Interspeech 2024
1-5 September 2024, Kos, Greece

Transcription-Free Fine-Tuning of Speech Separation Models for Noisy and Reverberant Multi-Speaker Automatic Speech Recognition

William Ravenscroft[1,2*], *George Close*[1], *Stefan Goetze*[1], *Thomas Hain*[1], *Mohammad Soleymanpour*[2], *Anurag Chowdhury*[2], *Mark C. Fuhs*[2]

[1]Department of Computer Science, The University of Sheffield, United Kingdom
[2]Solventum, USA

jwravesncroft1@sheffield.ac.uk

Abstract

One solution to automatic speech recognition (ASR) of overlapping speakers is to separate speech and then perform ASR on the separated signals. Commonly, the separator produces artefacts which often degrade ASR performance. Addressing this issue typically requires reference transcriptions to jointly train the separation and ASR networks. This is often not viable for training on real-world in-domain audio where reference transcript information is not always available. This paper proposes a *transcription-free* method for joint training using only audio signals. The proposed method uses embedding differences of pre-trained ASR encoders as a loss with a proposed modification to permutation invariant training (PIT) called *guided* PIT (GPIT). The method achieves a 6.4% improvement in word error rate (WER) measures over a signal-level loss and also shows enhancement improvements in perceptual measures such as short-time objective intelligibility (STOI).

Index Terms: speech recognition, speech separation, multi-speaker, adaptation, fine-tuning

1. Introduction

While significant progress has been made in recent years in multi-speaker ASR research, it remains a challenging problem [1, 2]. Many different approaches have been proposed to solving this problem [3, 4, 5]. These methods include entirely end-to-end single models [4] and modular approaches [6, 2, 5].

Modular approaches often include a variety of sub-components such as speech separation [7], speaker diarization [8] and ASR [6, 5]. Sometimes these sub-components, or *modules*, can be combined in an end-to-end fashion that allows for joint training of the modules [9, 3]. More recently, serialized output training (SOT) was proposed, a technique that enables the design of explicitly multispeaker ASR models with a single output sequence containing subsequences separated by a unique token for each speaker [4]. There is a benefit to having a simplified model that requires no fine-tuning on sub-components but this results in lower interpretability. A commonality of both approaches to multi-speaker ASR is that they are reliant on the existence of ground truth transcriptions of the speech of each speaker in the mixture. Whilst this is easily obtainable for simulated mixtures, it is often impossible to obtain for ‘real-world’ in-domain speech mixtures.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. This work was also supported in part by Solventum and Toshiba Cambridge Research Laboratory.

* Work done while an intern at Solventum, formerly part of 3M.

In [10, 11], computing embedding differences between self-supervised speech representations (SSSRs) was used as a loss for training speech enhancement networks to improve perceptual speech quality. Furthermore, in [12] it was shown that computing the difference between *senone* representations of ASR models can be used to adapt speech enhancement networks for robust ASR. Recent work [13, 14] has shown that features derived from ASR models can capture quality and intelligibility-related information.

In this work, a novel method for transcription-free fine-tuning of a separate-and-recognize modular approach is proposed. Embedding differences from a pre-trained ASR encoder are used to compute a loss from the reference speech and separated speech signals. Furthermore, a variant of the the standard PIT [15] algorithm, guided permutation invariant training (GPIT) is proposed to properly address speaker permutation solving of the ASR encoder embeddings within the proposed loss function. This recognition-based loss term is then used to fine-tune a speech separator. The proposed method notably improves multi-speaker WER performance over traditional signal-level losses across multiple ASR models, and also shows improved performance in intrusive perceptual measures. Another benefit of the proposed approach is that it allows for fine-tuning separators using truncated audio signals [16] as the full transcription is not required, potentially resulting in accelerated training, lower memory requirements and reduced computational expenditure. The proposed method also makes training on real-world data, with pseudo-reference audio used as input to the proposed loss function, possible [17] .

The remainder of this paper proceeds as follows: in Section 2, the proposed transcription-free fine-tuning method is described in detail. In Section 3, the experimental setup and data are described. Results and conclusions are given in Section 4 and Section 5, respectively.

