ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095115

CONFORMER-BASED TARGET-SPEAKER AUTOMATIC SPEECH RECOGNITION FOR SINGLE-CHANNEL AUDIO

Yang Zhang, Krishna C. Puvvada*, Vitaly Lavrukhin, Boris Ginsburg*

NVIDIA, USA

ABSTRACT

We propose CONF-TSASR, a non-autoregressive end-to-end time-frequency domain architecture for single-channel target-speaker automatic speech recognition (TS-ASR). The model consists of a TitaNet based speaker embedding module, a Conformer based masking as well as ASR modules. These modules are jointly optimized to transcribe a target-speaker, while ignoring speech from other speakers. For training we use Connectionist Temporal Classification (CTC) loss and introduce a scale-invariant spectrogram reconstruction loss to encourage the model better separate the target-speaker's spectrogram from mixture. We obtain state-of-the-art target-speaker word error rate (TS-WER) on WSJ0-2mix-extr (4.2%). Further, we report for the first time TS-WER on WSJ0-3mix-extr (12.4%), LibriSpeech2Mix (4.2%) and LibriSpeech3Mix (7.6%) datasets, establishing new benchmarks for TS-ASR. The proposed model will be open-sourced through NVIDIA NeMo toolkit.

***Index Terms*—** Target-speaker ASR, Conformer, multi-speaker ASR, source separation

1. INTRODUCTION

Target-speaker automatic speech recognition (TS-ASR) is the task to transcribe a specific speaker's speech in an overlapping multi-speaker environment given the speaker's profile - an auxiliary utterance (Fig. 1). Along with blind source separation (BSS) and multi-speaker ASR, TS-ASR constitutes a class of approaches for overlapped speech recognition.

BSS methods separate individual components from a speech mixture in time-domain [1, 2] which are passed on to a single-speaker ASR model for transcription as a second step. As the separation step of BSS is not optimized for ASR, this can be sub-optimal. Multi-speaker ASR approaches [3, 4, 5, 6] and their speaker-attributed variants (SA-ASR) [7, 8] generate transcripts as output and are optimized end-to-end for ASR. A characteristic of BSS models and their analogous multi-speaker ASR models is their multiple output branches, one per source. SA-ASR requires profiles of all speakers in a mixed utterance as auxiliary information.

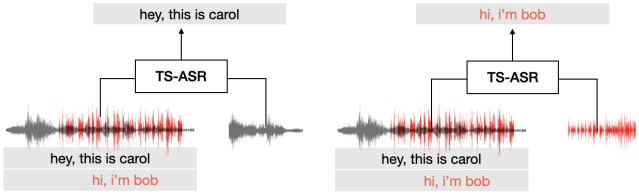

Fig. 1. Target-speaker ASR transcribes a specific speaker's part in a mixed utterance based on their clean speech sample (auxiliary utterance).

These seemingly similar approaches for overlapping speech recognition come with their own set of pros and cons. While BSS and their analogous multi-speaker ASR approaches do not require any auxiliary information, their major limitations include predefined number of output streams, permutation invariant training (PIT) [9] and speaker-tracing [10] for long audio inference. Further, having a different number of speakers during training and inference can greatly reduce their performance. In case of multi-speaker ASR, serialized output training (SOT) can overcome some of these limitations, but leaves much to be desired for in terms of performance [5]. In conjunction with SOT, SA-ASR uses speaker profile information to improve performance and not be limited by fixed number of outputs. But, it assumes availability of profiles for *all* speakers in an utterance. Nonetheless, it is well suited for transcribing meeting like scenarios. TS-ASR [10, 11], on the other hand, requires only one speaker profile of interest. It is apt for situations that require transcribing one target-speaker while ignoring interfering speakers. By design, TS-ASR doesn't suffer from permutation ambiguity and speaker-tracing. However, it requires one inference per speaker if used to transcribe multiple speakers.

In this paper, we propose Conformer-based TS-ASR model (CONF-TSASR) to address single-channel target-speaker ASR. Our approach adopts the SpeakerBeam [10] architecture and makes the following contributions:

- Improve SpeakerBeam using TitaNet [12] and Conformer [13] modules trained in an end-to-end fashion with CTC and novel spectrogram reconstruction loss.
- Achieve state-of-the-art target-speaker WER (TS-

*Equal contribution.

