Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification

Sandipana Dowerah, Romain Serizel, Denis Jouvet, M Mohammadamini, Driss Matrouf

Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification

Sandipana Dowerah[1], *Romain Serizel*[1], *Denis Jouvet*[1], *M. Mohammadamini*[2], *Driss Matrouf*[2]

[1]Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Laboratoire Informatique d'Avignon, Avignon University, France

firstname.lastnamer@loria.fr, firstname.lastname@univ-avignon.fr

Abstract

Today's smart devices using speaker verification are getting equipped with multiple microphones resulting in improving spatial ambiguity and directivity. However, unlike any other speech-based applications, the performance of speaker verification degrades in far-field scenarios due to the adverse effects of a noisy environment and room reverberation. This paper presents a novel multichannel speech enhancement module based on the diffusion probabilistic model. It is used as the front-end of the ECAPA-TDNN speaker verification system in far-field scenarios under a noisy-reverberant environment. The proposed system incorporates a two-stage training approach. In the first stage, both speech enhancement and speaker verification modules are trained individually. In the second stage, both the modules are combined to jointly trained them. We use similarity-preserving knowledge distillation loss that guides the network to produce similar activation for enhanced signals to that of clean speech signals. Using joint optimization with knowledge distillation loss achieved the best performance on both the evaluation composed of synthetic clips similar to those used at training and on unseen recorded clips from the VOiCES dataset.

Index Terms: multichannel speech enhancement, diffusion model, far-field speaker verification

1. Introduction

Speaker Verification (SV) aims at verifying the identity of speakers based on their voice characteristics. The use of neural networks in recent times has led to the successful implementation of SV under controlled conditions or close-talk scenarios for personalized authentication. The state-of-the-art SV systems (eg., Time Delay Neural Network [1], ResNet [2], ECAPA-TDNN [3]) commonly known as x-vector systems [4] have consistently improved SV performance in recent years but SV still suffers in far-field scenarios mainly due to long-range fading, complex environmental noises, and room reverberation. Several challenges like VOiCES from a distance challenge [5], Interspeech Far-field speaker verification challenge [6], etc have been organized over the years to address these issues.

Speech enhancement is the process of improving intelligibility and quality of speech by mapping distorted speech signals to clean signals. It can be used as a pre-processing to SV. Conventional speech enhancement methods compute the mapping of noisy and clean speech signals by first converting them into spectral features through short-time Fourier transform in time-frequency (T-F) domain. The mapping function of noisy-to-clean spectral features is then formulated by a direct mapping [7], or a masking function [8]. In multichannel scenarios, DNNs have been used to compute the T-F masks separating speech and noise from a mixture signal [9] which are then used to estimate the speech and noise covariance matrices for beamforming [10]. Although speech enhancement has been used for compensating adverse effects of noise robustness and reverberation as a front-end to speech recognition where joint optimization has been shown to improve the performance [11, 12], it is also investigated for SV with promising results. Among them, some jointly optimized or integrated weighted prediction error (WPE) and some variants of beamforming using speaker embedding model for reducing the error rate [13, 14, 15]. Shon et al. integrate speech enhancement and SV module into a single framework for SV [16]. Shi et al. used attention mechanism and cascaded speech enhancement network and speaker recognition by jointly optimizing their parameters using single loss function [17]. But, most of the previous studies either processed the speech enhancement module individually or the SV module was pre-trained and frozen during training of the speech enhancement. Moreover, most of them are for single-channel data and are invariably applied for multichannel with some additional processing, for instance, the multichannel signal is mapped to single-channel first by using BeamformIt [1][18] or embedding averaging.

