ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747771

MULTI-CHANNEL SPEAKER VERIFICATION WITH CONV-TASNET BASED BEAMFORMER

Ladislav Mošner, Oldřich Plchot, Lukáš Burget, Jan “Honza” Černocký

Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

ABSTRACT

We focus on the problem of speaker recognition in far-field multi-channel data. The main contribution is introducing an alternative way of predicting spatial covariance matrices (SCMs) for a beamformer from the time domain signal. We propose to use Conv-TasNet, a well-known source separation model, and we adapt it to perform speech enhancement by forcing it to separate speech and additive noise. We experiment with using the STFT of Conv-TasNet outputs to obtain SCMs of speech and noise, and finally, we fine-tune this multi-channel frontend w.r.t. speaker verification objective. We successfully tackle the problem of the lack of a realistic multi-channel training set by using simulated data of MultiSV corpus. The analysis is performed on its retransmitted and simulated test parts. We achieve consistent improvements with a 2.7 times smaller model than the baseline based on a scheme with mask estimating NN.

Index Terms— Conv-TasNet, beamforming, embedding extractor, speaker verification, MultiSV

1. INTRODUCTION

In the past years, the demand has been steadily increasing for smart appliances that process users’ commands, such as hands-free devices, smart speakers, TVs, and home assistants. They are often equipped with several microphones. Multiple sensors have the ability to provide spatial information that is especially useful in adverse noisy and reverberant conditions. Multi-channel signal processing is currently being researched and applied in various fields, such as source separation, and speech enhancement, with applications to automatic speech recognition, keyword spotting and speaker verification (SV).

Since the advent of speaker embeddings extracted by neural networks [1], single-channel SV has progressed tremendously. To achieve desirable generalization, state-of-the-art embedding extractors require a large amount of training data such as Voxceleb [2, 3], often inflated by multiple augmentations [4]. Even though some multi-channel data-collecting initiatives have emerged [5, 6, 7], it is prohibitively costly and time demanding to collect a multi-channnel dataset comparable to single-channel Voxceleb in terms of amount of speakers and hours of speech.

When dealing with multi-channel data, there is experimental evidence [8] that even a strong embedding extractor can benefit from multi-channel enhancement [9, 10, 11] more than from a single-channel one [12]. The type of speech enhancement, however, plays an important role: while various multi-channel enhancement and source separating networks directly predicting sources have been developed [13, 14, 15], the non-linear distortions and artifacts may hurt the performance of a downstream speaker verification [16] or even ASR [17]. A plausible option for multi-channel pre-processing is linear filtering performed by a beamformer. It turned out to be beneficial in multiple studies [18, 19, 11].

Successful speech enhancement [20] and separation models [20, 21, 22] were previously based on a prediction of time-frequency (T-F) spectral *masks* (representing the dominance of a desired signal in T-F bins) applied to the input spectrum. In accordance with their success, a traditional way of neural beamforming likewise utilizes masks [19, 23, 11]. In such scenarios, pooled per-channel masks are used for the estimation of second-order statistics (spatial covariance matrices – SCMs).

A trend to estimate enhanced/separated speech signals directly in either frequency [24] or time domains [25, 26, 27] has emerged, too. Whereas frequency domain used to dominate the field, recent studies increasingly employ time-signal processing. Following the direct prediction of desired outputs, some studies have also utilized such models for the estimation of SCMs required by beamforming in multi-channel settings [28, 29].

We propose a multi-channel speaker embedding extraction model for SV in far-field conditions with background noise and reverberation. Similarly to our baseline [11], the final model comprises neural-network-supported beamforming and single-channel embedding extractor. We focus on the multi-channel part of the system (beamforming). For it to enhance speech and suppress noise well, precise estimation of SCMs is required. To this end, motivated by the discussed trends, we depart from the *mask predictor* (used in the baseline). Instead, we aim at utilizing the speech modeling power of recent models. Similarly to [29], we base the SCM estimation on per-channel outputs of a time-domain model from the field of source separation. We opt for Conv-TasNet [25] for its modeling power and small size which we even decrease to obtain a small footprint network. Our contributions can be summarized as follows:

- Despite the utilization of Conv-TasNet, we propose an SCM estimation approach that is different from that used in [29]. We empirically found it performing consistently better in our task of SV, especially for retransmitted data and ad-hoc microphone arrays.
- We show significant improvements on simulated evaluation data and comparable or better performance on retransmitted data compared to the *mask predictor* based model, which has 2.7 times more parameters than the proposed network.
- We show the efficiency of the front-end (Conv-TasNet or mask predictor) fine-tuning in a joint model by optimizing speaker-discriminative loss. The fine-tuning brought average relative improvements of 8.0% and 6.4% for proposed and *mask predictor* based models, respectively.