WER) on WSJ0-2mix-extr[2] and report results on WSJ0-3mix-extr and LibriSpeechMix [14] for the first time.

- Study the effects of target-speaker SNR and length of auxiliary utterance on model performance.

2. CONFORMER-BASED TS-ASR ARCHITECTURE

The proposed single-channel CONF-TSASR model consists of three modules - TitaNet, MaskNet and an ASR module (Fig. 2). It takes two inputs - a mixed utterance and a clean auxiliary utterance from the target-speaker and transcribes only the target-speaker's speech from the mixed utterance. The auxiliary utterance is encoded into a 192-dim speaker embedding by TitaNet [12] – a speaker embedding extractor model based on ContextNet [15]. From the mixed utterance 80-dim log-Mel features are extracted every 10msec over a window of 25msec. These are further perturbed with SpecAugment [16] and sub-sampled by 4x using two convolutional layers. MaskNet takes the sub-sampled features (S_{mix}) and speaker embedding to produce a time-frequency mask which is multiplied with S_{mix} to estimate the target-speaker's time-frequency features ($\hat{S}_t$). Finally, a Conformer [13] ASR module is used to transcribe the target speaker's speech using $\hat{S}_t$. The entire model is optimized using CTC [17] loss and spectrogram reconstruction loss. The latter computes scale invariant SiSNR [18] between an upsampled $\hat{S}_t$ – the estimated spectrogram – and true spectrogram S_t. The spectrogram reconstruction loss is reserved for training where you have access to the individual sources.

In our experiments, both MaskNet and ASR module consist of 18 Conformer layers, each with a hidden dimension of 256 and feed-forward dimension of 1024. Multi-head attention consists of 4 heads and the kernel size of the convolutional module is 31. The speaker embedding is linearly projected to match MaskNet's hidden dimension of 256 and added to the input of every Conformer block. Both ASR module and TitaNet are initialized with pre-trained weights available in NVIDIA NeMo toolkit[1]. For TitaNet, we freeze its ContextNet encoder and only train the decoder. The CONF-TSASR model has 66.1M trainable parameters and 85.4M parameters in total including the frozen TitaNet encoder.

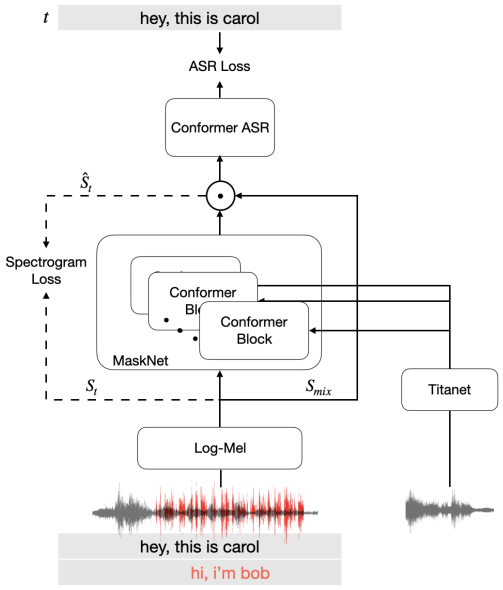

Fig. 2. Conformer-based CONF-TSASR model architecture. Feature extraction creates log-Mel spectrogram S_{mix} of the mixed utterance. A speaker embedding is extracted from an auxiliary utterance using TitaNet. The masking network learns a time-frequency mask for the target-speaker. The model is trained using CTC-loss and spectrogram reconstruction loss using the target-speaker's individual spectrogram S_t.