Diffusion probabilistic models (DPM) have shown impressive performance in image generation [19] and Text-to-speech systems [20]. This paper presents GradSE, a novel multichannel speech enhancement module based on DPM with score-based generative model [21]. The score-based generative model relies on computing gradients of the log probability density of noise on a large number of noise-perturbed data distributions. We named our proposed speech enhancement module GradSE as the main function of the neural network is used to compute the gradient of log probability density of noise. Recently, DiffuSE [22] and CDiffuSE [23] were proposed to recover clean speech signals from noisy signals based on Markov chains to provide a framework of denoising diffusion probabilistic models. Instead of Markov chains for diffusion used in DiffuSE and CDiffuSE, we opted for the scoring function to allow the forward diffusion process to transform clean signal to noisy signal. We provide multichannel noisy Mel spectrogram as input to GradSE to predict the Mel spectrogram of clean speech. As using Mel spectrogram to design the diffusion probabilistic models ease the joint optimization of the SV system with the speech enhancement module as a single pipeline. We use ECAPA-TDNN [3] based SV system for jointly optimizing with GradSE to compensate noisy and reverberant conditions in far-field multichannel scenarios. Section 2 explains the proposed model, section 3 describes the dataset, section 4 narrates the experimentation and section 5 illustrates the results and section 6 concludes.

[1]https://github.com/xanguera/BeamformIt

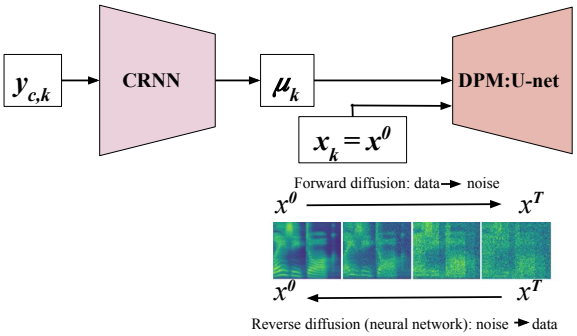

Figure 1: *Architecture of diffusion probabilistic model-based GradSE in training phase. $y_{c,k}$ is noisy multichannel Mel spectrogram from c number of channels with k number of Mel spectrogram frames. μ_k is output of encoder, x_k is clean Mel spectrogram frames, and x^0 is the starting of time-steps until x^T.*

2. System Overview

2.1. DPM-based Multichannel speech enhancement

DPM consists of forward and reverse diffusion processes. In scoring-based DPM, the training phase involves iteratively converting the Mel spectrogram of clean speech to a noisy spectrogram with the noise distribution represented by $\mathcal{N}(\mu, I)$ where μ is mean and I is unit variance. This is known as the forward diffusion process. A reverse diffusion process is used to gradually restore the clean input by predicting and removing the noise introduced in each step of the diffusion process.

The proposed GradSE architecture is composed of an encoder and a decoder network as shown in figure 1. Our model architecture is inspired from Grad-TTS framework [20]. We give noisy multichannel Mel spectrogram $y_{c,k}$ as input to the encoder, where c is the number of channels (microphones) and k is the number of Mel spectrogram frames. We use encoder to compute μ_k which is used to define the noise distribution $\mathcal{N}(\mu_k, I)$. For computing μ_k, we use convolutional recurrent neural network (CRNN) with a convolutional layer, batch normalization, ReLu activation function, and LSTM layers in the convolutional block.

For the DPM-based decoder network, we use U-net denoted by s_θ from Ronneberger et al. [24] . In the training phase, we conducted forward diffusion process by providing Mel spectrogram of clean speech x_k along with encoder output μ_k. Hence, DPM-based decoder s_θ learns to predict the gradient of log probability density of noise. In inference phase, we performed reverse diffusion process by providing encoder output μ_k as input to the decoder. After that, the decoder s_θ iteratively predicts the gradient of log probability density of noise reverse in time. This iterative process transforms the μ_k into Mel spectrogram of clean speech x_t. Thus, the diffusion process can be explained as given below;

$$dx^t = \frac{1}{2}(\mu - x^t - s_\theta(x^t, \mu, t)\beta_t dt \tag{1}$$

where x^t is Mel spectrogram of the clean speech at diffusion step t with predefined noise-scheduler and diffusion step horizon varies as $0 - T$. β_t is non-negative function which we refer as noise scheduler that controls the way noise is added in the diffusion forward process.

