN SECONDS

(a) Baseline system (Accuracy=49.0%)

0	1712	54	27	11	6	12	
1	645	551	549	254	166	198	
2	51	150	126	21	40	10	
3	41	45	277	238	22	8	
4	19	58	36	167	228		
5	87	41	164	75	105	447	
6	6			4	61	29	
	0	1	2	3	4	5	6

(b) Proposed system (Accuracy=68.5%)

0	1699	61	15	11	23	12	
1	182	1558	191	103	166	156	7
2	33	84	188	9	71	11	2
3	35	62	190	304	11	21	8
4	30	17	28	73	360		
5	103	55	152	50	92	467	
6	12		2	2	8	34	42
	0	1	2	3	4	5	6

## VIII. CONCLUSION

To analyze and extract features related to multiple simultaneous singing voices, this paper has discussed singer diarization, which is a technique to estimate who sings when from the songs with multiple unison singers. Since traditional speaker diarization methods are not suitable for singer diarization, this paper has proposed a new diarization framework. The framework is based on target-singer VAD, which recognizes whether the target singer is singing or not at each segment. While the traditional speaker diarization methods often handle overlapped segments by postprocessing, our framework performs the overlap detection before the main diarization process. This paper has also introduced Cosacorr score, a new acoustic feature to improve the performance of the overlap detection. The framework achieves singer diarization for polyphonic sound mixtures by exploiting the Spleeter singing voice separation technique. Moreover, this paper has utilized ArcFace to acquire highly discriminative singer representations. The results of the singer diarization experiments have shown that the proposed framework outperformed the baseline one that does not explicitly perform the overlap detection. The results have also shown that Cosacorr scores are effective for overlap detection.

Although the main contribution of this paper is to tackle this underexplored singer diarization task, there is still much room for improvement in the diarization error rate (DER). Future

work will include such improvement on the spectral clustering algorithm as well as the singer representations. Especially for singer representations, a more discriminative extractor can be constructed by using large speech corpora for pretraining the extractor. Another study is needed to determine whether speech corpora can be effectively utilized for training a singer representation extractor. There is also room for improvement in target-singer VAD. While target-singer VAD estimates the singing state of each singer individually, the accuracy can be improved by simultaneously estimating the singing state of all singers similarly to TS-VAD. Various applications based on singer diarization, such as a music listening interface that enables users to easily access singing voices of particular singers or unison singing voices, could also be developed in the future.

## REFERENCES

- [1] M. Goto, "Singing information processing," in *Proc. 12th Int. Conf. Signal Process.*, 2014, pp. 2431–2438.
- [2] E. J. Humphrey *et al.*, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 82–94, Jan. 2019.
- [3] A. Demetriou, A. Jansson, A. Kumar, and R. Bittner, "Vocals in music matter: The relevance of vocals in the minds of listeners," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 514–520.
- [4] S. Dixon, M. Goto, and M. Mauch, "Why singing is interesting," in *Proc. Tut. 16th Int. Soc. Music Inf. Retrieval Conf.*, 2015, p. 9.
- [5] M. Goto, T. Saitou, T. Nakano, and H. Fujihara, "Singing information processing based on singing voice modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 5506–5509.
- [6] M. Umberto, J. Bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 55–73, Nov. 2015.
- [7] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.
- [8] W.-H. Tsai and H.-M. Wang, "Automatic detection and tracking of target singer in multi-singer music recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 2004, pp. 221–224.
- [9] M. Thlithi, C. Barras, J. Pinquier, and T. Pellegrini, "Singer diarization: Application to ethnomusicological recordings," in *Proc. 5th Int. Workshop Folk Music Analysis*, 2015, pp. 124–125.
- [10] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2007, pp. 683–686.
- [11] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech*, 2019, pp. 4300–4304.
- [12] I. Medennikov *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech*, 2020, pp. 274–278.
- [13] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2013, pp. 1–9.
- [14] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, vol. 9, no. 4, pp. 357–363, Aug. 1990.
- [15] I. Ogawa and M. Morise, "Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs," *Acoustical Sci. Technol.*, vol. 42, no. 3, pp. 140–145, May 2021.
- [16] S. Watanabe *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th Int. Workshop Speech Process. Everyday Environ.*, 2020, pp. 1–7.
- [17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.
- [18] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 296–303.