The work was supported by Czech Ministry of Interior project No. VJ01010108 ”ROZKAZ”, Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, Czech Ministry of Education, Youth and Sports from project no. LTAIN19087 ”Multi-linguality in speech technologies”, and Horizon 2020 Marie Sklodowska-Curie grant ESPERANTO, No. 101007666. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project ”e-Infrastructure CZ – LM2018140”. Authors would like to thank Tsubasa Ochiai and Marc Delcroix from NTT for useful discussion.

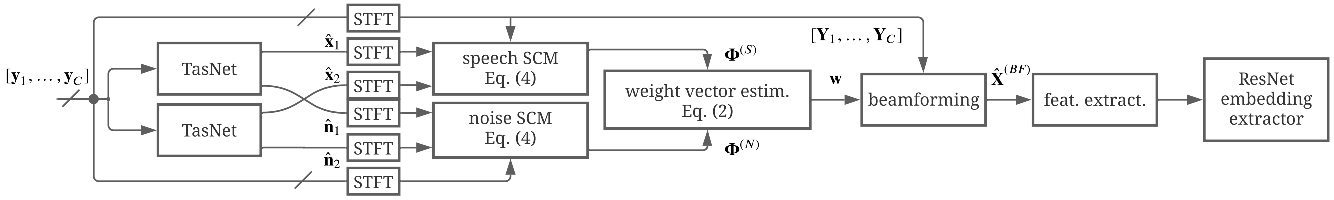

Fig. 1: Proposed multi-channel embedding extraction for SV based on beamforming and Conv-TasNet enhancement.

2. METHOD

As displayed in Figure 1, the architecture comprises two main components: beamforming-based front-end and speaker embedding extraction. At the first stage of front-end processing, four channels from a microphone array are independently fed to a per-channel operating Conv-TasNet [25]. It has been proven strong in source separation, and we aim at employing its modeling capacity in our scenario. We utilize it to split single-channel input mixture to corresponding speech and noise components. Resulting four estimates of speech and noise are subsequently used to estimate speech and noise SCMs, respectively. It is the quality of SCMs that affects beamformer speech enhancing ability. Any beamformer that makes use of SCMs, such as MVDR (minimum variance distortionless response) [30] or GEV (generalized eigenvalue) [31], may be used at this stage. We perform our analysis with the MVDR. Single-channel beamformer output is subject to feature extraction and passed to a ResNet-based embedding extractor. Finally, the speaker embeddings are compared with cosine similarity to yield SV scores.

2.1. Conv-TasNet for speech enhancement

The original Conv-TasNet [25] is a neural network designed to separate a time-domain mixture $\mathbf{y} \in \mathbb{R}^T$, with T representing the number of samples, into individual source signals: given that $\mathbf{y} = \sum_{i=1}^{I} \mathbf{x}_i$, the model should output predictions of I source signals $\{\hat{\mathbf{x}}_i\}_{i=1}^{I}$ in a non-specified order. To achieve this goal, it is trained to optimize a scale-invariant signal-to-distortion ratio (SI-SDR) in an utterance-level permutation-invariant training (PIT) manner [22].

Although Conv-TasNet was originally designed for speech source separation, we adapt it to separate single-speaker speech from an additive noise. Therefore, it can be viewed as a speech enhancement model applied to every channel $c = 1, \ldots, C$ of a microphone array.

We assume that every channel signal $\mathbf{y}_c \in \mathbb{R}^T$ can be decomposed into components $\mathbf{x}_c$ and $\mathbf{n}_c$. Component $\mathbf{x}_c$ is a time-domain speech signal impinging on the microphone c (reverberant speech), and $\mathbf{n}_c$ is a reverberant noise signal. Our speech enhancing model is supposed to separate the speech component from the noise component: $[\hat{\mathbf{x}}_c, \hat{\mathbf{n}}_c]^T = \text{Conv-TasNet}_{\text{enh}}(\mathbf{y}_c)$.