It is easier for decoding if we start from noise, which is already close to Mel spectrogram x_k of clean speech to train GradSE. Therefore, we use two-loss criteria, mean square error (MSE) loss and diffusion loss. We applied MSE loss on the encoder output with respect to Mel spectrogram of clean speech x_k. We use scoring-based DPM which uses Fisher divergence to define the diffusion loss [25]. Fisher divergence minimizes the divergence between the gradient of the log density of noisy data and the gradient predicted by DPM-based U-net decoder s_θ. Diffusion loss can be explained formally as,

$$\mathcal{L}_{diffusion} = \mathbb{E}_{p(x)}[\|\nabla log p_t(x^t) - s_\theta(x^t, \mu, t)\|_2^2] \tag{2}$$

where $\nabla log p_t(x^t)$ is gradient of log probability density of noise at step t and output of $s_\theta(x^t, \mu, t)$ at step t. Thus, diffusion loss enables s_θ to generate a better estimate of reverse trajectories of forward diffusion process.

2.2. Joint Optimization

We use ECAPA-TDNN based SV system [3], which demonstrates state-of-the-art performance compared to X-vector or ResNet systems. The proposed system incorporates a two-stage training approach. In the first phase, we trained GradSE and ECAPA-TDNN individually on the training dataset. In the second phase, we combined GradSE and ECAPA-TDNN systems by giving multichannel noisy Mel spectrogram as input to GradSE shown in figure 2. GradSE then performs a reverse diffusion process to remove the noise in input to reconstruct the target clean Mel spectrogram with a time-horizon of 20 time-steps, which means the network conducts 20 steps to reconstruct the target clean Mel spectrogram. After that, the output Mel spectrogram is passed through the ECAPA-TDNN network to generate speaker embedding. We provide generated speaker embedding to the classifier to derive the softmax probability distribution, which is later used for computing cross-entropy loss with target speaker labels.

2.3. Loss Function

We use knowledge distillation (KD) loss to derive the information to minimize the distance between speaker embedding from noisy signals and clean speech signals. For implementing KD loss, we use similarity-preserving KD [2] [26] loss which guides the network to produce similar activation for noisy signals to that of clean speech signals than to imitate the representation space of teacher model. Therefore, KD loss enables proposed joint optimization to generate embeddings closer to that generated by clean speech. Similarity-preserving KD loss is a novel form of KD that uses the pairwise activation similarities within each input mini-batch to supervise the training of a student network with a trained teacher network. In the proposed architecture, we use the same pre-trained ECAPA-TDNN network to generate output embedding on clean speech as a teacher network and proposed a jointly optimized model as a student network. In addition to GradSE, we use FaSNet based speech enhancement system for implementing a baseline system.

3. Dataset

3.1. RoboVoices

We use the dry speech (clean) data from the clean subset of Librispeech [27] corpus which is approximately 1000 hours of

[2]https://github.com/AberHu/Knowledge-Distillation-Zoo

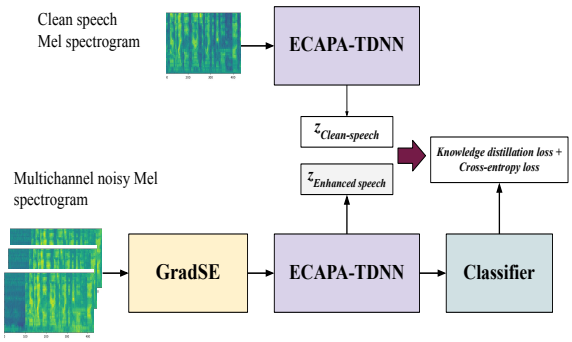

Figure 2: *Joint optimization of multichannel speech enhancement with speaker verification using knowledge distillation loss. Embedding generated by ECAPA-TDNN for clean speech Mel spectrogram is considered as the teacher network. Embedding generated by joint optimized network on multichannel noisy Mel spectrogram is considered as the student network.*

English speech data collected as part of the Librivox project. We have selected around 10000 files randomly from the clean training subset of Librispeech and truncated them to 10 seconds duration for the training set, contributing to 25 hours of speech data. For evaluation of the SV system, we use the Fabiole speech corpus [28]. Fabiole is a French speech corpus consisting of around 6882 audio files from 130 native French speakers. The minimum duration of the speech file is 1 second and the maximum is 46 seconds. The speech data of the corpus has been collected from different French radio and television shows. For creating each evaluation set, we have used 1200 speech files from Fabiole representing 2 hrs of evaluation material.