3. EXPERIMENTS

3.1. Datasets

We evaluated the proposed approach using two and three speaker mixtures created from WJS0 [19] and Librispeech [20] datasets following [21] and [14] respectively. To adapt these mixture datasets for TS-ASR, we augment them with a random auxiliary utterance from the target-speaker[2]. Briefly, a training example for two-speaker mixture was created by first randomly selecting two speakers and choosing one as the target-speaker. Among all utterances by the target-speaker two are chosen, one for creating the auxiliary utterance and one for the mixed utterance. To create the mixture we pick an utterance spoken by the other speaker. For WSJ0 mixtures, utterances were combined at a SNR uniformly chosen between 0 and 5 dB for each mixture [21]. For Librispeech mixtures, chosen utterances were combined without changing SNR to be consistent with [14]. For WSJ0 mixtures, the shorter utterance is both prepended and appended with random length of silence to match the length of the longest utterance in the mixture. In contrast, Librispeech mixtures are generated following [14], where the utterances were shifted by random delays before being added to the mixture. Delay values were chosen under the constraint that the start times of each utterance differed by 0.5sec or longer. Further, we augment the training data using speed and volume perturbation. For speed perturbation [22], the speed of each individual utterance is modified with a probability of 0.3 from its original rate to 95%, 97.5%, 100%, 102.5% or 105% before mixing. Volume perturbation [14] involves scaling the final mix-

[1] https://github.com/NVIDIA/NeMo

[2] https://github.com/xuchenglin28/speaker_extraction

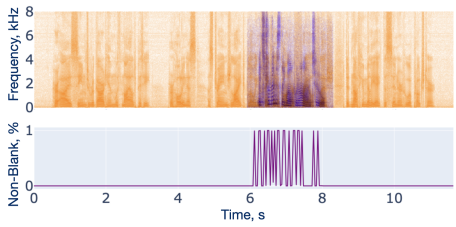

Fig. 3. A LibriSpeech2mix example. (top) Input mixture spectrogram. Target-speaker and overlapping speaker are shown in different colors. (bottom) Time-aligned non-blank emission probabilities for target-speaker using the proposed model. (Best viewed in color.)

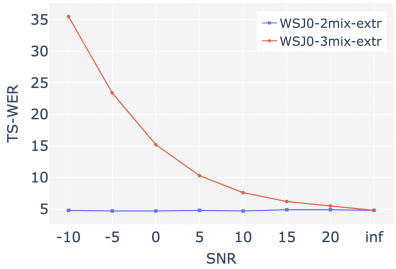

Fig. 4. TS-WER on WSJ0-2mix-extr and WSJ0-3mix-extr under different SNRs between target-speaker and interfering speakers using CONF-TSASR.

ture's volume by a random factor sampled from $[0.125, 2.0]$. In the following, we refer to two and three speaker mixtures of WSJ0 as WSJ0-2mix-extr and WSJ0-3mix-extr respectively, whereas LibriSpeech2mix and LibriSpeech3mix denotes Librispeech mixtures.

3.2. Training Setup

For training, we used 16 V100-32GB GPUs with a global batch size of 64. The model was trained for 60K and 480K updates for WSJ0-mix-extr and LibriSpeechMix respectively. We used AdamW with a peak learning rate of $3 * 10^{-4}$ and 0.01 weight decay. We used 10K and 25K warmup steps respectively with Cosine annealing and a minimum learning rate of $1 * 10^{-6}$. The relative weights of losses, when more than one is used, are tuned based on validation TS-WER.

3.3. WSJ0-mix-extr results

Table 1 compares performance of the proposed CONF-TSASR model with contemporary results on the WSJ0-mix-extr datasets. For baselines, we include a conventional ASR model (Conformer-CTC), SpeakerBeam [10], Exformer [23] and Conditional-Conformer-CTC [6]. The first model was trained on single-speaker, the second and third were trained on two speakers and the last model was trained on up to three speakers. SpeakerBeam can be regarded as a TS-ASR model which directly generates transcription for target-speaker. Exformer is a state-of-the-art target-speaker source separation model based on SepFormer [2] and thus requires an additional step to transcribe the model output. We updated the Exformer architecture with recent advancements to facilitate fairer comparisons. Namely, we replaced the original pretrained embedding network in Exformer with the same pretrained TitaNet [12] that was used in the CONF-TSASR. We also matched model size of the SepFormer with masking network in CONF-TSASR. Note that the model used to transcribe Exformer output and the "conventional ASR" baseline are the same. This in-turn is same as the one used for initializing ASR block in CONF-TSASR, except the former is further fine-tuned on training partition of WSJ0 dataset. Conditional-Conformer-CTC is a multi-speaker ASR model that uses conditional speaker chain to transcribe every speaker subsequently.

