INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Self-supervised learning with diffusion-based multichannel speech enhancement for speaker verification under noisy conditions

Sandipana Dowerah[*], *Ajinkya Kulkarni*[†], *Romain Serizel*[*], *Denis Jouvet*[*]

[*]Université de Lorraine, CNRS, Inria, Loria, F-54000, Nancy, France, [†]MBZUAI University, UAE

sandipana.dowerah@loria.fr, ajinkya.kulkarni@mbzuai.ac.ae, romain.serizel@loria.fr, denis.jouvet@inria.fr

Abstract

The paper introduces Diff-Filter, a multichannel speech enhancement approach based on the diffusion probabilistic model, for improving speaker verification performance under noisy and reverberant conditions. It also presents a new two-step training procedure that takes the benefit of self-supervised learning. In the first stage, the Diff-Filter is trained by conducting time-domain speech filtering using a scoring-based diffusion model. In the second stage, the Diff-Filter is jointly optimized with a pre-trained ECAPA-TDNN speaker verification model under a self-supervised learning framework. We present a novel loss based on equal error rate. This loss is used to conduct self-supervised learning on a dataset that is not labelled in terms of speakers. The proposed approach is evaluated on MultiSV, a multichannel speaker verification dataset, and shows significant improvements in performance under noisy multichannel conditions.

Index Terms: multichannel speech enhancement, diffusion probabilistic models, speaker verification, self-supervised learning

1. Introduction

Speaker verification (SV) aims to confirm the identity of a person based on his/her voice characteristics. SV has achieved significant performance gain in controlled or close-talk scenarios. However, it suffers from unsatisfactory performance in multichannel far-field scenarios. This is due to complex environmental settings as speech signals propagating in the long-range are subject to fading, absorption, room reverberation and complex environmental noises, which change the pressure level at different frequencies and degrade the signal quality. Speech enhancement (SE) can be used as a pre-processing to SV in noisy reverberant scenarios. Speech enhancement aims to enhance the quality and intelligibility of speech signals that are corrupted by noise and/or reverberation by estimating the original clean speech signal using various signal processing techniques. Multichannel speech enhancement aims to enhance distorted speech using multiple microphones and improve performance by taking advantage of the additional spatial information provided by these microphones compared to single-channel.

Generative models aim to learn the fundamental characteristics of speech, such as its spectral and temporal structure and can use this prior knowledge to identify clean speech from noisy or reverberant input signals that fall outside the learned distribution. [1, 2] used the raw waveform, or magnitude spectrum, as input for generative model-based speech enhancement. Generative adversarial networks (GAN) [3, 4], variational autoencoders (VAE) [5–7], and flow-based models [8] have been used to estimate the distribution of clean speech signals. Recently, diffusion-based models have also been studied for speech enhancement [9–11]. All these approaches share the concept of gradually converting input data into noise and training a neural network to invert this process for various noise scales based on the Markov chain.

DiffuSE [9] was proposed to recover the clean speech signal from the noisy signal based on Markov chains; it provides a framework for denoising diffusion probabilistic models. Lu et al. formulated the CDiffSE model using a generalized conditional diffusion probabilistic model that incorporates the observed noisy data into the model [10]. While CDiffSE and DiffSE employ U-net as their diffusion decoder network, our proposed work takes a different approach and uses Conv-TasNet as the diffusion decoder instead. Specifically, our method conducts speech enhancement on the time-domain representation of the signal. Zhang et al. extend the Diff-Wave vocoder [11] using a convolutional conditioner for denoising, and it is trained separately using a L1 loss for matching latent representations [12].Our proposed approach incorporates a conditioning network based on Conv-TasNet in addition to the diffusion decoder. This conditioning network provides an estimate of the clean and noisy signals, which are combined with the multichannel noisy signal and fed into the diffusion decoder. By doing so, the diffusion process is made easier as it can learn to remove the noise while taking into account the clean speech estimate provided by the conditioning network. Recently, some studies [13–15] have explored scoring-based diffusion models with stochastic differential equations (SDE) instead of Markov chains. SDE enables the controlling of selecting the reverse diffusion steps for enhancement [16]. The aforementioned works use only single-channel and have not been studied for SV.

