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, Romain Serizel*, Denis Jouvet*, Mohammad Mohammadamini† Driss Matrouf†*

*Université de Lorraine, CNRS, Inria, Loria, F-54000, Nancy, France
† Laboratoire Informatique d'Avignon, University of Avignon, France

ABSTRACT

Smart devices using speaker verification are getting equipped with multiple microphones, improving spatial ambiguity and directivity. However, unlike 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 diffusion probabilistic models-based multichannel speech enhancement as a front-end for the ECAPA-TDNN speaker verification system in a far-field noisy-reverberant scenario. The proposed approach incorporates a two-stage training approach. In the first stage, we individually train the speech enhancement and speaker verification modules. In the second stage, we combined both modules and trained them jointly. We use similarity-preserving knowledge distillation loss that guides the network to produce similar activation for enhanced signals like clean signals. Joint optimization achieved the best results on synthetic and VOiCES datasets.

Index Terms— multichannel speech enhancement, far-field speaker verification, deep neural network

1. INTRODUCTION

Speaker Verification (SV) aims to verify the identity of speakers based on their voice characteristics. In recent times, the use of neural networks 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 (e.g., 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. However, 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] and Interspeech Far-field speaker verification challenge [6] have been organized over the years to address these issues.

French National Research Agency supports this work in the framework of the ROBOVOX project (ANR-18-CE33-0014)). Experiments were partially carried out using the Grid5000 testbed supported by a scientific group of Inria including CNRS, RENATER and other Universities and organizations (see https://www.grid5000) hosted by the University of Lorraine.

Speech enhancement improves intelligibility and speech quality by mapping distorted speech signals to clean signals and 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 a short-time Fourier transform in the 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 apply to compute the T-F masks separating speech and noise from a mixture signal [9] and 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 an attention mechanism and cascaded speech enhancement network and speaker recognition by jointly optimizing their parameters using a single loss function [17]. However, most of the previous studies either processed the speech enhancement module individually or the SV module was pre-trained and frozen during the training of the speech enhancement. Moreover, most of them are for single-channel data and are invariably applied for multichannel with additional processing. For instance, the multichannel signal is mapped to a 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 a 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-

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

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

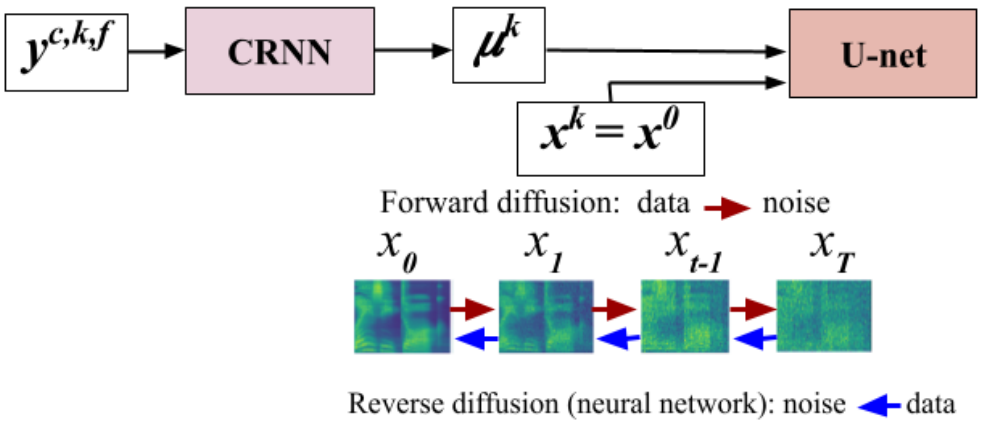

Fig. 1. *Proposed GradSE model in training phase, which is composed of an encoder (CRNN) and a DPM-based decoder (U-net). We give noisy multichannel Mel spectrogram as input to encoder from c number of channels with k number of Mel spectrogram frames with f dimension of mel spectrogram frames. The encoder output, μ^k which represents the noise distribution and clean speech Mel spectrogram x^k are given as input to the diffusion-based decoder. x_0 is the starting of time-steps until x_T. Red arrow depicts the forward diffusion process and blue arrow depicts the reverse diffusion process.*