We used realistic office noises from Freesound [3] [29]. The selected noise categories include door, keyboard, office, phone, background noise in the room, printer, fan, door knock, babble, and environmental noise. We divided the dataset into two sets: a training set of 3725 clips and an evaluation set of 1000 clips.

To simulate room effects, we have generated an RIR corpus of 10000 rooms for training and 3600 for evaluation with pyroomacoustics toolbox [30]. For training, 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. The RIR for the evaluation set was generated with the same room dimensions as in the training set but the absorption coefficient was selected to obtain an RT60 of 400 ms.

The final RoboVoices corpus for training and evaluation is created by first convolving the dry speech and noise with the simulated RIRs from different location in the same room. We then added the convolved dry speech and convolved noise to obtain the noisy signal. We randomly select the noise samples from Freesound and the dry speech from Librispeech for the training set. For training, the SNR is drawn randomly with a uniform distribution between $[0-10]$ dB. For the evaluation set, the generation process is similar except that we draw the SNR values in $5, 10, 20$dB, and the process is applied to each speech segment from the Fabiole dataset. In total, we have generated 10000 mixtures for training and 3600 mixtures for evaluation.

[3] https://freesound.org/

Table 1: *% EER on different utterance lengths on RoboVoices dataset. Performance is averaged over SNR conditions. Joint optim. in the table refers to joint optimization of speech enhancement and SV module. Confidence interval is* 0.1.

	Utterance length	Below 4 secs	Above 4 secs
	Model	**EER**	**EER**
	Dry clean speech	18.8	5.6
	Reverberated clean speech	21.1	7.5
	Noisy	27.8	9.5
	BLSTM Rank-1 [14]	27.4	9.2
	BLSTM MVDR Rank-1 [14]	27.5	9.4
	FaSNet	28.1	10.3
	FaSNet Rank-1 WPE	27.5	9.0
	GradSE	26.7	8.6
Joint optim.	FaSNet + ECAPA-TDNN	26.8	8.7
Joint optim.	GradSE + ECAPA-TDNN	26.2	8.3
Joint optim.	FaSNet + ECAPA-TDNN + KD loss	26.1	8.3
Joint optim.	GradSE + ECAPA-TDNN + KD loss	**25.8**	**7.9**

3.2. VOiCES

We also evaluate our approach on the VOiCES challenge 2019 dataset [5]. Among the 11 microphone positions in the evaluation set, we select 3 microphone positions of the same microphone types. The microphones are close enough to be considered as a compact microphone antenna. The identification of these microphone in the original corpus are 2, 4, and 9. The resulting virtual antenna is in mid-distance from the speaker.

4. Experimentation

4.1. Experimental set-up

4.1.1. Multichannel speech enhancement

We extract 40 dimensional Mel spectrogram features using torchaudio library with a window length of 400 samples, hop size of 160, and 512 FFT length. For GradSE, the CRNN encoder is implemented using a 2D convolutional block of kernel size 3×3, a stride of 1, and padding of 1 with 3 input channels and single output channel. We used 4 LSTM layers of 40 hidden dimensions. The encoder output is concatenated channel-wise and provided to DPM-based decoder. We use the same network configuration of U-net from Ronneberger [24]. GradSE is trained for 500 iterations with Adam optimizer using learning rate of $1e^{-4}$. We use batch size of 32 for training GradSE.