2. Transcription-Free Fine-Tuning Method

In this section, the main components of the proposed transcription-free fine-tuning method are described. A modularized approach to multi-speaker ASR is used; this approach is referred to here as the *separate-and-recognize* approach. A noisy reverberant mixture $x[i]$ of C speech signals $s_c[i]$, $c \in \{1, \ldots, C\}$ for time index i is defined as

$$x[i] = \sum_{c=1}^{C} h_c[i] * s_c[i] + \nu[i], \qquad (1)$$

where $*$ denotes the convolution operator, $h_c[i]$ is the room impulse response corresponding to speaker c and $\nu[i]$ denotes additive noise. In the separate-and-recognize approach, a sepa-

10.21437/Interspeech.2024-1264

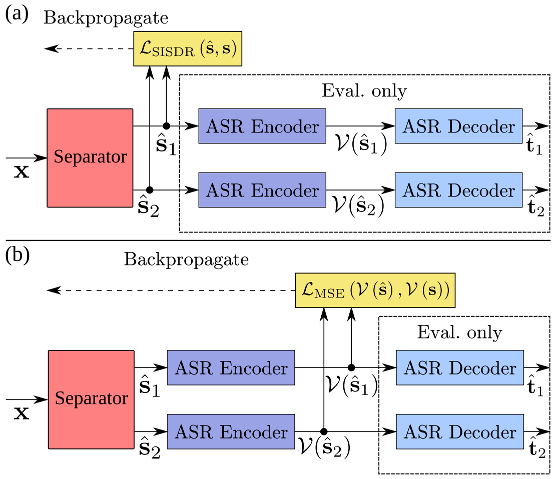

Figure 1: *(a) Baseline approach to training speech separators without ASR-based fine-tuning. (b) Proposed fine-tuning approach without using reference transcriptions. Solid lines indicate information flow; dashed lines the direction of gradient backpropagation. Figure exemplary for $C = 2$ speakers.*

ration model firstly separates the mixture signal $x[i]$ in C estimated speech signals $\hat{s}_c[i]$ and each separated signal is fed into an ASR module to attain predicted transcription of discrete character tokens $\hat{\mathbf{t}}_c$ [2].

The proposed fine-tuning method assumes the existence of pre-trained speech separation and ASR networks, where the speech separation model has been trained using a conventional signal-level objective function such as scale-invariant signal-to-distortion ratio (SISDR) [18]. In the proposed method, these two modules go through an additional number of addtional training epochs (ATEs), whereby the parameters of the ASR model are frozen and the embedding differences of the ASR encoder are backpropagated through the separation network where the parameters are updated at each step. The proposed approach is shown in Figure 1 compared to the baseline model trained with a purely signal-level SISDR loss.

2.1. ASR Encoder Loss

This section describes the proposed ASR encoder (AE) loss used for fine-tuning the speech separation model. In this work the ASR encoder used is a connectionist temporal classification (CTC)-based model [19] with a discrete number of possible output symbols N. The encoder network is defined as a function

$$\mathcal{V} : \mathbb{R}^{L_x} \mapsto \mathbb{R}^{L \times N} \tag{2}$$

where L_x is the length of an input speech signal $s_c[i]$, L is the output sequence length and N is the number of possible symbols, i.e. the number of characters that can be interpreted by a decoder or decoding function (minimum of 26 plus a word boundary token and a blank symbol for the Latin alphabet).

The proposed AE loss computes the difference between the encoder output of the predicted speech signals $\hat{s}_c$ and reference speech signal $s_c, \forall c \in \{1, \ldots, C\}$. The loss is the mean square error (MSE) of the two output sequences defined as

$$\mathcal{L}_{\text{AE}}(\mathbf{s}_c, \hat{\mathbf{s}}_c) = \frac{1}{LN} \sum_{\ell=1}^{L} \sum_{n=1}^{N} (\mathcal{V}(\hat{\mathbf{s}}_c)_{\ell,n} - \mathcal{V}(\mathbf{s}_c)_{\ell,n})^2 \tag{3}$$

where s_c is the target audio and $\hat{s}_c$ is the predicted output audio of the seperator for speaker c. The $\mathcal{L}_{\text{AE}}$ loss is analogous to a conventional time-frequency based loss, but instead of comparing references and outputs in terms of continuous time/frequency bins, the proposed loss instead compares in terms of discrete time/character label logits.