perturbed data distributions. We named our proposed speech enhancement module GradSE as the main function of the neural network used to compute the gradient of the 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 for diffusion probabilistic models. The denoising diffusion probabilistic model is developed by training the Markov chain for forward and reverse diffusions. These Markov chains fixed the Markov chain, thus leading to slower inference speed. DiffuSE and CDiffuSE use fixed Markov chains for training diffusion models under the framework of DPM. On the other hand, scoring-based diffusion models (GradSE) implement using Stochastic calculus (Stochastic differential equation (SDE)). SDE provides not only an easy-to-use framework for training DPMs [24] but also controls the selection of the number of reverse diffusion steps for enhancement over the noisy multichannel Mel spectrogram. In GradSE, inference phase (reverse diffusion), sampling is conducted from conditional noise distribution. Also, it is easier to decode if we sampled from the noise closer to the target Mel spectrogram. We use ECAPA-TDNN [3] based SV system for jointly optimizing with GradSE to compensate for noisy and reverberant conditions in far-field multichannel scenarios. To further improve the performance of the jointly optimized system, we propose using a novel similarity-preserving knowledge distillation technique to minimize the distance between speaker embeddings obtained from the proposed system and clean speech signals. Section 2 explains the proposed model, section 3 describes the dataset, section 4 narrates the experimentation, section 5 illustrates the results, and section 6 concludes.

2. SYSTEM OVERVIEW

2.1. DPM-based Multichannel speech enhancement

DPMs were introduced to represent the complex data distribution using stochastic calculus [25]. DPMs consist of two processes, namely, (i) a forward diffusion process and (ii) a reverse diffusion process. The forward diffusion process is built by iteratively deconstructing data until we obtain a simple distribution, such as the Gaussian distribution, $\mathcal{N}(0, I)$ with zero mean and unit variance. In the reverse diffusion process, DPM reconstructs the data by sampling noise from the Gaussian distribution, parameterizing reverse diffusion with a neural network. Consequently, for the reverse diffusion process, the neural network predicts the trajectories of the forward diffusion process in reverse to generate the data from the sampled noise.

This paper introduces GradSE, a novel DPM-based multichannel speech enhancement for SV. The architecture of GradSE is inspired from Grad-TTS framework [20] and WaveGrad [26] systems. GradSE comprises an encoder and a decoder network, as shown in Figure 1. We used a convolutional recurrent neural network (CRNN) to implement the encoder network. CRNN network comprises a convolutional layer, batch normalization, ReLU activation function, and LSTM layers. The encoder network defines conditional noise distribution in the diffusion process. We used the U-net network denoted by s_θ from Ronneberger et al. [27] to implement the decoder network. The decoder network carries out the forward and reverse diffusion processes. The forward diffusion process uses for training the GradSE, and the reverse diffusion process is for the inference phase.

GradSE is a scoring-based diffusion probabilistic model,

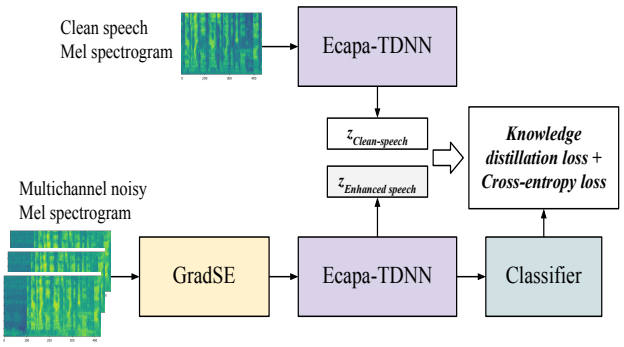

Fig. 2. *Joint optimization of multichannel speech enhancement with SV using KD loss. Embedding generated by ECAPA-TDNN for clean speech Mel spectrogram is considered as the teacher network. Embedding generated by joint optimized network on multichannel noisy Mel spectrogram is considered as the student network.*

where scoring refers to the gradients of the log probability density of the noise [28, 29]. Song et al. [24] illustrated that the diffusion process based on the Markov chain is an approximated trajectory of the stochastic process satisfying SDE [30]. GradSE formulates a reverse diffusion process by an SDE solver using the first-order Euler scheme [31] and matches the trajectories of reverse diffusion to the forward diffusion but in reverse time order.

