Differences Between Singer and Speaker Verification: Training Singer Feature Representation Extractor Utilizing Singing Voice Characteristics

Sayaka Toma*, Tomoki Ariga*, Yosuke Higuchi*, Ichiju Hayasaka†, Rie Shigyo† and Tetsuji Ogawa*
* Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan
† DAIICHIKOSHO CO., LTD., Tokyo, Japan
E-mail: toma@pcl.cs.waseda.ac.jp

***Abstract*—We aimed to construct a robust feature extractor for singer verification using a straightforward method that leverages the unique characteristics of singing voices. The speaker feature extractor based on ECAPA-TDNN needs to be trained to identify numerous speakers using voice data that vary within the same speaker and to distinguish voices that sound similar but belong to different speakers. However, collecting singing voice data is challenging, particularly when constructing a corpus with a large number of singing samples from the same singer. In this study, we address this problem by adopting a simple approach: segmenting singing voices before inputting them into ECAPA-TDNN during training. The validity of this approach arises not merely from increasing the amount of data, but from the inherent characteristics of singing voices. Specifically, compared to spoken voices, singing voices exhibit less acoustic variation over short periods but greater variation over long periods. By utilizing short segments of singing voices for training, we can develop a feature extractor that is robust to acoustic variations within the same singer. Our experiments demonstrate the effectiveness of the proposed approach of segmenting singing voices into three-second intervals for training and provide insights that this method does not yield the same benefits for spoken voices, highlighting its unique effectiveness for singer verification using singing voices.**

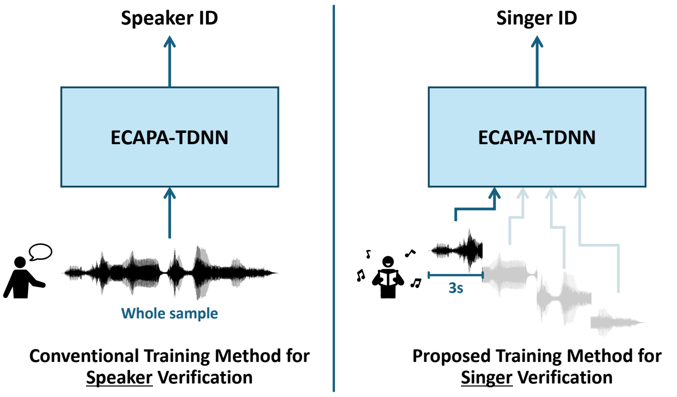
Speaker ID
ECAPA-TDNN
Whole sample
Conventional Training Method for Speaker Verification
Singer ID
ECAPA-TDNN
3s
Proposed Training Method for Singer Verification

Fig. 1. Proposed strategy for building singer feature extractor robust to acoustic variability, compared to conventional strategy for training speaker feature extractor. In our approach, singing voice is first segmented into short, stationary signal intervals (e.g., three seconds) and then used to train ECAPA-TDNN to perform singer identification using these segments.

I. Introduction

Speaker verification technology, which determines whether different voice inputs belong to the same speaker, has been extensively studied for practical applications. However, most of these studies focus on spoken speech [1]–[5]. Conversely, achieving high accuracy in singer verification using singing voices could enable the creation and viewing of singing histories without requiring user registration. Additionally, singer verification can be seen as a technology that facilitates the comparative evaluation of acoustically similar singing, allowing for the recommendation of songs that are easier for users to sing (i.e., songs frequently sung by people with similar voices). Nevertheless, in contrast to the well-developed and extensively researched speaker verification technology for spoken speech, singer verification technology for singing voices has not been sufficiently explored [6]–[8].

In general, speaker verification involves using deep neural networks (DNNs) such as ECAPA-TDNN [9] or ResNet [10], [11], which are designed to identify numerous speakers from long-duration voice inputs. The similarity of the obtained speaker feature representations is then calculated using cosine distance [3] or probabilistic linear discriminant analysis (PLDA) [12], [13]. Specifically, speaker feature representations are obtained by accumulating and statistically processing local acoustic information extracted through convolutional layers over extended periods (attentive pooling) [14]. Ideally, the speaker feature extraction process should compact the acoustic data distribution of the same speaker (reducing false rejections) and separate the acoustic data distribution of different speakers (reducing false acceptances). To accurately achieve such functionality, it is crucial to train DNNs to identify numerous speakers using voice data that vary within the same speaker and voices that sound similar but belong to different speakers. However, for singing voices, there is no well-prepared corpus that meets these requirements, and it is particularly challenging to provide a large number of singing samples from the same singer.