2.2. Guided Permutation Invariant Training (GPIT)

Empirically it was found that the standard PIT algorithm [20] was not robust enough against distortions in the ASR encoder features when applying the proposed AE loss function in (3), resulting in inaccuracies in resolving speaker permutations. This resulted in models trained using the AE loss diverging and giving 100% WER in evaluations. Thus, a modified scheme referred to as guided permutation invariant training (GPIT) (not to be confused with Graph-PIT [21]) is proposed that uses an alternate signal-level loss to guide the permutation solving, and then the AE loss is applied to the minimum permutation estimates and references. The permutation-solving formula is defined the same as for PIT. The predicted permutation $\hat{\phi}$ is determined by

$$\hat{\phi} = \arg\min_{\phi \in \boldsymbol{\Phi}} \sum_{c=1}^{C} \mathcal{L}_{\text{guide}}\left(\hat{\mathbf{s}}_c^{(\phi)}, \mathbf{s}_c\right) \tag{4}$$

where ϕ is a permutation in the set $\boldsymbol{\Phi}$ of $\Phi = C!$ possible permutations and $\hat{\mathbf{s}}_c^{(\phi)}$ denotes the ϕth possible permutation of $\hat{\mathbf{s}}_c$. $\mathcal{L}_{\text{guide}}$ denotes the *permutation guiding* loss. In this paper $\mathcal{L}_{\text{guide}} = \mathcal{L}_{\text{SISDR}}$ is chosen. Using the predicted permutation the desired loss, in this case $\mathcal{L}_{\text{AE}}(\cdot)$ from (3), is calculated s.t. the final loss value is

$$\mathcal{L}_{\text{GPIT}}\left(\mathbf{s}_c, \hat{\mathbf{s}}_c, \hat{\phi}\right) = \sum_{c=1}^{C} \mathcal{L}_{\text{AE}}\left(\mathbf{s}_c, \hat{\mathbf{s}}_c^{(\hat{\phi})}\right). \tag{5}$$

3. Experimental Setup

3.1. Data

The WHAMR [22] corpus is used for all experiments in this work. This corpus is an extension of the WSJ0-2Mix corpus [23] which in turn is an extension of the WSJ0 corpus [24]. WHAMR is a corpus of noisy, reverberant, two-speaker mixtures created from overlapping two artificially reverbed speaker utterances from WSJ0 at *mixing signal-to-noise ratios (SNRs)* of 0-5 dB and then adding an ambient noise source at an SNR of −6-3 dB relative to the loudest speaker. The 16kHz *max* (non-truncated) configuration of the corpus is used. Dynamic mixing [25] is also used, whereby the training set is uniquely simulated at each epoch to improve data diversity and model generalization. Signal lengths are also limited to 4s in length to speed up training as well as reduce computational expenditure and memory consumption [16]. Further analysis of training signal length (TSL) limits is given in Section 4.4.

3.2. Speech Separator

The TD-Conformer-XL model is chosen for speech separation as it gives close to state-of-the-art performance on the WHAMR benchmark and is open source[1] [26]. The configuration is based on the best-performing model on WHAMR in [26] with $S = 1$ subsampling layers, a kernel size of $P = 64$ and feature dimension $B = 1024$. To compensate for the larger sampling rate used in this paper (8kHz vs 16kHz here), the number of subsampling layers is set to $S = 2$. This model is *pre-trained* on

[1] TD-Conformer training recipe available at `https://github.com/jwr1995/PubSep`

Table 1: *ASR and speech enhancement performance comparison of the AE loss in (3) to baseline SISDR loss models as well as oracle signal $s_c[i]$ and mixture signal $x[i]$ after a set number of ATEs. Δ indicates improvement over the mixture signal. Bold font indicates the best result for each metric. Arrows indicate performance increase.*