In brief, the neural network minimizes the scoring function in the forward diffusion process and learns to deconstruct the clean speech Mel spectrogram to noise distribution. And in the reverse diffusion process, the neural network predicts the gradients of the log density of noise, thus enabling forward diffusion to reconstruct the clean speech Mel spectrogram in reverse-time order. GradSE's forward diffusion transforms the clean speech Mel spectrogram distribution into a noise distribution defined as $\mathcal{N}(\mu, I)$, where the mean is μ, and I is unit variance. Thus, the encoder output allows the conditional generative modeling by conditioning the diffusion process's terminal distribution with mean denoted by μ. After obtaining μ, it is parameterized with latent variable sampled from Gaussian distribution, thus creating conditional Gaussian distribution depending on the μ. Afterward, for given parameterized μ as input to the U-net decoder, the decoder performs reverse diffusion to transform latent variables into estimates of the target Mel spectrogram.

2.1.1. Training

We give noisy multichannel Mel spectrogram $y^{c,k,f}$ as input to the encoder, where c is the number of channels (microphones), k is the number of Mel spectrogram frames, and f is the dimension of the mel spectrogram frames. Next, we use an encoder to compute μ^k, which is then used to estimate the noise distribution $\mathcal{N}(\mu^k, I)$. For simplicity of notation, we denoted encoder output, μ^k as μ. Finally, we give the noise distribution of the encoder to the DPM-based decoder, which is then used to perform the forward diffusion process. In the forward diffusion process, the decoder network iteratively deconstructs the Mel spectrogram of clean speech $x^k = x_0$ to noise x_T, where T is the terminal time horizon shown in Figure 1.

We parameterize the terminal noise distribution of the forward diffusion process by μ. Therefore, the decoder network learns the trajectories of the forward diffusion through the scoring function. The forward diffusion process can be explained mathematically as below;

$$x_0 \sim \mathcal{P}_{data} \Longrightarrow x_T = \tau(x_0) \sim \mathcal{N}(\mu, I) \tag{1}$$

where, $\mathcal{P}_{data}$ is data distribution of clean speech Mel spectrogram and forward diffusion process, τ to slowly deconstruct x_0 by adding noise to simple distribution defined by $\mathcal{N}(\mu, I)$.

2.1.2. Inference

In the inference phase, enhancement is performed by the reverse diffusion process. We provide the noisy multichannel Mel spectrogram as input to the encoder network, which gives encoder output μ. Furthermore, we carried out the reverse diffusion process by providing encoder output μ as input to the decoder. Thus noise distribution is parameterized through encoder output μ. The reverse diffusion process reconstructs the Mel spectrogram of clean speech by gradually removing the noise sampled from the distribution $\mathcal{N}(\mu, I)$.

In each step of reverse diffusion, the reverse trajectories of the forward diffusion are defined by SDE with an estimated scoring function from the decoder network. This iterative process transforms the encoder output μ into the Mel spectrogram of the clean speech x_0. The reverse diffusion process is explained as given below;

$$x_T \sim \mathcal{N}(\mu, I) \Longrightarrow \tau_{-1}(x_T) \sim \mathcal{P}_{data} \tag{2}$$

where, x_T denotes the noise sampled from terminal noise distribution defined by $\mathcal{N}(\mu, I)$ with μ as encoder output. τ_{-1} denotes reverse diffusion process to construct the Mel spectrogram of the clean speech from data distribution, $\mathcal{P}_{data}$.

2.1.3. Loss function

We use two-loss criteria, mean square error (MSE) and diffusion loss. We applied MSE loss on the encoder output concerning the Mel spectrogram of clean speech x^k. It is easier for decoding if we start from noise, which is already close to Mel spectrogram x^k of the clean speech to train GradSE. We used MSE to ensure the training process's stability and provide smooth global optima in the optimization process.