In contrast, this study seeks to address the problem using a straightforward approach that effectively leverages the unique characteristics of singing voices. Singing and spoken voices have significant differences in their properties. Spoken voices exhibit substantial acoustic variation over short periods due to phonetic differences, but this variation seldom changes dramatically. Conversely, singing voices show relatively less acoustic variation over short periods but display greater acoustic variation over long periods due to singing techniques and song structures. To capitalize on these characteristics, this study proposes dividing singing voices into short, steady-state signals and using these as inputs to train an ECAPA-TDNN for singer identification (as illustrated in Figure 1). Given the nature of singing voices, it is expected that although the numerous short signals are individually steady-state, they will exhibit significant variation in the embedding space produced by the ECAPA-TDNN. Consequently, training the singer feature extractor in this manner is anticipated to yield robust singer feature representations that are resilient to acoustic variations within the same singer and sensitive to acoustically similar inputs from different singers.

By comparing the results of training ECAPA-TDNN using entire songs versus segmented singing voices, we aim to clarify the effectiveness of the proposed approach. Additionally, we will conduct experiments by segmenting spoken voice inputs and feeding them into ECAPA-TDNN to demonstrate that the proposed approach specifically addresses issues unique to singing voice verification. The contributions of this study are as follows:

- Provide insights into the differences between singing and spoken voices.
- Present a simple approach to significantly reduce false rejections in singer verification.
- Demonstrate the effectiveness of the proposed approach through experiments using a large-scale karaoke singing voice dataset and provide insights into the differences from speaker verification using spoken voices.

The findings from this study are expected to be valuable for researchers and engineers who wish to develop singer verification systems more easily.

The remainder of this paper is organized as follows. Section II explores the acoustic differences between spoken and singing voices. Section III details the singer verification system employed in this study. Section IV presents the singer verification experiments conducted using karaoke singing voices and evaluates the effectiveness of training the feature extractor with segmented voice inputs. Finally, Section V provides this study's conclusion.

II. Differences in Acoustic Variations between Spoken and Singing Voice

In general, singing voices exhibit less acoustic variation over short periods compared to spoken voices. Figure 2 shows histograms of acoustic variation within short segments of spoken and singing voices. The acoustic variation for short segments was measured by calculating the variance of mel-frequency cepstral coefficients (MFCCs) using a 25-millisecond frame size and a 10-millisecond frame shift for each three-second segment. These variance values were then averaged for each speaker or singer, with the analysis performed on data from 100 individuals. The histograms indicate that acoustic variation within short segments is smaller for singing voices than for spoken voices.

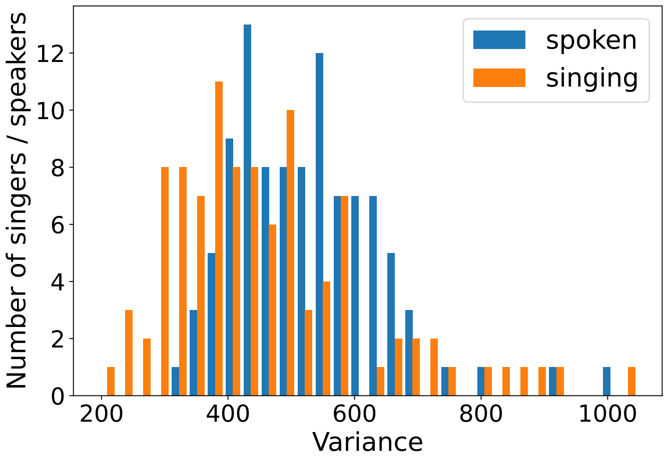

Fig. 2. Histogram of intra-singer and intra-speaker acoustic variations over short periods for singing and spoken voices.