Self-supervised learning is a powerful machine learning technique that enables models to learn from unlabeled data by leveraging the inherent structure or patterns in the data itself without the need for explicit supervision from labelled data. In the context of speaker verification tasks, few approaches have conducted contrastive learning for self-supervised learning [17–20]. The loss function design for SV mainly focuses on speaker classification loss function, and verification loss [21]. Furthermore, the contrastive learning framework enables the online creation of verification labels. In order to exploit the multichannel speech data without explicit speaker labels, we propose to use the equal error rate (EER) evaluation metric as a loss function to optimize the speaker embedding representation on the verification task.

In this paper, we present a diffusion probabilistic model (DPM)-based two-stage multichannel speech enhancement approach as a pre-processing to SV. We named our approach Diff-Filter as it mimics the behaviour of Rank-1 multichannel Wiener filter (MWF). In the first stage, we train the Diff-Filter

10.21437/Interspeech.2023-1890

by conducting time-domain speech filtering using a scoring-based diffusion model. In the second stage of training, we jointly optimize the Diff-Filter with a pre-trained ECAPA-TDNN SV model under a self-supervised learning framework. We evaluate our results on MultiSV, a multichannel SV dataset, and show that our proposed approach significantly improves SV performance under multichannel noisy conditions.

2. Proposed Approach

In this section, we present the proposed approach for developing a robust multichannel SV system in a noisy environment. In the first phase, we trained the ECAPA-TDNN [22] based SV system and a multichannel speech enhancement system separately. We used this pre-training of speech enhancement and SV for training the jointly optimized system using self-supervised learning. We jointly optimized Diff-Filter and ECAPA-TDNN with the proposed EER loss as a verification loss to optimize the binary classification with speaker embedding representations. Diff-Filter is a scoring-based diffusion probabilistic model where Conv-TasNet architecture is utilized for conducting the diffusion process. Diff-Filter is trained to provide Rank-1 MWF clean speech signal for a given multichannel noisy input signal. We used a conditioning network to provide the estimates of clean and noise signals as additional input to the diffusion decoder, thus conditioning the sampling process from terminal distribution aware of noise to be removed from the noisy multichannel signal.

2.1. Diff-Filter

This section presents a novel way to train a multichannel speech enhancement system as a DPM-based filtering method named Diff-Filter. We termed the proposed system Diff-Filter, as it replicates the functionality of the Rank-1 MWF filter to provide a clean speech signal. The proposed Diff-Filter comprises a diffusion-based decoder network and a conditioning network, as shown in Figure 1. We used Conv-TasNet [23] as an external conditioning network. The conditioning network is used to compute the estimates of the clean speech signal, s, and noise in time-domain representation, N. We provide conditioning network output estimates along with the multichannel noisy speech signal as input to the Diff-Filter system. In the forward and reverse diffusion process, terminal noise distribution is defined as $\mathcal{N}(\mu, I)$, where the mean is μ, and I is unit variance. We parameterized the mean μ of terminal noise distribution of the diffusion process using noisy multichannel input, y.

Similar to [13], we incorporated scoring-based diffusion probabilistic model, in which the diffusion decoder learns the trajectories of forward diffusion in reverse time order. In the training phase, the forward diffusion process is conducted by iteratively deconstructing Rank-1 MWF clean speech estimate signal to the terminal distribution defined by noisy multichannel signal. Furthermore, terminal distribution is also conditioned with estimates of the clean speech signal and noise signal provided by the conditioning network. The usage of clean speech estimate and noise estimate in the diffusion process assists in conducting noise-aware speech enhancement. We used stochastic differential equations (SDE) to learn the gradients of the forward diffusion process as shown in Figure 1, where $s_\theta(X_t, \mu, s, N, t))$ denotes the diffusion decoder network with t as the diffusion time step and β as noise scheduler. During the inference phase solving the SDE describing dynamics of the reverse diffusion with a simple first-order Euler-Maruyama

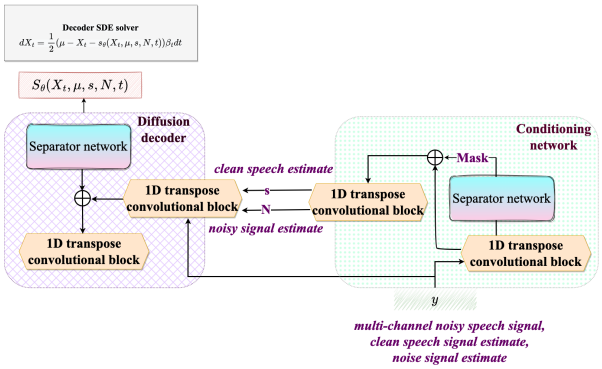

Figure 1: *Architecture of Diff-Filter consists of a diffusion decoder and a conditioning network. The clean speech signal and noise signal estimated from the conditioning network are channel-wise concentrated with noisy multichannel signal along with Rank-1 MWF clean speech signal in the diffusion decoder.*

scheme [16, 24].