We use scoring-based DPM, which uses Fisher divergence to define the diffusion loss [32]. Fisher divergence minimizes

the divergence between the gradient of the log density of the noise and the gradient predicted by the DPM-based U-net decoder s_θ. Thus, diffusion loss enables s_θ to generate a better estimate of reverse trajectories of the forward diffusion process. The diffusion loss can be explained formally as,

$$\mathcal{L}_{diffusion} = E_{p(x)}[\|\nabla log p_t(x_t) - s_\theta(x_t, \mu, t)\|_2^2] \quad (3)$$

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.

2.2. Joint Optimization

We optimize the front-end speech enhancement with the SV jointly in a single framework. We use ECAPA-TDNN based SV system [3], which demonstrates state-of-the-art performance compared to TDNN [1] or ResNet [2] 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 both GradSE and ECAPA-TDNN systems by giving multichannel noisy Mel spectrogram as input to GradSE as 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 reverse diffusion steps to reconstruct the target clean Mel spectrogram.

The enhanced Mel spectrogram from GradSE is then passed through the ECAPA-TDNN network. Next, we extract the last hidden output of the ECAPA-TDNN network as speaker embedding, where the ECAPA-TDNN network is trained for a classification task. Then, we provide the 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. Finally, for the joint optimization, the error gradients are passed through ECAPA-TDNN and GradSE in the backpropagation pass.

2.3. Knowledge Distillation Loss

The knowledge distillation (KD) is used as a model compression technique that extracts the knowledge of a large pre-trained neural network model (teacher) and transfers it to a small neural network model (student) [33]. Traditionally knowledge distillation (KD) is used for improving inference speed and reducing the model parameters. Moreover, the distillation loss determines the process of reducing the divergence between the output distribution of the teacher network and the student network.

We propose to use the novel form of KD known as similarity-preserving knowledge distillation (KD) loss introduced by Tung et al. [34] [2]. Similarity-preserving KD loss is motivated by the idea that semantically similar inputs tend to obtain similar activation patterns in a trained neural network [34]. The main idea of similarity-preserving KD loss is to use the pairwise activation similarities within each input mini-batch to supervise the training of a student network with a trained teacher network. The similarity-preserving KD loss requires the student network only to maintain the pairwise similarities in its own representation space rather than replicating the teacher network's representation space.

We apply similarity-preserving KD loss to facilitate the jointly optimized system to generate embeddings closer to those generated by clean speech. Thus, similarity-preserving KD loss derives the information to minimize the distance between speaker embeddings obtained from the proposed system ($z_{Enhanced-speech}$) and that of clean speech signals ($z_{Clean-speech}$), as shown in Figure 2. We use the ECAPA-TDNN model as a teacher network with embeddings obtained on clean speech signals and our proposed joint optimization of the GradSE and ECAPA-TDNN as the student network. The similarity-preserving KD loss assists the student model in matching the performance of enhanced signal embedding to the embeddings from a clean speech signal. Thus, KD technique allows the proposed system to learn the robust latent space of speaker representation in noisy scenarios.

3. DATASET

3.1. RoboVoices

We use dry clean speech data from the clean subset of Librispeech [35] corpus, which is approximately 1000 hours of English speech data collected as part of the Librivox project. In addition, 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 Fabiole speech corpus [36]. 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 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 realistic office noises from Freesound [3] [37]. 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 [38]. For training, the room length

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

[3] https://freesound.org/

Table 1. *% EER on different utterance lengths and SNR on RoboVoices dataset. Performance is averaged over SNR conditions for utterance lengths . Joint optim. refers to joint optimization of both modules. Confidence interval is 0.1.*