In summary, given multi-channel mixture signals $\{\mathbf{y}_c\}_{c=1}^{C}$, the Conv-TasNet-based speech enhancer is used to obtain estimates of speech $\{\hat{\mathbf{x}}_c\}_{c=1}^{C}$ and estimates of noise $\{\hat{\mathbf{n}}_c\}_{c=1}^{C}$ for each channel independently as schematically displayed in Figure 1.

2.2. Beamforming in frequency domain

In order to perform beamforming in the frequency domain, time-domain inputs need to be transformed employing short-time Fourier transform (STFT). Let $\mathbf{Y}_c \in \mathbb{C}^{\mathcal{T} \times F}$ be an STFT representation of $\mathbf{y}_c$, where F is the number of frequency bins and $\mathcal{T}$ represents the number of frames. For every frame index t and frequency index f, we then construct a vector $\mathbf{Y}_{t,f} = [Y_{t,f,c=1}, \ldots, Y_{t,f,c=C}]^T \in \mathbb{C}^C$ grouping channels together, where $Y_{t,f,c} \in \mathbb{C}$ is a complex value of a time-frequency bin of channel c. A beamformer performs linear filtering, therefore, the enhanced STFT value $\hat{X}_{t,f}^{(BF)}$ is obtained as

$$\hat{X}_{t,f}^{(BF)} = \mathbf{w}_f^H \mathbf{Y}_{t,f}, \tag{1}$$

where $\mathbf{w}_f \in \mathbb{C}^C$ is a time-independent beamforming weight vector and $(\cdot)^H$ represents the conjugate (or Hermitian) transpose. The way $\mathbf{w}_f$ is computed depends on the criterion that the beamformer optimizes. In this study, we adopt the MVDR beamformer. Subjected to a unity gain constraint in the desired direction (distortion-less constraint), the MVDR minimizes the power of the output. The minimization leads to a closed-form solution. We use a formulation with SCMs [32]:

$$\mathbf{w}_f^{\text{MVDR}} = \frac{\left(\mathbf{\Phi}_f^{(N)}\right)^{-1} \mathbf{\Phi}_f^{(S)}}{\text{Tr}\left(\left(\mathbf{\Phi}_f^{(N)}\right)^{-1} \mathbf{\Phi}_f^{(S)}\right)} \mathbf{u}, \tag{2}$$

where $\mathbf{\Phi}_f^{(S)}$ is a frequency-dependent SCM of speech, $\mathbf{\Phi}_f^{(N)}$ is the SCM of noise, and $\mathbf{u} \in \{0,1\}^C$ is a one-hot vector encoding the reference microphone. We always aim at obtaining enhanced speech at the first microphone, hence we keep $\mathbf{u} = [1, 0, \ldots, 0]^T$ constant.

2.3. Spatial covariance matrix estimation

Let $\hat{\mathbf{X}}_c \in \mathbb{C}^{\mathcal{T} \times F}$ and $\hat{\mathbf{N}}_c \in \mathbb{C}^{\mathcal{T} \times F}$ be the STFT representations of Conv-TasNet enhanced speech signal $\hat{\mathbf{x}}_c$ and the estimated noise $\hat{\mathbf{n}}_c$, respectively. We propose an approach to SCM estimation that is based on input masking rather than using $\hat{\mathbf{x}}_c$, $\hat{\mathbf{n}}_c$ directly. Our assumption is that it has the potential to neglect wrong predictions for some channels. Using Conv-TasNet estimates, we first compute channel-dependent real ratio masks resembling ideal ratio masks (IRM) [20]

$$\mathbf{M}_c^{(S)} = \left(\frac{|\hat{\mathbf{X}}_c|^2}{|\hat{\mathbf{X}}_c|^2 + |\hat{\mathbf{N}}_c|^2}\right)^{\beta}, \; \mathbf{M}_c^{(N)} = \left(\frac{|\hat{\mathbf{N}}_c|^2}{|\hat{\mathbf{X}}_c|^2 + |\hat{\mathbf{N}}_c|^2}\right)^{\beta}, \tag{3}$$