We trained the Diff-Filter using a two-stage training process. First, we conditioned the diffusion encoder with a target clean speech signal, a target noise signal, and a noisy multichannel signal. The main purpose of pre-training the diffusion decoder is to ensure that diffusion model parameters converge in optimal minima direction using target clean speech and noise. In the second stage of training, we used the clean speech and noise estimated by the conditioning network.

In the inference phase, multichannel noisy signals and estimates of speech and noise signals obtained from the conditioning network are provided to the diffusion decoder to estimate the reverse trajectories of forward diffusion. The reverse diffusion process iteratively reconstructs the Rank-1 MWF filter output of clean speech by sampling latent variables from conditional terminal distribution.

2.2. Self-supervised learning for multichannel SV

We jointly optimized the Diff-Filter and ECAPA-TDNN as a multichannel SV. In joint optimization, the multichannel noisy signal is first given to Diff-Filter, which provides a single-channel Rank-1-MWF filtered clean speech signal. The error gradient passes through ECAPA-TDNN and Diff-Filter as a single unit because we back-propagate through both models that we are jointly training. To conduct the self-supervised learning with the unlabelled dataset, we created an RIR simulated dataset and applied it to clean speech from the LibriSpeech dataset [25] detailed in section 3.

As shown in Figure 2, utterances 1 and 2 are given to the jointly optimized network composed of Diff-Filter and ECAPA-TDNN. The utterances 1 and 2 are multichannel noisy signals, where utterance 2 is either the data-augmented multichannel noisy signal from utterance 1 or a randomly selected multichannel noisy signal from a different speaker. During the training, we used the self-supervised contrastive learning framework, where both utterances, utterances 1 and 2, are given to the same jointly optimized network. Verification labels are generated as 0 or 1 if utterance 2 is data-augmented from utterance 2 or not, where 1 represents that utterances 1 and 2 are from the same speaker, 0 otherwise. For data augmentation in self-supervised learning, we used speed perturbation by 0.9, 1.1 factor only and masking the 1 sec part of the noisy multichannel signal.

We propose to use an EER as a loss function to train the

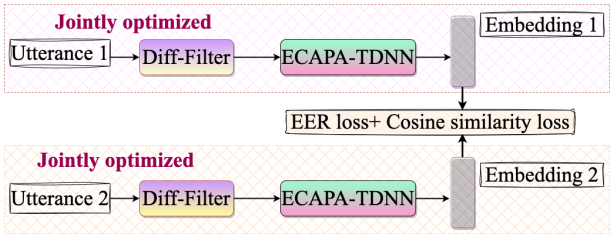

Figure 2: *Framework of self-supervised learning, where* $utterance_1$ *and* $utterance_2$ *are noisy multichannel signals given to the jointly optimized network of ECAPA-TDNN to obtain the speaker embeddings.*

jointly optimized network. The EER is the location on a receiver operating characteristic curve where the false acceptance rate and false rejection rate are equal. First, we computed the cosine similarity distance between $embedding_1$ and $embedding_2$ for a given batch. Then, false acceptance rate (FAR) and false rejection rate (FRR) are estimated based on cosine scores and verification labels using torchmetrics[1]. We estimated EER for the given batch size from FAR and FRR as stated in Equation 1, where $\mathcal{L}_{EER}$ ranges from value 0 to 1.

$$\mathcal{L}_{EER} = FAR\Big[argmin|FRR - FAR|\Big] \tag{1}$$

We also estimated cosine similarity loss between embeddings: $embedding_1$ and $embedding_2$ [26] as shown below in Equation 2.

$$\mathcal{L}_{cosine} = \begin{cases} 1 - cos(emb_1, emb_2) & label = 1 \\ \max(0, \cos(\text{emb}_1, emb_2) - M) & else \end{cases} \tag{2}$$

where emb_1 and emb_2 refers to embeddings extracted on utt_1 and utt_2 respectively, M refers to regularizer of value 0.2 and cos refers to cosine angle between emb_1 and emb_2.