As expected, the conventional ASR model trained on single-speaker data performs poorly on WSJ0-mix (36.7% WER on WSJ0-2mix), whereas SpeakerBeam and Exformer+Conformer-CTC show 30.6% and 13.2% TS-WER respectively. Conditional-Conformer-CTC reaches a WER of 19.9% on WSJ0-2mix and 34.3% on WSJ0-3mix. In comparison, CONF-TSASR trained on up to two-speaker mixtures reaches 4.8% TS-WER and 4.2% TS-WER with additional spectrogram reconstruction loss. Exformer results show optimizing for SiSNR does not necessarily give the best result for transcription. Also, shifting away from time-frequency domain to time-domain is not only unnecessary for target-speaker speech recognition but also decreases model efficiency due to more time steps. CONF-TSAR model reaches TS-WER of 4.8% on WSJ0-2mix-extr and 12.4% on WSJ0-3mix-extr when trained on up to three speakers, suggesting that the proposed model is able to transcribe target-speaker from single-channel input in spite of two distracting speakers. To our knowledge, this is the best TS-WER reported on WSJ0-2mix-extr and the first study to report TS-WER on WSJ0-3mix-extr.

Fig. 4 shows the sensitivity of CONF-TSASR w.r.t. SNR between target and overlapping speakers. We observe that the performance is more sensitive to SNR when the mixture contains three overlapping speakers compared to just two speakers.

3.4. LibriSpeechMix results

Tables 2 & 3 report TS-WER results for the first time on LibriSpeechMix datasets. Due to lack of previous TS-ASR results on LibriSpeechMix dataset in literature, we use SOT-Conformer-AED [7] and its streaming version t-SOT [8] as reference.[3] These are different, yet closely related multi-

[3]While [11] tackles TS-ASR, our work is not directly comparable to theirs as they focuses on streaming and evaluate on Japanese corpus.

Model	Params, M	N	Loss	Learn Embedding	2-mix	3-mix
CONF-TSASR	85	2	CTC	no	8.6	-
		2	CTC	yes	4.8	-
		2	CTC+Spec	yes	**4.2**	-
		3	CTC	yes	5.4	13.8
		3	CTC+Spec	yes	**4.8**	**12.4**
SpeakerBeam [10]	n/a	2	Cross Entropy	yes	30.6	-
Exformer [23] + Conformer-CTC	80	2	SiSNR, CTC	no	13.2	-
Conditional-Conformer-CTC [6]	n/a	3	CTC	yes	19.9**	34.3**
Conformer-CTC	29	1	CTC	no	36.7*	54*

Table 1. TS-WER of different models on WSJ0-2mix-extr and WSJ0-3mix-extr. The model was trained on N-maximum speakers in train mixture. * denotes best WER on individual transcript. ** denotes WER for multi-speaker ASR model.

Model	Params, M	L (sec)	2-mix
CONF-TSASR CTC	85	7.5	5.1
		15	4.6
CONF-TSASR CTC+Spec	85	7.5	4.5
		15	**4.2**
SOT-Confomer-AED [7]	129	15	4.9†
t-SOT TT-18 [8]	82	15	5.2‡
t-SOT TT-36 [8]	139	15	4.4‡

Table 2. TS-WER of CONF-TSASR, SA-WER†, and permutation invariant SA-WER‡ of related multi-speaker references on LibriSpeechMix, trained for up to 2 speakers. Notation: L - Length of auxiliary utterance

Model	L (sec)	2-mix	3-mix
CONF-TSASR CTC	7.5	7	9.7
	15	6	8.4
CONF-TSASR CTC+Spec	7.5	6.3	9
	15	**5.4**	**7.6**
SOT-Confomer-AED [7]	15	6.8†	9.6†
SOT-Confomer-AED [7] SD	15	**6.4**†	**8.5**†

Table 3. TS-WER of CONF-TSASR, and permutation invariant SA-WER‡ of related multi-speaker references on LibriSpeechMix. CONF-TSASR was trained for up to 3 speakers. Notation: L - length of auxiliary utterance, SD - speaker deduplication [7]

speaker transcription models. They are based on transformer encoder-decoder architecture and use SOT [5]. They differ with the proposed model in the following (non-exhaustive) ways. 1) They transcribe all speakers in a given utterance. 2) Have knowledge of speaker profiles for all possible speakers in a given utterance. 3) Use 10 auxiliary utterances during training and 2 during evaluation (each with avg. length of 7.5 sec) to calculate speaker profiles. 4) SOT-Conformer-AED reports speaker-attributed WER (SA-WER) [14] on LibriSpeech2Mix and LibriSpeech3Mix, while t-SOT reports permutation-invariant SA-WER [8] on LibriSpeech2Mix. In contrast, the proposed model 1) Transcribes only target-speaker, 2) Is not aware of profiles for other speakers in utterance, 3) Uses only one auxiliary utterance during training and two during evaluation to calculate speaker profiles, and 4) Reports WER only for target speaker (TS-WER).