where β is a scalar set to 0.5 since we empirically found it performing the best. Vectors $\hat{\mathbf{X}}_{t,f} = [\hat{X}_{t,f,c=1}, \ldots, \hat{X}_{t,f,c=C}]^T \in \mathbb{C}^C$ and $\hat{\mathbf{N}}_{t,f} = [\hat{N}_{t,f,c=1}, \ldots, \hat{N}_{t,f,c=C}]^T \in \mathbb{C}^C$ are constructed from time-freqency-channel bins of STFTs $\{\hat{\mathbf{X}}_c\}_{c=1}^{C}$ and $\{\hat{\mathbf{N}}_c\}_{c=1}^{C}$, respectively. Following [33], we combine channel-dependent masks using product pooling $\mathbf{M}^{(\nu)} = \prod_{c=1}^{C} \mathbf{M}_c^{(\nu)}$, with ν being either S or N. Values of $\mathbf{M}^{(\nu)} \in \mathbb{R}^{\mathcal{T} \times F}$ represent the prevalence of speech or noise in time-frequency bins of Conv-TasNet outputs. Finally, values of combined masks weigh the outer products when estimating SCMs:

$$\mathbf{\Phi}_f^{(\nu)} := \frac{\sum_{t=1}^{\mathcal{T}} M_{t,f}^{(\nu)} \mathbf{Y}_{t,f} \mathbf{Y}_{t,f}^H}{\sum_{t=1}^{\mathcal{T}} M_{t,f}^{(\nu)}}. \tag{4}$$

2.4. Speaker embedding extractor

Speaker embedding network extracts utterance-level embeddings given enhanced single-channel audio (1). Due to frame length (64 ms vs. 25 ms) and shift (16 ms vs. 10 ms) mismatch between the beamforming front-end and embedding extractor, the signals are re-framed through the time domain. Subsequently, log-Mel filter bank energy features (*fbanks*) are extracted and fed to the extractor.

The embedding extractor architecture is based on ResNet34 [34], with slight modifications described in Section 3.3. Following the x-vector extractor approach [1], outputs of the last residual block are subject to statistical pooling. The obtained statistics are projected to speaker-descriptive 256-dimensional embedding. The model is trained to optimize additive margin (AM) softmax [35].

3. EXPERIMENTAL SETUP

3.1. Training data

We use two training datasets – one for embedding extractor and one for front-end. The single-channel embedding extractor is trained on the development part of Voxceleb 2 [3] with reverberation and noise augmentations defined by the Kaldi recipe.

Since front-end models (Conv-TasNet and mask predictor) require speech and noise references, we use training data from the MultiSV[1] corpus [36] which we have recently released. It comprises simulated four-channel reverberated and noisy training recordings. It comes with reverberant speech and reverberant noise references for training as well as with speaker labels. Speech signals of 1,000 speakers that enter the simulation were selected from Voxceleb 2 dev recordings exceeding 20 dB SNR. Distractors cover various domains – music from FMA small [37], noises (without music and babble) from MUSAN [38], and selection of noises from Freesound.org (fan, HVAC, office sounds, etc). They were added to speech signals with SNRs uniformly sampled from [3, 20] dB. Room impulse responses were created by the image source method (ISM) [39], and RT60 ranges from 0.3 to 0.9 s.

Since front-end models process signals independently, the multichannel nature of MultiSV is disregarded. However, it is utilized in joint training, which will be detailed later, as all four channels are used in beamforming whose output is fed to the embedding extractor.

3.2. Evaluation multi-channel data

We use trial sets defined by MultiSV for evaluation. They are based on trials for a single-channel VOiCES challenge [40] where the data are selected from the VOiCES corpus [5]. Modification enabling multi-channel SV evaluation was presented in [41]. MultiSV extends it by adding other conditions. We adopt the scenario of single-channel enrollment and multi-channel test. For a thorough evaluation, we employ both simulated and retransmitted data. They are equal in terms of source speech originating from LibriSpeech [42].

Following [36], the retransmitted evaluation sets are referred to as *dev retr SRE*, *eval retr v1 SRE*, and *eval retr v2 SRE*. Enrollment single-channel recordings are the same as those used in the VOiCES challenge [40] – reverberant (dev) or combination of reverberant and clean (eval). Test segments are four-channel recordings of ad-hoc microphone arrays. They contain retransmitted music (only in *dev retr*), television, babble, and none (diffuse background) noises. Versions of the "eval" set differ in employed microphones. *V1* comprises arrays with large inter-microphone distance where distant sensors might have very low SNR. *V2* contains more compact arrays. Trial definitions *dev retr CE* and *eval retr CE* differ from their SRE counterparts only in terms of clarity of enrollment segments. CE stands for "clean enrollment". The development set comprises 196 speakers and 996,448 trials (with 5,024 target ones). The evaluation set comprises 100 different speakers and 973,929 trials (with 9,939 target ones).