3. Dataset Preparation

We used various datasets at different stages while developing the proposed approach for multichannel SV in noisy conditions. We used the MultiSV dataset [27] for training the Diff-Filter, which consists of 4 channel speech utterances room simulated impulse response with background noises from Music, MUSAN, and freesound.org[2]. The training dataset of MultiSV is simulated using the VoxCeleb2 dataset [28]. Consistent with the Diff-Filter training data source, we utilized the VoxCeleb2 dataset with standard Kaldi-based data augmentation techniques for training ECAPA-TDNN single-channel SV. We opted for the VoxCeleb2 dataset for joint training as MultiSV is a labelled dataset, and the core of self-supervised learning is to explore the unlabelled dataset.

To jointly optimized the network, we first simulated a room impulse dataset and applied it to the clean speech from the LibriSpeech dataset without taking into account the speaker information, thus creating an unlabelled multichannel SV dataset. The pyroomacoustic toolbox[3] is used for room simulation with 4 channels. The room length was drawn randomly between $[3, 8]$ m, the width was chosen between $[3, 5]$ m, and the height was chosen between $[2, 3]$ m. The absorption coefficient was drawn randomly such that the room's RT60 was between $[200, 600]$ ms. The minimum distance between a source and the wall is 1.5 m and 1 m between the wall and the microphones. We generated a total of 50000 training samples for self-supervised learning.

To evaluate the proposed work, we used two multichannel trial protocols from the MultiSV dataset, namely MRE and MRE hard trial protocols. The evaluation set of MultiSV is retransmitted development set derived from the VOiCES dataset. In addition to MultiSV evaluation data, we also created an internal evaluation set using Fabiole corpus [29], a French speech corpus consisting of around 6882 audio files from 130 native French speakers. The speech data of Fabiole has been collected from different French radio and TV shows. For creating each evaluation set, we have used 1200 speech files from Fabiole representing 2 hrs of evaluation material. We used the same configuration for room impulse response simulation as used for creating the training dataset for the self-supervised learning phase. We designed the evaluation set with various RIR scenarios to be used for both speech enhancement and SV.

4. Experimentation set-up

4.1. Multichannel speech enhancement

The model is trained using two loss functions, diffusion loss and scale invariant signal to distortion (SI-SDR) loss [30]. The diffusion loss is defined by Fisher divergence as a way to compute the scoring function, which is the gradient of change in log probability density in each diffusion step [31]. The second loss function, SI-SDR loss, is applied to the output of the conditioning network to ensure that the diffusion model ingrains the intrinsic information about clean speech estimate and noise estimate in time-domain representation. In training, we provided speech segments of a fixed length of 4 seconds of duration.

We set the initial weight of 0.001 on SI-SDR loss. Then, we increased the initial weight by 0.0001 after every 5 epoch till it reached 1. For the two-stage training approach, first, we trained the network for 100 epochs with a learning rate of 1e-2 and reduced the learning rate over the epochs with a factor of 0.85 after every 5 epoch. We used Adam optimizer for two-stage training with a batch size of 2. In the second stage of training, the system is trained with a learning rate of $1e-4$ for 500 epochs.

We used Conv-TasNet architecture to develop both diffusion decoder and conditioning network, with modification of replacing PReLU activation function with GeLU [32]. The implementation of networks using Conv-TasNet includes 512 filters in the convolutional block and transpose convolutional block (N), 20 lengths of filters (L), 256 channels in a bottleneck, and the residual paths 1×1 convolutional blocks. Each convolutional block's kernel size (P) is set to 3, and the number of convolutional blocks in each repeat is 8. Also, we adopted global layer normalization with a non-causal strategy for Diff-Filter implementation. To ensure a stable learning process, we used gradient clipping with a maximum L2-norm of 5.

We conducted self-supervised training on the proposed approach in a contrastive learning framework for 50k iterations with a batch size of 4. In each batch of self-supervised training, we kept equal distribution of verification labels as 0 and 1. We used Adam optimizer with a learning rate of $1e-3$ with weight decay of $1e-4$ for every 1000 iteration.