ASR Input	Loss	ATEs	CP-WER ↓	Δ ↑	ORC-WER ↓	Δ ↑	SISDR ↑	PESQ ↑	STOI ↑	SRMR ↑
Oracles s_c	-	-	7.7	-	7.7	-	-	-	-	-
Mixture x	-	-	85.7	-	81.8	-	-7.2 dB	1.05	27.1	2.99
Estimates $\hat{s}_c$	$\mathcal{L}_{\text{SISDR}}$	-	44.2	41.5	44.2	37.6	5.5 dB	1.41	69.3	7.14
Estimates $\hat{s}_c$	$\mathcal{L}_{\text{SISDR}}$	30	43.5	42.3	43.4	38.4	**5.6** dB	1.44	71.9	7.21
Estimates $\hat{s}_c$	$\mathcal{L}_{\text{AE}}$ (proposed)	30	**37.1**	**48.7**	**37.1**	**44.8**	5.1 dB	**1.54**	**75.0**	**7.27**

4s audio clips with a learning rate of $\eta_{\text{PT}} = 5 \times 10^{-5}$ over 150 epochs. The learning rate is halved after 90 epochs if there is no improvement in 3 epochs.

3.3. Speech Recognizers

The large SSSR-based Wav2Vec2 model [27] fine-tuned on 960 hours of LibriSpeech data with a CTC loss is chosen[2] as the primary model used to obtain the embeddings $\mathcal{V}(\cdot)$ in the AE loss in (3), as well as for evaluating ASR performance. The Wav2Vec2 model is a large transformer SSSR model that is initially trained to predict speech representations from time-domain signals via self-supervision where the model is trained to predict contextualized representations across masked portions of the input signal from the unmasked signal context [27]. The fine-tuning of Wav2Vec2 involves training an additional linear layer to map these representations to a CTC specific representations, in this case, grapheme-based representations where each logit represents a possible grapheme. Note that the fine-tuning in this model is unrelated to the proposed fine-tuning method described in Section 2. In the large Wav2Vec2 model used in the following experiments, the number of labels is $N = 30$ (28 characters interpreted verbatim plus word boundary token "|" and blank symbol "-"). In the CTC decoder of the Wav2Vec2 ASR model, an open-source 4-gram Librispeech language model is used[3].

In addition, the large Whisper model[4] [28] is used as an *unseen* ASR evaluation model for evaluating ASR performance on a model that was not used in the fine-tuning of the separator. Whisper is a weakly supervised speech foundation model that uses multi-task training. The weakly supervised and multi-task approaches are designed to make the ASR model generalisation well across many acoustic conditions and speaker types [28]. Thus this well-generalising ASR model is thought to make it more challenging for the proposed approach to achieve improvements over the baseline SISDR models.

3.4. Fine-Tuning

For fine-tuning with the AE loss and GPIT, a learning rate of $\eta_{\text{FT}} = 2 \times 10^{-7}$ is used. Fine-tuning is performed over 30 addtional training epochs (ATEs) with the learning rate fixed for all epochs. Exploring different learning rate strategies and optimizing this hyperparameter is beyond the scope of this paper and left to future work.

[2]Torchaudio Wav2Vec2 Large model available at `https://pytorch.org/audio/0.10.0/pipelines.html#torchaudio.pipelines.WAV2VEC2_ASR_LARGE_960H`

[3]Librispeech language model and other resources available at: `https://www.openslr.org/11/`.

[4]In this work, Whisper 'large-v2' is used, which can be downloaded from: `https://github.com/openai/whisper`.

3.5. Evaluation Metrics

Several options are available for multi-speaker WER evaluation [29]. In this paper, two such definitions are chosen: concatenated minimum-power word error rate (CP-WER) [6] and optimal reference combination word error rate (ORC-WER) [3]. The key difference between the two measures is that CP-WER penalizes output speaker channel switches and ORC-WER is unconcerned with whether a given speaker is output on a single channel or multiple channels, so long as the ASR model is still able to estimate the word accurately. CP-WER is thus the more important measure as, ideally, in the proposed system, the goal is to have one speaker for each output channel. However, the addition of ORC-WER provides additional insight into the overall intelligibility of the speech regardless of which channel(s) a given speaker gets output to. Intrusive speech enhancement measures are also used to observe any benefit gained in these using the proposed approach. STOI [30] is used to measure speech intelligibility, perceptual evaluation of speech quality (PESQ) [31] is used to measure speech quality and speech-to-reverberation modulation energy ratio (SRMR) [32] is used to assess reverberant effects.