Simulated trial sets are labeled as *dev simu* and *eval simu*. Microphone array recordings were obtained by simulation using source speech signals. Background noises contain MUSAN music and noises (not present in training data), and distractors from Freesound.org. RT60 reverberation times of ISM-generated RIRs were uniformly drawn from the interval $[0.3, 0.9]$ s. Mixing SNRs were uniformly drawn from $[3, 20]$ dB.

3.3. Models and hyperparameters

Original best-performing Conv-TasNet model in [25] comprises 5.1M parameters. To obtain more practical small-footprint model of 1.2M parameters, we altered hyperparameters (respecting the original naming convention) as follows: N = 256, L = 40, B = 128, H = 192, P = 3, X = 7, and R = 3. The rationale behind the increased length of filters (L) in the encoder and decoder is to keep approximately the same time span of bases while using data sampled at 16 kHz (compared to 8 kHz). The original Conv-TasNet for source separation is trained with SI-SDR objective in a PIT fashion. In our case, sources (speech or noise) are well-defined, therefore PIT is not required. As opposed to SI-SDR, we optimize SNR. It is because scale invariance in training could cause a different dynamic range of enhanced speech and noise. As noted in [29], preserved scale information is important for correct subsequent SCM estimation.

ResNet-based [34] embedding extractor requires 40D *fbank* features. Its stages comprise conv. layers of 64, 128, 256, 256 channels, and the last part of the model is tailored towards embedding prediction. The scale of the AM softmax loss [35] was set to 30. The margin was continuously increased during training up to 0.2.

4. EXPERIMENTS

We present SV results in terms of equal error rate (EER [%]) and minimum detection cost (MinDCF), where the prior probability of a target trial P_{tar} is set to 0.01 following the VOiCES challenge [40].

4.1. Baseline

The baseline comprises frequency domain *mask predictor* [11] and the same embedding extractor as the proposed model. Given the magnitude spectra at the input, the front-end directly estimates per-channel speech and noise *masks* using NN (as opposed to speech and noise signals). They are combined by averaging and used for SCM estimation according to (4). The mask predictor is trained to optimize a binary cross-entropy between outputs and ideal binary masks [20]. The model comprises a long short-term memory (LSTM) layer (providing outputs of the same dimensionality as inputs, i.e. 513 – number of frequency bins) followed by 2 fully connected (FC) layers with 513 neurons and two parallel FC layers predicting masks. It is noteworthy that the results obtained with the baseline are not readily comparable with results in [11]. The reasons are as follows: following the evolution in SV, we switched our embedding extractor from simpler time delay NN model (TDNN) with cross-entropy to ResNet with the AM loss. We also employ cosine-similarity scoring instead of PLDA. Training data has changed as well.

[1] https://github.com/BUTSpeechFIT/MultiSV

Table 1: Evaluation on trials with retransmitted multi-channel test segments. Enrollment utterances are retransmitted (SRE) or clean (CE).

Front end	train type	params	dev retr SRE		eval retr v1 SRE		eval retr v2 SRE		dev retr CE		eval retr CE	
			EER [%]	MinDCF	EER [%]	MinDCF	EER [%]	MinDCF	EER [%]	MinDCF	EER [%]	MinDCF
mask predictor	sep.	3.2M	0.98	0.124	4.47	0.354	2.04	0.197	0.92	0.111	4.25	0.326
proposed		1.2M	0.97	0.119	4.39	0.346	2.08	0.188	0.90	0.104	4.24	0.330
mask predictor	joint	3.2M	0.86	0.109	4.38	0.329	2.22	0.182	0.84	0.087	4.32	0.315
proposed		1.2M	**0.82**	**0.108**	**4.26**	**0.324**	**1.98**	**0.176**	**0.80**	**0.085**	**4.20**	**0.314**

Table 2: Proof-of-concept experiments on simulated data.

Front end	dev simu		eval simu	
	EER [%]	MinDCF	EER [%]	MinDCF
mask predictor	1.41	0.162	2.02	0.195
propopsed	**1.17**	**0.147**	**1.91**	**0.176**

4.2. Proof of concept

The first experiment aims to evaluate Conv-TasNet's ability to provide good representations of speech and noise, leading to better estimation of SCMs in a controlled environment. To this end, we opted for an evaluation data that is similar to the training data – both are simulated. However, differences in speakers, speaking styles (read and spontaneous speech), utterances, and noises make the evaluation sets challenging. As per results in Table 2, we observe consistent improvements in all metrics over the baseline with the proposed front-end. Relative improvements range from 5.4% to 21.1%.