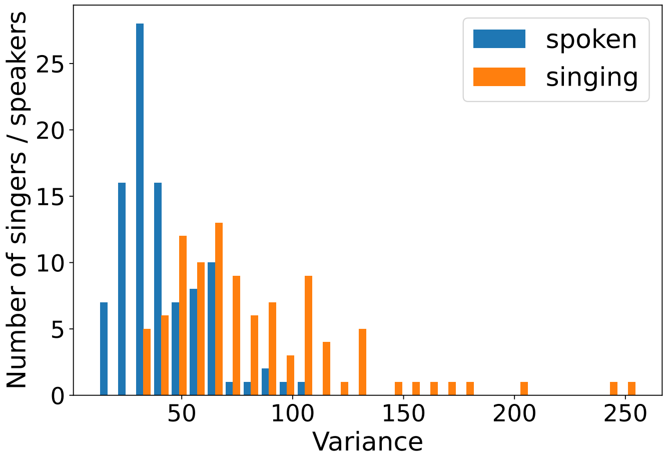

Fig. 3. Histogram of intra-singer and intra-speaker acoustic variations over long periods for singing and spoken voices.

Conversely, due to significant changes in melody and pitch throughout different parts of a song, acoustic variation over long periods is greater for singing voices compared to spoken voices. Figure 3 represents histograms of acoustic variation over long segments, which were computed as the variance within each speaker of the mean MFCCs calculated for each three-second frame. This graph demonstrates that acoustic variation over long periods is greater for singing voices than for spoken voices.

TABLE I
STRUCTURE OF ECAPA-TDNN.

Index	DNN Module	Input Dim.	Output Dim.
1	TDNNBlock	$80 \times T$	$1024 \times T$
2	SE-Res2Block	$1024 \times T$	$1024 \times T$
3	SE-Res2Block	$1024 \times T$	$1024 \times T$
4	SE-Res2Block	$1024 \times T$	$1024 \times T$
5	TDNNBlock	$3072 \times T$	$3072 \times T$
6	AttentiveStatisticsPooling	$3072 \times T$	256
7	Conv1d	256	192

III. SINGER VERIFICATION

A. ECAPA-TDNN

We focus on utilizing a widely adopted speaker verification system for singer verification, leveraging ECAPA-TDNN [9] for embedding singer information from singing voice data and employing cosine similarity scoring for verifying singers [3].

The ECAPA-TDNN-based feature extractor, configured with the parameters detailed in Table I, is first trained using our singing voice database (see Section IV-A). The intermediate layer outputs are then extracted as singer embeddings, which are subsequently used for cosine similarity scoring. ECAPA-TDNN incorporates attentive pooling and Squeeze-and-Excitation (SE) blocks [15] to enhance the learning of relevant features within the model. Additionally, by combining the outputs not only from the preceding layer but also from shallower layers during the pooling process, ECAPA-TDNN is expected to capture more robust singer embeddings.

Increasing the number of identification classes in this network is expected to capture finer distinctions among speakers, potentially reducing the false acceptance rate. However, if the data available for each singer is insufficient, it may hinder the model's generalization, potentially leading to an increased false rejection rate.

B. Proposed Training Strategy for Singer Feature Extractor

In speaker verification tasks involving spoken voice, shortening the input length for ECAPA-TDNN is generally not recommended. Spoken voice exhibits significant acoustic variations over short periods, as phonemes change within a few hundred milliseconds, but the overall acoustic variation across an entire utterance is relatively small. In contrast, this study on singer verification seeks to train the ECAPA-TDNN using singing voice samples segmented into short durations. As discussed in Section II, singing voice typically shows minimal acoustic variation over short periods but significant variation over the duration of an entire song. Because of this characteristic, short segments of the same song may display considerable variability in the embedding space. By intentionally increasing the acoustic variability within the same singer's samples during training, we aim to create an optimal feature embedding for singer verification: it suppresses intra-singer acoustic variations while effectively distinguishing between different singers.

In this approach, whether the entire song or segmented portions are used as input, the total amount of training data remains the same. However, capturing the precise acoustic variations across an entire song in the averaged embedding representation after TDNN pooling is likely to be limited. Segmenting the singing voice offers a simple and effective approach to overcoming this limitation.

IV. SINGER VERIFICATION EXPERIMENTS

To evaluate the effectiveness of the proposed approach, we conducted singer verification experiments using a custom-built singing voice corpus. In the first experiment, we compared the verification performance of ECAPA-TDNN when trained with segmented singing voice inputs versus entire song inputs. In the second experiment, to elucidate the differences between singing and spoken voices, we compared the verification performance for spoken voices using segmented versus unsegmented inputs during the training of ECAPA-TDNN.