	Testing environment	EER	EER	EER/SNR		
	Utterance length	**Below 4 secs**	**Above 4 secs**	**5**	**10**	**20**
	Dry clean speech	18.8	5.6	5.6	5.6	5.6
	Reverberated clean speech	21.1	7.5	7.5	7.5	7.5
	Noisy	27.8	9.5	11.2	9.4	7.8
	BLSTM Rank-1 [14]	27.4	9.2	10.9	9.0	7.8
	BLSTM MVDR Rank-1 [14]	27.5	9.4	10.8	9.1	7.7
	FaSNet	28.1	10.3	12.4	10.5	8.0
	FaSNet Rank-1 WPE	27.5	9.0	10.5	8.8	7.7
	GradSE	26.7	8.6	10.2	8.5	7.2
Joint optim.	FaSNet + ECAPA-TDNN	26.8	8.7	10.1	8.4	7.5
	GradSE + ECAPA-TDNN	26.2	8.3	9.8	8.0	7.1
	FaSNet + ECAPA-TDNN + KD loss	26.1	8.3	9.9	8.0	7.1
	GradSE + ECAPA-TDNN + KD loss	**25.8**	**7.9**	**9.2**	**7.7**	**6.8**

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 exact room dimensions as in the training set. However, 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 locations in the same room. We then added the convolved dry speech and convolved noise to obtain the noisy signal. Subsequently, we select the noise samples randomly 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.

3.2. VOiCES

We also evaluate our approaches on the publicly available Voices Obscured in Complex Environmental Settings (VOiCES) challenge 2019 Eval 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 a compact microphone antenna. The identification of these microphones in the original corpus is 2, 4, and 9. We select the signal from these three microphones confirming that all three are in mid-distance from the speaker and are close to building a "virtual" microphone antenna.

4. EXPERIMENTATION

4.1. Experimental set-up

4.1.1. Multichannel speech enhancement

We extract 40 dimensional Mel spectrogram features using the 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 a single output channel. We used 4 LSTM layers of 40 hidden dimensions. The encoder output is concatenated channel-wise and provided to the DPM-based decoder. We use the same network configuration of U-net from Ronneberger [27]. GradSE is trained for 500 iterations, using a batch size of 32, and a learning rate of $1e^{-4}$.

4.1.2. Speaker verification

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

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			
	Testing environment	**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
	GradSE ECAPA-TDNN	3.8	6.4	6.0	6.2
	FaSNet + ECAPA-TDNN + KD loss	3.9	6.6	6.1	6.3
	GradSE + ECAPA-TDNN + KD loss	**3.8**	**6.2**	**5.9**	**6.1**

ECAPA-TDNN network extracted using the same procedure as used for GradSE. We used cosine scoring system for verification purpose from extracted embedding..

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 the classifier and target speaker labels and similarity preservation KD loss. We performed the joint optimization using the 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 a batch size of 64 and trained for 20k iterations. We opted for 20 steps for reverse diffusion for speech enhancement after analyzing the trade-off between performance on SV and inference speed.

4.2. Evaluation

We compute an equal error rate (EER) to evaluate our system. All metrics are presented with a 95 % confidence interval using the bootstrap algorithm [39]. 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 the signals estimated with different speech enhancement algorithms.

5. RESULTS AND ANALYSIS

We present the evaluation results for SV in terms of EER. Table 1 shows the results of our experiments on the RoboVoices dataset. To compute the EER for both the utterance lengths, the SNR conditions are averaged. 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 Dowerah et al. [40] and consider them as baselines. The baseline BLSTM-based and FaSNet-based models are trained with the same data as used for the proposed approaches. 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 consistently improves the performance on both the utterance lengths as well as on all the SNR conditions. 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 [40].

Table 2 reports the results for different distractor noise conditions on 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 the proposed GradSE in all the noise conditions. Babble seems to be the most challenging 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 both multichannel 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

We introduced GradSE, a novel multichannel speech enhancement approach based on diffusion probabilistic models for far-field speaker verification. In order to facilitate speaker verification in adverse conditions, we applied a two-stage approach in which the front-end multichannel speech enhancement is trained separately at first and then jointly optimized with the back-end speaker verification. To the best of our knowledge, this is the first study to apply the diffusion probabilistic models for multichannel speech enhancement as a front-end to speaker verification.

We explored various experimentations, and GradSE consistently improved the performance over state-of-the-art pre-processing approaches. Moreover, the joint optimization of the whole system (GradSE speech enhancement and ECAPA-TDNN SV) using knowledge distillation loss achieved excellent results over separately trained models on the synthetic dataset as well as on the VOiCES dataset. In the future, we would like to investigate various teacher model architectures, such as wav2vec and UniSpeech, under a multi-teacher knowledge distillation setting.