To make the TS-WER results on LibriSpeechMix comparable to SA-WER, we transcribe all speakers in a mixed utterance with each speaker as target, one at a time, with CONF-TSASR model. CONF-TSASR trained on up to three speakers achieves 5.4% TS-WER on LibriSpeech2Mix and 7.6% on LibriSpeech3Mix (Table 3). When trained on only up to two-speaker mixtures, the performance improves to 4.2% TS-WER on LibriSpeech2Mix (Table 2). Both two and three-speaker results show that adding spectrogram loss (CTC+Spec) significantly outperforms using merely CTC loss. When evaluated using only one auxiliary utterance (7.5 sec) as speaker profile, the proposed model exhibits slight performance deterioration (e.g. 7.6% vs. 9% in Table 3) highlighting the importance of robust speaker profiles. Training model with CTC loss [17] provides an auxiliary benefit of obtaining time-aligned token output probabilities for target-speaker (Fig. 3, bottom).

4. CONCLUSION

We present CONF-TSASR, an end-to-end state-of-the-art single-channel target-speaker speech recognition model. It consists of three modules. TitaNet – extracts a speaker embedding from a target-speaker's auxiliary utterance. MaskNet – generates a time-frequency mask for a target-speaker using Conformer. ASR module – transcribes the masked speech features using Conformer. The model is trained with CTC and spectrogram loss. We obtain state-of-the-art results on WSJ0-2mix-extr and establish new benchmarks for WSJ0-3mix-extr and LibriSpeechMix datasets. The model can be both used for fully and partially overlapped speech, requires as little as one auxiliary utterance and is non-autoregressive. Model will be open-sourced through NVIDIA NeMo toolkit[4].

[4]https://github.com/NVIDIA/NeMo

5. REFERENCES

[1] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, 2019.

[2] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP*, 2021.

[3] Xuankai Chang, Yanmin Qian, Kai Yu, and Shinji Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," in *ICASSP*, 2019.

[4] Ilya Sklyar, Anna Piunova, and Yulan Liu, "Streaming multi-speaker asr with rnn-t," in *ICASSP*, 2021.

[5] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," *Interspeech*, 2020.

[6] Pengcheng Guo, Xuankai Chang, Shinji Watanabe, and Lei Xie, "Multi-speaker ASR combining non-autoregressive conformer CTC and conditional speaker chain," *Interspeech*, 2021.

[7] Naoyuki Kanda, Guoli Ye, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka, "End-to-end speaker-attributed ASR with transformer," *Interspeech*, 2021.

[8] Naoyuki Kanda, Jian Wu, Yu Wu, Xiong Xiao, Zhong Meng, Xiaofei Wang, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Takuya Yoshioka, "Streaming multi-talker ASR with token-level serialized output training," *Interspeech*, 2022.

[9] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[10] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černocký, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[11] Takafumi Moriya, Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, and Takahiro Shinozaki, "Streaming target-speaker asr with neural transducer," *Interspeech*, 2022.

[12] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg, "TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP*, 2022.

[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech*, 2020.

[14] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka, "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," *Interspeech*, 2020.

[15] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu, "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," *Interspeech*, 2020.

[16] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, 2019.

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[18] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *ICASSP*, 2019.

[19] John Garofolo, David Graff, Doug Paul, and David Pallett, "CSR-I (WSJ0) complete," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[21] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, 2016.

[22] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015.

[23] Zhepei Wang, Ritwik Giri, Shrikant Venkataramani, Umut Isik, Jean-Marc Valin, Paris Smaragdis, Mike Goodwin, and Arvindh Krishnaswamy, "Semi-supervised time domain target speaker extraction with attention," *arXiv:2206.09072*, 2022.