4. Results

4.1. Results on clean targets

The performance of the AE loss function (3) using the large Wav2Vec2 model is shown in Table 1. The TD-Conformer separation network trained with the proposed AE loss for 30 ATEs is compared to the TD-Conformer separator before fine-tuning and the TD-Conformer trained with an additional 30 epochs, but using the standard signal-level SISDR loss [26, 18]. The model trained with the proposed AE loss significantly outperforms both models in ASR performance (CP-WER and ORC-WER). The AE loss-based model also outperforms the others in terms of the speech enhancement metrics PESQ, STOI and SRMR. This is a powerful finding, as often improved perceptual performance leads to degraded ASR performance and vice-versa. It is consistent however with prior work [10] which similarly compares SSSR output representations in a speech enhancement system loss function.

4.2. Generalization to an Unseen ASR System

The generalization of the improvements found using the proposed AE loss in Table 1 is analysed in this subsection by reevaluating models on the large Whisper ASR model [28] that was not used in the fine-tuning stages. The results in Table 2 show a consistent improvement in both CP-WER and ORC-WER for the AE loss over the SISDR loss trained with additional epochs demonstrating that at least some of the improvements can generalise from one ASR model to another.

Table 2: *ASR performance for the re-evaluation of models in Table 1 using the large Whisper model [28]. Bold indicates the best-performing trained model. Δ indicates improvement over the mixture signal.*

ASR Input	Loss	ATEs	CP-WER ↓	ORC-WER ↓
Oracles s_c	-	-	10.9	10.9
Mixture x	-	-	63.3	60.0
Estimates $\hat{s}_c$	$\mathcal{L}_{\text{SISDR}}$	-	29.5	29.4
Estimates $\hat{s}_c$	$\mathcal{L}_{\text{SISDR}}$	30	29.3	29.2
Estimates $\hat{s}_c$	$\mathcal{L}_{\text{AE}}$	30	**26.8**	**26.7**

4.3. Joint SISDR Loss Weighting

To further assess the impact of the proposed AE loss $\mathcal{L}_{\text{AE}}$ against the SISDR loss $\mathcal{L}_{\text{SISDR}}$ [26, 18], a series of experiments fine-tuning the separator with a joint weighted loss of the two different loss types is carried out. The joint loss is defined as

$$\mathcal{L}_{\text{Joint}} = (1-\alpha)\mathcal{L}_{\text{AE}} + \alpha\mathcal{L}_{\text{SISDR}} \tag{6}$$

where α controls the weighting of the two loss terms. Five models are fine-tuned with values of $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ for 30 ATEs.

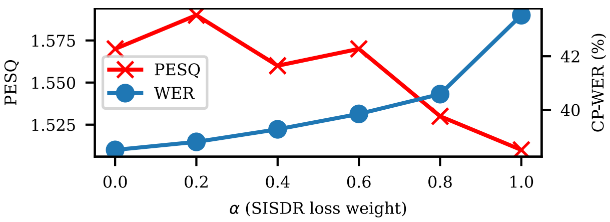

Figure 2: *ASR performance (CP-WER) and objective perceptual quality of test audio for models trained with differing weight α between loss terms in (6).*

Figure 2 shows the performance of the models fine-tuned using (6) in terms of CP-WER and PESQ on the WHAMR test set. The ASR performance of the models trained with lower values of α is significantly better than those trained with higher values, as the influence of the proposed $\mathcal{L}_{\text{AE}}$ function on the models' training is higher for lower values of α. The biggest jump in improvement is between $\alpha = 1.0$ (i.e no use of $\mathcal{L}_{\text{AE}}$) and $\alpha = 0.8$, suggesting that even a small weighting of the proposed loss can significantly improve ASR performance. Perceptual quality in terms of PESQ does not change as uniformly or significantly for lower α values, but a trend is still apparent; models with a greater weighting of the proposed loss $\mathcal{L}_{\text{AE}}$ produce higher quality audio.

