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

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

To cite this version:

Sandipana Dowerah, Romain Serizel, Denis Jouvet, M Mohammadamini, Driss Matrouf. Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification. IEEE SLT 2022, Jan 2023, Doha, Qatar. hal-03671583v2

HAL Id: hal-03671583
https://hal.science/hal-03671583v2
Submitted on 27 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
JOINT OPTIMIZATION OF DIFFUSION PROBABILISTIC-BASED MULTICHANNEL SPEECH ENHANCEMENT WITH FAR-FIELD SPEAKER VERIFICATION

Sandipana Dowerah*, Romain Serizel*, Denis Jouvet*, Mohammad Mohammadmamin† 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.

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 [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-
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, \( \mu^k \) 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.

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, \( N(0, I) \) 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 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.

GradSE is a scoring-based diffusion probabilistic model,
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 $\mu$, 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 $\mu$. After obtaining $\mu$, it is parameterized with latent variable sampled from Gaussian distribution, thus creating conditional Gaussian distribution depending on the $\mu$. Afterward, for given parameterized $\mu$ 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 $\mu^k$, which is then used to estimate the noise distribution $\mathcal{N}(\mu^k, I)$. For simplicity of notation, we denoted encoder output, $\mu^k$ as $\mu$. 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 $\mu$. 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:

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

where, $\mathcal{P}_{data}$ is data distribution of clean speech Mel spectrogram and forward diffusion process, $\tau$ 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 $\mu$. Furthermore, we carried out the reverse diffusion process by providing encoder output $\mu$ as input to the decoder. Thus noise distribution is parameterized through encoder output $\mu$. The reverse diffusion process reconstructs the Mel spectrogram of clean speech by gradually removing the noise sampled from the data 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 $\mu$ into the Mel spectrogram of the clean speech $x_0$. The reverse diffusion process is explained as given below:

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

where, $x_T$ denotes the noise sampled from terminal noise distribution defined by $\mathcal{N}(\mu, I)$ with $\mu$ as encoder output. $\tau_-$ 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 \( s_q \). Thus, diffusion loss enables \( s_q \) to generate a better estimate of reverse trajectories of the forward diffusion process. The diffusion loss can be explained formally as,

\[
\mathcal{L}_{\text{diffusion}} = E_p(x_t)[\|\nabla \log p_t(x_t) - s_q(x_t, \mu, t)\|_2^2] \tag{3}
\]

where \( \nabla \log p_t(x_t) \) is gradient of log probability density of noise at step \( t \) and output of \( s_q(x_t, \mu, t) \) 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 (\( \mathcal{E}_{\text{Enhanced-speech}} \)) and that of clean speech signals (\( \mathcal{E}_{\text{Clean-speech}} \)), 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 3000 for evaluation with pyroomacoustics toolbox [38]. For training, the room length

\(^2\)https://github.com/AberHu/Knowledge-Distillation-Zoo

\(^3\)https://freesound.org/
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 \(3 \times 3\), 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 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.

<table>
<thead>
<tr>
<th>Testing environment</th>
<th>EER</th>
<th>EER</th>
<th>EER/SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uterance length</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below 4 secs</td>
<td>18.8</td>
<td>5.6</td>
<td>5.6</td>
</tr>
<tr>
<td>Above 4 secs</td>
<td>21.1</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>27.8</td>
<td>9.5</td>
<td>9.4</td>
</tr>
<tr>
<td>BLSTM Rank-1 [14]</td>
<td>27.4</td>
<td>9.2</td>
<td>10.9</td>
</tr>
<tr>
<td></td>
<td>27.5</td>
<td>9.4</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>28.1</td>
<td>10.3</td>
<td>12.4</td>
</tr>
<tr>
<td>FaSNet</td>
<td>27.5</td>
<td>9.0</td>
<td>10.5</td>
</tr>
<tr>
<td>FaSNet Rank-1 WPE</td>
<td>26.7</td>
<td>8.6</td>
<td>10.2</td>
</tr>
<tr>
<td>GradSE</td>
<td>26.8</td>
<td>8.7</td>
<td>10.1</td>
</tr>
<tr>
<td>GradSE + ECAPA-TDNN</td>
<td>26.2</td>
<td>8.3</td>
<td>9.8</td>
</tr>
<tr>
<td>GradSE + ECAPA-TDNN + KD loss</td>
<td>26.1</td>
<td>8.3</td>
<td>9.9</td>
</tr>
<tr>
<td>GradSE + ECAPA-TDNN + KD loss</td>
<td><strong>25.8</strong></td>
<td><strong>7.9</strong></td>
<td><strong>9.2</strong></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Testing environment</th>
<th>Clean</th>
<th>Babble</th>
<th>TV</th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>4.1</td>
<td>7.9</td>
<td>7.1</td>
<td>7.2</td>
</tr>
<tr>
<td>BLSTM Rank-1</td>
<td>4.4</td>
<td>7.8</td>
<td>7.4</td>
<td>7.9</td>
</tr>
<tr>
<td>FaSNet</td>
<td>4.2</td>
<td>6.9</td>
<td>6.4</td>
<td>6.8</td>
</tr>
<tr>
<td>FaSNet Rank-1 WPE</td>
<td>3.9</td>
<td>6.7</td>
<td>6.2</td>
<td>6.6</td>
</tr>
<tr>
<td>GradSE</td>
<td>3.8</td>
<td>6.2</td>
<td>5.9</td>
<td>6.1</td>
</tr>
</tbody>
</table>

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