[1] https://torchmetrics.readthedocs.io
[2] https://freesound.org/
[3] https://github.com/LCAV/pyroomacoustics

Table 1: *Evaluation of proposed approach on MultiSV dataset for MRE and MRE hard as multichannel trial protocol, where J. op. refers to a jointly optimized system, and SSL refers to the system trained using self-supervised learning.*

	SE	SV	MRE	MRE hard
	Mask [27]	Resnet	3.91	5.37
	ConvTasNet [27]	Resnet	3.71	4.61
	Unprocessed	ECAPA-TDNN	5.84	10.27
	Oracle Rank1-MWF	ECAPA-TDNN	1.64	3.12
	ConvTasNet	ECAPA-TDNN	3.73	4.52
	Diff-Filter	ECAPA-TDNN	3.57	4.36
J. op.	Diff-Filter	ECAPA-TDNN	3.24	4.26
J. op.	Diff-Filter	ECAPA-TDNN (SSL)	**3.07**	**3.19**

Table 2: *Evaluation of proposed approach on room impulse simulated data on Fabiole dataset, where we used ECAPA-TDNN as SV system and J. op. refer to the jointly optimized system, and SSL refers to the system trained using self-supervised learning.*

	SE System	EER	SIR	SDR
	Unprocessed	9.23	15.11	2.01
	Oracle Rank-1 MWF	5.91	24.73	7.24
	ConvTasnet	7.87	23.21	6.12
	Diff-Filter	7.83	23.78	6.69
J. op.	Diff-Filter	7.54	24.11	6.93
J. op.	Diff-Filter (SSL)	**6.27**	**24.37**	**7.02**

4.2. Speaker verification

We used ECAPA-TDNN as a single-channel SV system from [22]. We used the VoxCeleb2 dev dataset for training ECAPA-TDNN. As SV systems often benefit from data augmentation, we used a combination of different data-augmentation techniques, such as Kaldi recipes of data-augmentation (using MUSAN [33] and room impulse response dataset[4]) and speed perturbation by changing the tempo of speech.

Besides squeeze and excitation block, the attention module of ECAPA-TDNN is set to 128. The scale dimension in Res2Block is set to 8. We extracted 256 dimension speaker embedding from the ECAPA-TDNN network. Initially, we trained the ECAPA-TDNN network with a cyclic learning rate varying between $1e-8$ and $1e-3$ using the triangular policy with Adam optimizer. The ECAPA-TDNN network is trained with angular margin softmax with a margin of 0.3 and softmax pre-scaling of 30, 100k iterations. We provided the Mel spectrogram as an input to ECAPA-TDNN. We extracted 40-dimensional Mel spectrogram features using the torchaudio library with a window length of 400 samples, hop size of 160, and 512 FFT length/ Mel spectrogram features of 40 dimensions as input to the ECAPA-TDNN network. We used a cosine scoring system for verification purposes from extracted embedding.

5. Results and Discussion

We compared the performance of the proposed approach with Conv-TasNet as baseline multichannel speech enhancement used as a front end to the ECAPA-TDNN system. For establishing baseline Conv-TasNet, we trained under the same training data used by the Diff-Filter system. Also, we used the same network configuration for Conv-TasNet as for the conditioning network of Diff-Filter. In addition to this, we also computed performance with oracle Rank-1 MWF in order to analyze the filtering approach based on the diffusion probabilistic model. We used EER as an evaluation metric to evaluate the multichannel SV systems on MRE and MRE hard trials from the MultiSV dataset. We compute signal-to-inference ratio (SIR), signal-to-distortion ratio (SDR), and EER on a Fabiole-based multichannel evaluation set. We used MIR eval tool[5] to compute the SIR and SDR metrics. The usage of SIR and SDR metrics provides insight into the performance of the multichannel speech enhancement system as a front end to the SV system.

In Table 1, the Diff-Filter front-end outperforms the Conv-TasNet without additional post-training using joint optimization or self-supervised learning. We observed that the proposed approach showed better results on both trials MRE and MRE hard compared to baseline results presented in [27], where the Resnet-based SV system was used. We obtained the best results on the proposed approach trained under a self-supervised learning framework, which shows an efficient generalization of speaker representation under noisy conditions using an unlabelled speaker dataset. In the case of the MRE hard protocol, it has performance close to the multichannel speech enhancement baseline using Oracle Rank-1-MWF. On the other hand, the performance of the proposed approach had a significant margin in performance difference with oracle Rank-1 MWF. Table 2 illustrates consistent performance improvement by the proposed approach on both trials sets on the Fabiole-based evaluation set. SDR and SIR seem to be closely co-related with EER. With a SIR of 24.37, the proposed joint optimized approach with self-supervised learning achieves the best performance among all the speech enhancement systems. Similarly, with an SDR of 7.02, the proposed joint optimization approach with self-supervised learning achieves the best performance among all the systems evaluated. SIR and SDR