A. Database

To conduct a singer verification experiment, we compiled a comprehensive database of Japanese singing voices. This dataset includes 4950 songs performed by 1666 amateur singers in karaoke settings. Of the 1666 participants, 1000 singers performed only one song, 225 sang two songs, and 136 sang three songs. The recordings were made using directional microphones installed in karaoke booths, capturing the singing voices with minor contamination from background music, other voices, and reverberation.

B. Experimental Setups

The singing voices used in the experiment were sampled at 16 kHz and quantized into 16-bit data. To mitigate the impact of noise, particularly from background music, we performed data augmentation for training singer embeddings using rirs-noises [16] and MUSAN [17], which include human voices, environmental sounds, and music as noise sources. The training of the ECAPA-TDNN model [9] utilized our singing voice dataset, which comprised 3500 songs performed by 441 singers (each singing three or more songs).

For the enrollment and verification phases, we used singing voice data from 225 singers, with each singer contributing two songs—one for enrollment and the other for verification. The total number of trials was determined by the combinations of enrollment and verification data for the 225 singers, resulting in 225×225 trials. Notably, the enrollment and verification data were sampled from different songs. We computed similarity scores for all possible pairs of data, and the evaluation was based on the Equal Error Rate (EER) and the minimum Decision Cost Function (minDCF). Both the training and inference processes were conducted using the SpeechBrain toolkit [18].

C. Experiment 1

1) Experimental Procedure: To assess the effectiveness of dividing singing voice data into segments of N seconds for training a singer feature extractor, we compared the verification performance for segment lengths of 3 seconds, 10 seconds, and

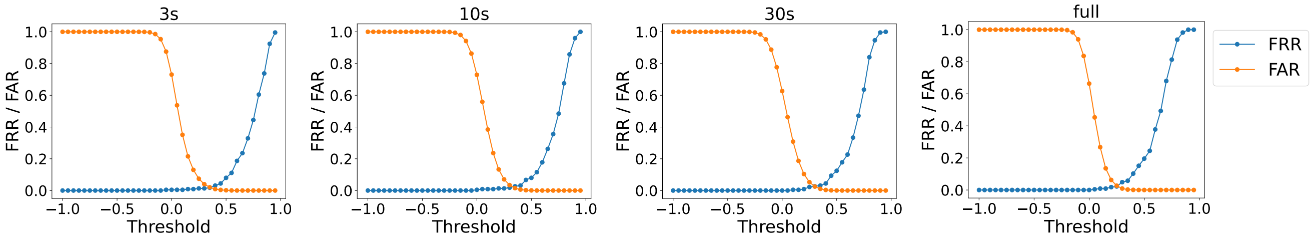

Fig. 4. False rejection rate (FRR) and false acceptance rate (FAR) as function of threshold for cosine similarity.

TABLE II
EQUAL ERROR RATE FOR VARIOUS LENGTHS OF TRAINING AND VERIFICATION DATA. ENTIRE SONG (FULL) WAS USED FOR ENROLLMENT DATA.

Length of Training Data	**Length of Verification Data**		
	full	30s	10s
full	1.72	2.92	9.33
30s	2.22	2.85	7.53
10s	1.78	**2.67**	4.44
3s	**1.33**	2.76	**3.97**

TABLE III
NUMBER OF FALSE REJECTIONS AND MINDCF. C_{MISS} AND C_{FA} REPRESENT COSTS ASSOCIATED WITH MISSED DETECTION ERRORS AND FALSE ALARMS, RESPECTIVELY.

Length of Training Data	**Number of False Rejections**	**minDCF** ($C_{\text{miss}} = 2$, $C_{\text{fa}} = 1$)
full	29	0.15
30s	21	0.13
10s	22	**0.11**
3s	23	0.14

30 seconds with the performance obtained without segmentation. We ensured that the sample count of the training data remained constant regardless of N by overlapping the segmentation windows. Additionally, to examine the effects of varying input audio lengths between enrollment and verification data, we evaluated verification performance using the entire song, the first 10 seconds, and the first 30 seconds for the verification audio, while the entire song was used for the enrollment audio.