4.3. Retransmitted evaluation data

We provide results obtained on more realistic retransmitted data. Such an evaluation is often missing, but we find it important. Conducted experiments assess the ability to generalize to unseen acoustic conditions and background noises. Television and babble noises are new to the system (although, they might share some properties with music noise to some extent). What also makes the data difficult is the fact that the assumptions that hold for simulated training data do not need to hold for retransmitted evaluation data. Noise and source additivity, and ray acoustics are examples of the assumptions.

According to the results in Table 1, retransmitted data poses a challenge to both systems, and the clear dominance of Conv-TasNet based system from the previous experiment has diminished. On the other hand, despite a significantly smaller size of the proposed model, it is still able to perform on par with or slightly better than the mask-predicting model. Comparing *dev retr SRE* and *CE* (its clean-enrollment counterpart) as well as *eval retr v1 SRE* with *eval retr CE*, the results suggest that, on average, speaker embeddings extracted from beamformed signals are more similar to embeddings of clean enrollment segments. It holds for both systems and is appealing as this scenario is practically useful. As expected, microphone arrays that are compact, and do not suffer from outlier signals of bad quality, provide considerably better results. This outcome stems from the comparison of *eval retr v1 SRE* and *eval retr v2 SRE*.

4.4. Joint model fine-tuning

Since loss functions optimized by speech enhancement models might not be directly related to the quality of SCM estimation, and in turn to speaker embeddings, we also propose a front-end fine-tuning scheme. We join both separately trained models and optimize speaker-discriminative AM loss. This is enabled because all components (including beamforming) of the models are differentiable. During the fine-tuning phase, the ResNet extractor is fixed, and only the front-end parameters are updated. We re-utilize data used for separate front-end training, MultiSV, as it also includes speaker labels. Prior to this fine-tuning, the last layer must have been redefined and trained since the number of speakers differs from the number of speakers in the embedding extractor training set.

The lower part of Table 1 shows considerable improvements by fine-tuning. Our Conv-TasNet based system benefits from this phase more as the average relative improvement is 8.0% compared to 6.4% for the mask predictor based model. Overall, our model yields the best numbers across the board. It suggests its power and generality.

4.5. Discussion

Apart from the presented approach to SCMs estimation, we also experimented with a more straightforward method directly utilizing Conv-TasNet outputs to estimate speech and noise SCMs similarly to [29], i.e. without masking. Even though we obtained 1.23% EER, 0.155 MinDCF on *dev simu*, and 1.99% EER, 0.179 MinDCF on *eval simu*, the proposed approach consistently outperforms it, especially on retransmitted data. It is worth noting that retransmitted data was not explored in [29].

We also explored the GEV beamformer and we observed similar outcomes. Owing to the space constraints, we only present results with the MVDR beamformer.

In this study, Conv-TasNet enhances individual channels independently. Utilization of multiple channels to enhance one of them might be helpful. A simple concatenation of channels did not improve results in our preliminary experiments.

It is noteworthy that the data used for fine-tuning is simulated. During this phase, speaker labels suffice, and no speech-noise decomposition is required. We hypothesize that real data could boost performance even more.

5. CONCLUSIONS

We proposed a new multi-channel speaker embedding extractor for far-field SV with noise and reverberation. It consists of NN-boosted beamforming and a ResNet-based extractor. We proposed to use Conv-TasNet for speech enhancement and a beamforming-related SCM prediction based on its outputs, which is different from that used in [29]. We obtained significant improvements on simulated MultiSV. Analysis on more difficult retransmitted MultiSV shows comparable or better performance over the baseline while using 2.7 times less parameters for front-end. We also demonstrate the effectiveness of a Conv-TasNet fine-tuning in a joint model.

As part of future work, we aim to analyze fine-tuning of both models simultaneously. We will also explore a better way of making use of multi-channel information during the enhancement phase.

6. REFERENCES

[1] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Proc. Interspeech*, 2017.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a Large-scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017.

[3] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018.

[4] A. Nagrani, J. S. Chung, J. Huh, et al., "VoxSRC 2020: The Second VoxCeleb Speaker Recognition Challenge," *arXiv e-prints*, p. 2012.06867, 2020.