4.4. Training Signal Length Analysis

A benefit of the transcription-free method is the ability to train on arbitrary signal lengths. In transcription-dependant losses, this is non-trivial due to the requirement of truncating the transcription in alignment with the audio. The impact of applying TSL limits [16, 33] is analysed in Table 3. The results show that using longer signal lengths gives slightly better performance, most likely due to the ASR Encoder having more context. This is opposite to the SISDR-based evaluations in [16] where improved SISDR performance could be obtained on WHAMR using shorter signal lengths.

Table 3: *Comparison of the impact of the TSL limit on ASR performance of models fine-tuned using the proposed AE loss over 30 ATEs. Best results are shown in bold.*

TSL limit (s)	CP-WER ↓	ORC-WER ↓	SISDR ↑
2.0	38.3	38.4	**5.0**
4.0	37.1	37.1	5.1
8.0	**36.3**	**36.2**	5.2

4.5. ASR Encoder Visualization

A visual comparison of Wav2Vec2 ASR Encoder output representations $\mathcal{V}(\cdot)$ used in (3) of audio outputs from models trained using SISDR loss (middle) and AE loss (bottom) can be seen in Figure 3. The ASR encoder output representation $\mathcal{V}(\cdot)$ of the respective reference audio is visualized in the top panel. Notably, the AE loss reduces the uncertainty in the logits when compared to the SISDR loss. In both representations of the estimated signals, there is some bleed-through of interfering noise and speakers at $L \gtrapprox 180$, but this effect is reduced in the AE loss suggesting that the additional context provided by the encoder network may improve separation performance and reduce more general distortions.

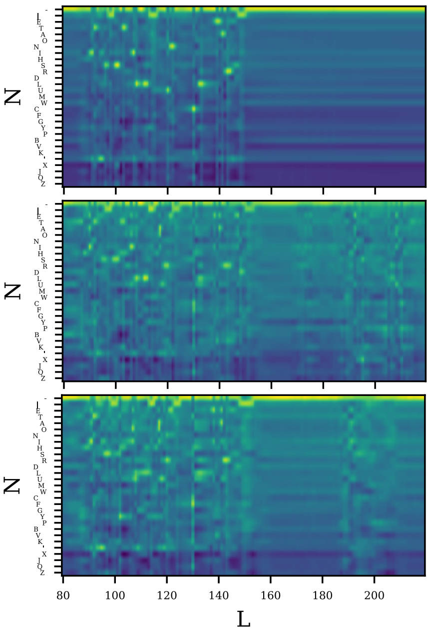

Figure 3: *Wav2Vec2 ASR Encoder output representations for reference audio $\mathcal{V}(\mathbf{s}_c)$ (top), and $\mathcal{V}(\hat{\mathbf{s}}_c)$ for models with baseline $\mathcal{L}_{\text{SISDR}}$ (middle) and proposed $\mathcal{L}_{\text{AE}}$ fine-tuning (bottom) losses*

5. Conclusions

This paper presented a novel method for fine-tuning speech separation models for multi-speaker ASR without the need for reference transcriptions. It was shown that the proposed loss leveraging pre-trained ASR encoder representations for fine-tuning separators results in improved ASR performance over a standard signal-level-based loss. It was also demonstrated that the proposed method produces improved performance in speech enhancement metrics such as PESQ and STOI which is often not the case for many other approaches. Finally, it was shown that the performance gained by the proposed method generalises to other ASR models that weren't used in the fine-tuning stage.

6. References

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, “Far-field automatic speech recognition,” *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[2] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, “Monaural source separation: From anechoic to reverberant environments,” in *IWAENC 2022*, Sep. 2022, pp. 1–5.

[3] I. Sklyar, A. Piunova, X. Zheng, and Y. Liu, “Multi-turn rnn-t for streaming recognition of multi-party speech,” in *ICASSP 2022*, May 2022, pp. 8402–8406.

[4] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, “Serialized Output Training for End-to-End Overlapped Speech Recognition,” in *Proc. Interspeech 2020*, 2020, pp. 2797–2801.

[5] T. von Neumann, C. Boeddeker, T. Cord-Landwehr, M. Delcroix, and R. Haeb-Umbach, “Meeting recognition with continuous speech separation and transcription-supported diarization,” 2023. [Online]. Available: https://arxiv.org/abs/2309.16482

[6] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, “CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings,” in *CHiME 2020*, May 2020, pp. 1–7.