4.1.2. Speaker verification

We use ECAPA-TDNN model architecture introduced by Desplanques et al. [4]. The squeeze and excitation block and attention module of ECAPA-TDNN is set to 128 and 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 on VoxCeleb1 and VoxCeleb2 dataset with a cyclic learning rate varying between 1e-8 and 1e-3 using the triangular policy with Adam optimizer. We trained the ECAPA-TDNN network with angular margin softmax with a margin of 0.3 and softmax pre-scaling of 30. We trained the ECAPA-TDNN network for 100k iterations and used Mel spectrogram features of 40 dimensions as input to ECAPA-TDNN network, extracted using the same procedure as used for

Table 2: *%EER on different noise conditions of the VOiCES Eval dataset. Joint optim. in the table refers to joint optimization of speech enhancement and SV module. Confidence interval is 0.2.*

		Noise conditions			
	Model	**Clean**	**Babble**	**TV**	**Music**
	Unprocessed	4.1	8.8	7.8	7.9
	BLSTM Rank-1	4.1	7.9	7.1	7.2
	FaSNet	4.4	7.8	7.4	7.9
	FaSNet Rank-1 WPE	4.2	6.9	6.4	6.8
	GradSE	3.9	6.7	6.2	6.6
Joint optim.	FaSNet ECAPA-TDNN	4.0	6.6	6.2	6.5
Joint optim.	GradSE ECAPA-TDNN	3.8	6.4	6.0	6.2
Joint optim.	FaSNet + ECAPA-TDNN + KD loss	3.9	6.6	6.1	6.3
Joint optim.	GradSE + ECAPA-TDNN + KD loss	**3.8**	**6.2**	**5.9**	**6.1**

the GradSE module.

4.1.3. Joint Optimization

After training GradSE and ECAPA-TDNN individually, we jointly optimized both the networks using cross-entropy loss on predicted labels by classifier and target speaker labels and similarity preservation KD loss. We performed the joint optimization using Adam optimizer and cyclic learning rate scheduler varying between 1e-3 and 1e-1 using the triangular2 policy. During the joint optimization process, we trained the network with angular margin softmax with a margin of 0.4 and softmax pre-scaling of 30. We used batch size of 64, and trained for 20k iterations. We opted for 20 steps for reverse diffusion for speech enhancement after analysing the trade-off between performance on SV and inference speed.

4.2. Evaluation

We compute equal error rate (EER) to evaluate our system. All metrics are presented with a 95 % confidence interval using the bootstrap algorithm [31]. We consider different conditions corresponding to different steps in the acoustic propagation process: dry clean speech, reverberated clean speech, and Noisy (mixture of reverberated noise and speech). We compute EER on these conditions, and on the signals estimated with different speech enhancement algorithms.

5. Results and Analysis

Table 1 shows the evaluation results for SV in terms of EER depending on utterance lengths. The performance for both the utterance lengths are averaged over SNR conditions. For a comprehensive comparison we include other state-of-the-art pre-processing techniques in our experiments. We implement the BLSTM-based models from Taherian et al. [14] and FaSNet-based models from our previous work [32] and consider them as baselines. First, we compare GradSE to the separately trained pre-processing approaches, then joint optimization is done using both speech enhancement and SV module. Joint optimization yields better performance on both the utterance lengths. Using KD loss further enhances the SV performance. In terms of the two joint optimized models, the proposed GradSE-based model outperforms the FaSNet-based model. With joint optimization we observe an absolute error reduction of 2%. FaSNet is good at speech enhancement but when applied to SV, the performance degrades mainly due to the artifacts FaSNet introduced during training as observed in our previous work [32].

Table 2 reports the results for different distractor noise conditions on publicly available VOiCES Eval dataset. As expected all the approaches achieved the best performance in the condition without any distractor noise (Clean in the table). The baseline BLSTM Rank-1 performs poorly compared to proposed GradSE in all the noise conditions. Babble seems to be the most difficult condition with an equal error rate of 8.8% without any pre-processing due to overlapping speech interference as well as its similarity to the desired clean speech. However, the error rate is reduced to 6.7% by GradSE alone and 6.2% with joint optimization of GradSE and ECAPA-TDNN using KD loss. Joint optimization of speech enhancement and SV using KD loss outperforms all the other approaches on all the noise conditions even though the model was trained on a synthetic dataset.