The proposed approach shows consistent performance on both SV and multichannel speech enhancement tasks. The usage of self-supervised learning eases the network optimization for generalization from the unlabelled distribution. As one of the primary evaluation metrics for the SV task is EER, the adaptation of EER loss without speaker labels in self-supervised training elevates the intraclass speaker representation while increasing the interclass speaker representation. The usage of a conditioning network allowed the diffusion process allowed to perform a noise-aware reverse diffusion process. The usage of Conv-TasNet as a diffusion decoder enabled to perform the stepwise noise removal on time-domain signal representation, thus inherently considering the phase information.

6. Conclusion

In this work, we proposed Diff-Filter, a multichannel speech enhancement approach as a front end to SV. We improved the performance of the proposed Diff-Filter by jointly optimizing it with ECAPA-TDNN-based SV and further training under self-supervised contrastive learning. We presented EER loss in self-supervised learning to exploit the unlabelled speaker dataset. The obtained results have shown significant improvement in performance on the MultiSV dataset compared to state-of-the-art systems. In order to measure speech enhancement performance, we used SIR and SDR evaluation metrics. The results computed on the simulated evaluation set (derived from Fabiole) showed results in-line with performance on the MultiSV evaluation set. In future, we will conduct further experimentation with Diff-Filter to observe the efficiency of different tasks such as source separation, speaker diarization etc.

[4]https://www.openslr.org/28/

[5]https://craffel.github.io/mir_eval/

7. References

[1] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Interspeech*, 2017.

[2] D. Michelsanti and Z. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Interspeech*, 2017.

[3] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5039–5043, 2018.

[4] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," *ArXiv*, vol. abs/1905.04874, 2019.

[5] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 716–720, 2017.

[6] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2018.

[7] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder," in *Interspeech*, 2020.

[8] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1104–1117, 2020.

[9] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 659–666, 2021.

[10] Y.-J. Lu, Z. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7402–7406, 2022.

[11] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *ArXiv*, vol. abs/2009.09761, 2020.

[12] J. Zhang, S. Jayasuriya, and V. Berisha, "Restoring degraded speech via a modified diffusion model," in *Interspeech*, 2021.

[13] S. Dowerah, R. Serizel, D. Jouvet, M. Mohammadamini, and D. Matrouf, "Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification," in *IEEE SLT 2022*, 2023.

[14] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[15] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," *arXiv preprint arXiv:2203.17004*, 2022.

[16] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," in *NeurIPS*, 2021.

[17] J. Kang, J. Huh, H.-S. Heo, and J. S. Chung, "Augmentation adversarial training for self-supervised speaker representation learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1253–1262, 2022.

[18] B. Han, Z. Chen, and Y. Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," in *Interspeech*, 2022.

[19] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6723–6727, 2020.

[20] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6829–6833, 2020.

[21] V. Mingote, A. Miguel, A. O. Giménez, and E. L. SOLANO, "Log-likelihood-ratio cost function as objective loss for speaker verification systems," in *Interspeech*, 2021.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech*, 2020.

[23] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, 2018.

[24] P. E. Kloeden and E. Platen, "The numerical solution of stochastic differential equations," *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, vol. 20, pp. 8 – 12, 1977.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *ICASSP*, 2015.

[26] S. Novoselov, V. Shchemelinin, A. Shulipa, A. Kozlov, and I. Kremnev, "Triplet loss based cosine similarity metric learning for text-independent speaker recognition," in *Interspeech*, 2018.

[27] L. Moner, O. Plchot, L. Burget, and J. H. ernocký, "Multisv: Dataset for far-field multi-channel speaker verification," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7977–7981, 2021.

[28] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, 2020.

[29] M. Ajili, J.-F. Bonastre, J. Kahn, S. Rossato, and G. Bernard, "Fabiole, a speech database for forensic speaker comparison," in *International Conference on Language Resources and Evaluation*, 2016.

[30] S. Li, H. Liu, Y. Zhou, and Z. Luo, "A si-sdr loss function based monaural source separation," in *2020 15th IEEE International Conference on Signal Processing (ICSP)*, vol. 1, 2020, pp. 356–360.

[31] S. Lyu, "Interpretation and generalization of score matching," in *Conference on Uncertainty in Artificial Intelligence*, 2009.

[32] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv: Learning*, 2016.

[33] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.