[7] W. Ravenscroft, S. Goetze, and T. Hain, “Att-TasNet: Attending to Encodings in Time-Domain Audio Speech Separation of Noisy, Reverberant Speech Mixtures,” *Frontiers in Signal Processing*, vol. 2, 2022.

[8] J. Clarke, Y. Gotoh, and S. Goetze, “Improving audiovisual active speaker detection in egocentric recordings with the data-efficient image transformer,” in *ASRU 2023*, Dec. 2023, pp. 1–8.

[9] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, “MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition,” in *ASRU 2019*, Dec. 2019, pp. 237–244.

[10] G. Close, W. Ravenscroft, T. Hain, and S. Goetze, “Perceive and predict: Self-supervised speech representation based loss functions for speech enhancement,” in *ICASSP 2023*, Jun. 2023, pp. 1–5.

[11] ——, “Multi-CMGAN+/+: Leveraging Multi-Objective Speech Quality Metric Prediction for Speech Enhancement,” in *ICASSP 2024*, Apr. 2024, pp. 351–355.

[12] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, “Spectral feature mapping with mimic loss for robust speech recognition,” in *ICASSP 2018*, Apr. 2018, pp. 5609–5613.

[13] Z. Tu, N. Ma, and J. Barker, “Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners,” in *Proc. Interspeech 2022*, Sep. 2022, pp. 3488–3492.

[14] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, “Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate asr features and human memory models,” in *ICASSP 24*, May 2024, pp. 306–310.

[15] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *ICASSP 2017*, Mar. 2017, pp. 241–245.

[16] W. Ravenscroft, S. Goetze, and T. Hain, “On Data Sampling Strategies for Training Neural Network Speech Separation Models,” in *EUSIPCO 2023*, Sep. 2023, pp. 331–335.

[17] J. Kalda, C. Pagés, R. Marxer, T. Alumäe, and H. Bredin, “PixIT: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings,” 2024. [Online]. Available: https://arxiv.org/abs/2403.02288

[18] Y. Luo and N. Mesgarani, “Tasnet: Time-domain audio separation network for real-time, single-channel speech separation,” in *ICASSP 2018*, Apr. 2018, pp. 696–700.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML 2006*, Jun. 2006, p. 369–376.

[20] M. Kolbaek, D. Yu, Z.-H. Tan, J. Jensen, 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, p. 1901–1913, Oct. 2017.

[21] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, “Graph-PIT: Generalized Permutation Invariant Training for Continuous Separation of Arbitrary Numbers of Speakers,” in *Proc. Interspeech 2021*, 2021, pp. 3490–3494.

[22] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, “WHAMR!: Noisy and Reverberant Single-Channel Speech Separation,” in *ICASSP 2020*, May 2020, pp. 696–700.

[23] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-Channel Multi-Speaker Separation Using Deep Clustering,” in *Proc. Interspeech 2016*, Sep. 2016, pp. 545–549.

[24] D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR corpus,” in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. [Online]. Available: https://aclanthology.org/H92-1073

[25] N. Zeghidour and D. Grangier, “Wavesplit: End-to-end speech separation by speaker clustering,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[26] W. Ravenscroft, S. Goetze, and T. Hain, “On time domain conformer models for monaural speech separation in noisy reverberant acoustic environments,” in *ASRU 2023*, 2023, pp. 1–7.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *ICML 2023*, Jul. 2023, pp. 28 492–28 518.

[29] T. von Neumann, C. Boeddeker, K. Kinoshita, M. Delcroix, and R. Haeb-Umbach, “On word error rate definitions and their efficient computation for multi-speaker speech recognition systems,” in *ICASSP 2023*, Jun. 2023, pp. 1–5.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *ICASSP 2010*, Mar. 2010, pp. 4214–4217.

[31] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *ICASSP 2001*, May 2001, pp. 749–752.

[32] T. H. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[33] W. Ravenscroft, S. Goetze, and T. Hain, “Receptive field analysis of temporal convolutional networks for monaural speech dereverberation,” in *EUSIPCO 2022*, Aug. 2022, pp. 80–84.