- [19] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *Proc. Interspeech*, 2020, pp. 269–273.
- [20] H.-S. Heo *et al.*, "NAVER CLOVA submission to the third DIHARD challenge," 2021. [Online]. Available: [https://dihardchallenge.github.io/dihard3/system\\_descriptions/dihard3\\_system\\_description\\_team713.pdf](https://dihardchallenge.github.io/dihard3/system_descriptions/dihard3_system_description_team713.pdf)
- [21] J.-W. Jung, H.-S. Heo, Y. Kwon, J. S. Chung, and B.-J. Lee, "Three-class overlapped speech detection using a convolutional recurrent neural network," in *Proc. Interspeech*, 2021, pp. 3086–3090.
- [22] X. Xiao *et al.*, "Microsoft speaker diarization system for the VoxCeleb speaker recognition challenge 2020," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5824–5828.
- [23] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.
- [24] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using MFCC features and support vector machine," in *Proc. Int. Conf. Speech Comput.*, vol. 2, 2007, pp. 556–561.
- [25] A. Temko, D. Macho, and C. Nadeu, "Enhanced SVM training for robust speech activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 2007, pp. 1025–1028.
- [26] E. Rentzeperis, A. Stergiou, C. Boukis, A. Pnevmatikakis, and L. C. Polymenakos, "The 2006 athens information technology speech activity detection and speaker diarization systems," in *Proc. Mach. Learn. Multimodal Interact.*, 2006, pp. 385–395.
- [27] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
- [28] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7378–7382.
- [29] F. Eyben, F. W€eninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 483–487.
- [30] B. Lehner, G. Widmer, and R. Sonnleitner, "On the reduction of false positives in singing voice detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 7480–7484.
- [31] S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 121–125.
- [32] J. Schlüter and B. Lehner, "Zero-mean convolutions for level-invariant singing voice detection," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 321–326.
- [33] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, 1998, pp. 127–132.
- [34] J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, "Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, 2006, pp. 521–524.
- [35] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization," in *Proc. Interspeech*, 2006, pp. 2178–2181.
- [36] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Process. Lett.*, vol. 27, pp. 381–385, Dec. 2019.
- [37] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
- [38] J. Prazak and J. Silovsky, "Speaker diarization using PLDA-based speaker clustering," in *Proc. 6th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Syst.*, vol. 1, 2011, pp. 347–350.
- [39] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 413–417.
- [40] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4930–4934.
- [41] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5329–5333.
- [42] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker," in *Proc. Interspeech*, 2021, pp. 3555–3559.
- [43] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw.*, vol. 5, no. 50, pp. 2154–2157, Jun. 2020.
- [44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
- [45] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6738–6746.
- [46] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5265–5274.
- [47] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, "Personal VAD: Speaker-conditioned voice activity detection," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 433–439.
- [48] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research," *Speech Lang. Process. Tech. Committee Newslett.*, vol. 1, no. 4, pp. 1–32, Sep. 2013.
- [49] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, S. Renals, S. Bengio, and J. G. Fiscus, Eds., Berlin, Germany: Springer, 2006, pp. 309–322.



**Hitoshi Suda** received the Ph.D. degree in engineering from The University of Tokyo, Tokyo, Japan, in 2022. His research interests include voice conversion, speech synthesis, and speaker diarization. He is a Member of the International Speech Communication Association.



**Daisuke Saito** (Member, IEEE) received the B.E., M.S., and Dr.Eng. degrees from The University of Tokyo, Tokyo, Japan, in 2006, 2008, and 2011, respectively. He is currently an Associate Professor with the Graduate School of Engineering, The University of Tokyo. His research interests include various areas of speech engineering, including voice conversion, speech synthesis, acoustic analysis, speaker recognition, and speech recognition. From 2010 to 2011, he was a Research Fellow (DC2) of the Japan Society for the promotion of science. Dr. Saito is a Member of the International Speech Communication Association (ISCA), Acoustical Society of Japan (ASJ), Information Processing Society of Japan, Institute of Electronics, Information and Communication Engineers, and Institute of Image Information and Television Engineers. He was the recipient of the ISCA Award for the Best Student Paper of INTERSPEECH 2011, Awaya Award from the ASJ in 2012, and Itakura Award from ASJ in 2014.



**Satoru Fukayama** (Member, IEEE) received the Ph.D. degree in information science and technology from The University of Tokyo, Tokyo, Japan, in 2013. He is currently a Senior Researcher with the National Institute of Advanced Industrial Science and Technology, Japan. His research interests include signal processing and music information retrieval, especially music generation with probabilistic models. He was the recipient of awards, including IPSJ Yamashita SIG Research Award, several best presentation awards, and Specially Selected Paper Award from the Information Processing Society of Japan.



**Tomoyasu Nakano** received the Ph.D. degree in informatics from the University of Tsukuba, Tsukuba, Japan, in 2008. He is currently a Senior Researcher with the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. His research interests include singing information processing, human-computer interaction, and music information retrieval. He was the recipient of several awards, including the IPSJ Yamashita SIG Research Award from the Information Processing Society of Japan, Best Paper Award from the Sound and Music

Computing Conference 2013, and Honorable Mention Poster Award from the IEEE Pacific Visualization Symposium 2018. He is a Member of the IPSJ and Acoustical Society of Japan.



**Masataka Goto** received the Doctor of Engineering degree from Waseda University, Tokyo, Japan, in 1998. He is currently a Prime Senior Researcher with the National Institute of Advanced Industrial Science and Technology, Japan. Over the past 30 years, he has authored or coauthored more than 300 papers in refereed journals and international conferences. He was a Committee Member of more than 120 scientific societies and conferences, including the General Chair of ISMIR 2009 and 2014. As the Research Director, he began OngaACCEL Project in 2016 and

RecMus Project in 2021, which are five-year JST-funded research projects, such as ACCEL and CREST related to music technologies. He was the recipient of 57 awards, including several best paper awards, best presentation awards, the 10th Japan Academy Medal, and 10th JSPS PRIZE.