6. Conclusion

This paper described a novel diffusion probabilistic-based multichannel speech enhancement module for speaker verification in far-field/distant noisy-reverberant scenarios. Our speech enhancement module consisted of an encoder and a diffusion probabilistic model-based decoder and ECAPA-TDNN-based speaker verification system. We compared our proposed system to the state-of-the-art pre-processing techniques. We found that our separately trained speech enhancement system obtained best performance on a synthetic dataset and on VOiCES challenge dataset. Furthermore, we showed that joint optimization of the whole system (GradSE speech enhancement and ECAPA-TDNN speaker verification) using knowledge distillation loss achieved excellent results over separately trained models. The proposed joint optimized model achieved superior performance across noise conditions on VOiCES challenge dataset even though the model was trained on a synthetic dataset. In future, we would like to explore the diffusion probabilistic models to estimate the time-frequency masks for a multichannel noisy-reverberant input.

7. Acknowledgements

The French National Research Agency is funding this research as part of the project Robust voice identification for mobile security robots (ANR-18-CE33-0014). Grid5000 testbed hosted by the University of Lorraine and supported by a scientific interest group hosted by Inria including CNRS, RENATER, and several universities as well as other organizations (see https://www.grid5000), was used in part for the experiments presented in this paper.

8. References

[1] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva, A. Ivanov, A. Kozlov, T. Pekhovsky, and Y. N. Matveev, "Deep speaker embeddings for far-field speaker recognition on short utterances," *Odyssey*, 2020.

[2] D. Cai, X. Qin, W. Cai, and M. Li, "The dku system for the speaker recognition task of the 2019 voices from a distance challenge," in *INTERSPEECH*, 2019.

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

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP*, 2018.

[5] M. K. Nandwana, J. van Hout, C. Richey, M. McLaren, M. Barrios, and A. Lawson, "The voices from a distance challenge 2019," in *INTERSPEECH*, 2019.

[6] X. Qin, M. Li, H. Bu, W. Rao, R. K. Das, S. S. Narayanan, and H. Li, "The interspeech 2020 far-field speaker verification challenge," *ArXiv*, 2020.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *INTERSPEECH*, 2013.

[8] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[9] N. Furnon, R. Serizel, S. Essid, and I. Illina, "Dnn-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[10] J. Heymann, L. Drude, and R. Häb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," *ICASSP*, 2016.

[11] L. Drude, C. Böddeker, J. Heymann, R. Häb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *INTERSPEECH*, 2018.

[12] J. Heymann, L. Drude, R. Häb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online asr," *ICASSP*, 2019.

[13] J.-Y. Yang and J.-H. Chang, "Joint optimization of neural acoustic beamforming and dereverberation with x-vectors for robust speaker verification," in *INTERSPEECH*, 2019.

[14] H. Taherian, Z.-Q. Wang, and D. Wang, "Deep learning based multi-channel speaker recognition in noisy and reverberant environments," in *INTERSPEECH*, 2019.

[15] L. Mošner, P. Matějka, O. Novotný, and J. H. Černocký, "Dereverberation and beamforming in far-field speaker recognition," in *ICASSP*, 2018.

[16] S. Shon, H. Tang, and J. R. Glass, "Voiceid loss: Speech enhancement for speaker verification," in *INTERSPEECH*, 2019.

[17] Y. Shi, Q. Huang, and T. Hain, "Robust speaker recognition using speech enhancement and attention model," *ArXiv*, 2020.

[18] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2007.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, 2020.

[20] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*, 2021.

[21] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *MLR*, 2015.

[22] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *ArXiv*, 2021.

[23] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, and Y. T. Cheng Yu, "Conditional diffusion probabilistic model for speech enhancement," *ArXiv*, 2022.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[25] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2020.

[26] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *ICCV*, 2019.

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

[28] M. Ajili, J. Bonastre, J. Kahn, S. Rossato, and G. Bernard, "Fabiole, a speech database for forensic speaker comparison," in *LREC*, 2016.

[29] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *ISMIR*, 2017.

[30] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *ICASSP*, 2018.

[31] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *ICASSP*, 2004.

[32] S. Dowerah, R. Serizel, D. Jouvet, M. Mohammadamini, and D. Matrouf, "Compensating noise and reverberation in far-field Multichannel Speaker Verification," 2022, working paper or preprint. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03619903