7. REFERENCES

[1] Aleksei Gusev, V. Volokhov, Tseren Andzhukaev, Sergey Novoselov, G. Lavrentyeva, M. Volkova, Alice Gazizullina, Andrey Shulipa, Artem Gorlanov, Anastasia Avdeeva, Artem Ivanov, Alexander Kozlov, Timur Pekhovsky, and Yuri N. Matveev, "Deep speaker embeddings for far-field speaker recognition on short utterances," *Odyssey*, 2020.

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

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

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

[5] Mahesh Kumar Nandwana, Julien van Hout, Colleen Richey, Mitchell McLaren, M. Barrios, and A. Lawson, "The voices from a distance challenge 2019," in *INTERSPEECH*, 2019.

[6] Xiaoyi Qin, Ming Li, Hui Bu, Wei Rao, Rohan Kumar Das, Shrikanth S. Narayanan, and Haizhou Li, "The interspeech 2020 far-field speaker verification challenge," *INTERSPEECH*, 2020.

[7] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder," in *INTERSPEECH*, 2013.

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

[9] Nicolas Furnon, Romain Serizel, Slim Essid, and Irina 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] Jahn Heymann, Lukas Drude, and Reinhold Häb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," *ICASSP*, 2016.

[11] Lukas Drude, Christoph Böddeker, Jahn Heymann, Reinhold Häb-Umbach, Keisuke Kinoshita, Marc Delcroix, and Tomohiro Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *INTERSPEECH*, 2018.

[12] Jahn Heymann, Lukas Drude, Reinhold Häb-Umbach, Keisuke Kinoshita, and Tomohiro Nakatani, "Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online asr," *ICASSP*, 2019.

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

[14] Hassan Taherian, Zhong-Qiu Wang, and DeLiang Wang, "Deep learning based multi-channel speaker recognition in noisy and reverberant environments," in *INTERSPEECH*, 2019.

[15] Ladislav Mošner, Pavel Matějka, Ondřej Novotný, and Jan Honza Černocký, "Dereverberation and beamforming in far-field speaker recognition," in *ICASSP*, 2018.

[16] Suwon Shon, Hao Tang, and James R. Glass, "Voiceid loss: Speech enhancement for speaker verification," in *INTERSPEECH*, 2019.

[17] Yanpei Shi, Qiang Huang, and Thomas Hain, "Robust speaker recognition using speech enhancement and attention model," *Odyssey*, 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] Jonathan Ho, Ajay Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, 2020.

[20] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*, 2021.

[21] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *MLR*, 2015.

[22] Yen-Ju Lu, Yu Tsao, and Shinji Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *APSIPA ASC*, 2021.

[23] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, and Yu Tsao Cheng Yu, "Conditional diffusion probabilistic model for speech enhancement," *ICASSP*, 2022.

[24] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon, "Maximum likelihood training of score-based diffusion models," in *NeurIPS*, 2021.

[25] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.

[26] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan, "Wavegrad: Estimating gradients for waveform generation," in *ICLR*, 2021.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[28] Yang Song and Stefano Ermon, "Generative modeling by estimating gradients of the data distribution," *NeurIPS*, 2019.

[29] Yang Song and Stefano Ermon, "Improved techniques for training score-based generative models," *NeurIPS*, 2020.

[30] Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2020.

[31] Peter E. Kloeden and Eckhard Platen, "Numerical solution of stochastic differential equations," in *Applications of Mathematics book series*, 1977.

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2020.

[33] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," *ArXiv*, 2015.

[34] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation," *ICCV*, 2019.

[35] Vassil Panayotov, Guoguo Chen, Daniel Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *ICASSP*, 2015.

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

[37] Eduardo Fonseca, Jordi Pons, Xavier Favory, F. Font, D. Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *ISMIR*, 2017.

[38] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *ICASSP*, 2018.

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

[40] 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, url: https://hal.archives-ouvertes.fr/hal-03619903.