2) Experimental Results: Table II presents the verification performance for various lengths of training and verification data. The results demonstrate that training the feature extractor with short, segmented singing voice data (three seconds) directly enhances verification performance, regardless of the input length of the verification data. Additionally, using an entire song for both enrollment and verification data yielded the best performance, with an EER of 1.33. These findings suggest that while using longer-duration data for enrollment and verification enhances the reliability of the verification process, a feature extractor trained on short, segmented audio can effectively capture precise (capable of distinguishing between similar but different singers) and robust (suppressing acoustic variations of the same singer) singer features.

TABLE IV
EER FOR TRAINING DATA LENGTH IN SPEAKER VERIFICATION.

Length of Training Data	**EER**
full	1.48
3s	**1.41**

When applying singer verification to build and reference individual singing histories in karaoke, minimizing false rejections is essential. In particular, ECAPA-TDNN trained with short, segmented singing voice data effectively reduced false rejections. Figure 4 illustrates the false rejection rate (FRR) and false acceptance rate (FAR) as the threshold for cosine similarity in verification is varied in increments of 0.05 for each input length of the singing voice data used in training (entire song, 3 seconds, 10 seconds, 30 seconds). The results show that the integral of the FRR curve is smallest when the training data input length is 3 seconds, and it increases progressively as the input length extends to 10 seconds, 30 seconds, and an entire song.

Table III presents the minDCF values, adjusted to prioritize the reduction of false rejections (with the false rejection rate weighted twice as heavily as the false acceptance rate), along with the corresponding number of false rejections at the adjusted threshold settings. The results show a decrease in false rejections when using segmented inputs of 30 seconds, 10 seconds, and 3 seconds, compared to using the full-length song, with 10-second segments achieving the best minDCF performance. Although Table II indicates that the EER worsens with 30-second and 10-second segments compared to full-length inputs, it is important to recognize that these segmentations help reduce false rejections. This suggests that the observed deterioration in EER is likely due to an increase in false acceptances.

D. Experiment 2

1) Experimental Procedure: To evaluate whether similar effects can be achieved with spoken voice data, we conducted an experiment similar to Experiment 1, where we segmented spoken voice data and input it into the ECAPA-TDNN model. Using audio segments longer than six seconds from the VoxCeleb dataset, we compared the performance of training with three-second segments to that of training with full-length audio.

2) Experimental Results: The results are presented in Table IV. For spoken voice, the verification performance remained consistent regardless of whether the audio was segmented. Due to the relatively stable acoustic characteristics of spoken voice over long durations, segmenting the audio and inputting it into the ECAPA-TDNN did not enhance the extraction of robust speaker features against acoustic variability, and thus did not impact verification performance. These findings suggest that the observed improvement in verification performance from training with short segments is specific to singing voice data, which exhibits substantial acoustic variation over long durations.

V. Conclusion

In this study, we proposed a straightforward method for extracting robust feature representations for singer verification by leveraging the distinct characteristics of singing voice, which shows minimal acoustic variation over short periods but significant variation over longer durations compared to spoken voice. We demonstrated that segmenting the input into shorter segments for training the feature extractor can effectively reduce false rejections in singer verification. Furthermore, we established that this effect is unique to singing voice and does not extend to spoken voice.

References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[4] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Proceedings of Interspeech*, vol. 2017, 2017, pp. 999–1003.

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[6] L. Regnier and G. Peeters, "Singer verification: Singer model .vs. song model," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 437–440.

[7] L. Wang, B. Wang, G. Tan, *et al.*, "ATRemix: An auto-tune remix dataset for singer recognition," in *Proceedings of Chinese Conference on Biometric Recognition*, 2022, pp. 348–355.

[8] H. Yakura, K. Watanabe, and M. Goto, "Self-supervised contrastive learning for singing voices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1614–1623, 2022.

[9] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proceedings of Interspeech*, 2020, pp. 3830–3834.

[10] C. Li, X. Ma, B. Jiang, *et al.*, "Deep speaker: An end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[11] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 301–307.

[12] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV European Conference on Computer Vision*, 2006, pp. 531–542.

[13] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE international conference on computer vision*, 2007, pp. 1–8.

[14] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proceedings of Interspeech*, 2018, pp. 3573–3577.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[16] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[17] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[18] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.