[5] C. Richey, M. A. Barrios, Z. Armstrong, et al., "Voices Obscured in Complex Environmental Settings (VOICES) Corpus," *arXiv e-prints*, p. arXiv:1804.05053, Apr 2018.

[6] D. Garcia-Romero, D. Snyder, S. Watanabe, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker Recognition Benchmark Using the CHiME-5 Corpus," in *Proc. Interspeech*, 2019.

[7] X. Qin, H. Bu, and M. Li, "HI-MIA: A Far-Field Text-Dependent Speaker Verification Database and the Baselines," in *ICASSP*, 2020.

[8] H. Taherian, Z. Wang, J. Chang, and D. Wang, "Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, 2020.

[9] H. Taherian, Z.-Qiu Wang, and D. Wang, "Deep Learning Based Multi-Channel Speaker Recognition in Noisy and Reverberant Environments," in *Proc. Interspeech*, 2019.

[10] J.-Y. Yang and J.-H. Chang, "Joint Optimization of Neural Acoustic Beamforming and Dereverberation with x-Vectors for Robust Speaker Verification," in *Proc. Interspeech*, 2019.

[11] L. Mošner, O. Plchot, J. Rohdin, L. Burget, and J. Černocký, "Speaker Verification with Application-Aware Beamforming," in *Proc. ASRU*, 2019.

[12] S. Shon, H. Tang, and J. Glass, "VoiceID Loss: Speech Enhancement for Speaker Verification," in *Proc. Interspeech*, 2019.

[13] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation," in *Proc. Interspeech*, 2019.

[14] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On End-to-end Multi-channel Time Domain Speech Separation in Reverberant Environments," in *ICASSP*, 2020.

[15] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-Attention Dense U-Net for Multichannel Speech Enhancement," in *ICASSP*, 2020.

[16] S. Omid Sadjadi and J. H. L. Hansen, "Assessment of Single-Channel Speech Enhancement Techniques for Speaker Identification Under Mismatched Conditions ," in *Proc. Interspeech*, 2010.

[17] T. Yoshioka, N. Ito, M. Delcroix, et al., "The NTT CHiME-3 system: Advances in Speech Enhancement and Recognition for Mobile Multi-microphone Devices," in *Proc. ASRU*, 2015.

[18] J. Heymann, L. Drude, and R. Haeb-Umbach, "A Generic Neural Acoustic Beamforming Architecture for Robust Multi-channel Speech Processing," *Computer Speech & Language*, vol. 46, 2017.

[19] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *Proc. Interspeech*, 2016.

[20] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, 2018.

[21] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *ICASSP*, 2016.

[22] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 10, Oct. 2017.

[23] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural Network Based Spectral Mask Estimation for Acoustic Beamforming," in *ICASSP*, 2016.

[24] K. Tan and D. Wang, "Complex Spectral Mapping with a Convolutional Recurrent Network for Monaural Speech Enhancement," in *ICASSP*, 2019.

[25] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, 2019.

[26] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *ICASSP*, 2020.

[27] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP*, 2021.

[28] Z. Q. Wang and D. Wang, "Multi-Microphone Complex Spectral Mapping for Speech Dereverberation," in *ICASSP*, 2020.

[29] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain Audio Separation Network Meets Frequency-domain Beamformer," in *ICASSP*, 2020.

[30] J. Capon, "High-resolution Frequency-wavenumber Spectrum Analysis," *Proceedings of the IEEE*, vol. 57, no. 8, 1969.

[31] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, July 2007.

[32] M. Souden, J. Benesty, and S. Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, 2010.

[33] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint Training of Complex Ratio Mask Based Beamformer and Acoustic Model for Noise Robust ASR," in *ICASSP*, 2019.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive Margin Softmax for Face Verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, 2018.

[36] L. Mošner, O. Plchot, L. Burget, and J. Černocký, "MultiSV: Dataset for Far-Field Multi-Channel Speaker Verification," in *ICASSP*, 2022.

[37] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[38] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv e-prints*, p. arXiv:1510.08484v1, 2015.

[39] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, 1979.

[40] M. K. Nandwana, J. van Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The VOiCES from a Distance Challenge 2019 Evaluation Plan," *arXiv e-prints*, p. arXiv:1902.10828, Feb 2019.

[41] L. Mošner, O. Plchot, J. Rohdin, and J. Černocký, "Utilizing VOiCES Dataset for Multichannel Speaker Verification with Beamforming," in *Proc. Odyssey*, 2020.

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