2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 978-1-6654-3739-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASRU51503.2021.9688231

SPEAKER CONDITIONING OF ACOUSTIC MODELS USING AFFINE TRANSFORMATION FOR MULTI-SPEAKER SPEECH RECOGNITION

Midia Yousefi, John H.L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, USA

ABSTRACT

This study addresses the problem of single-channel Automatic Speech Recognition of a target speaker within an overlap speech scenario. In the proposed method, the hidden representations in the acoustic model are modulated by speaker auxiliary information to recognize only the desired speaker. Affine transformation layers are inserted into the acoustic model network to integrate speaker information with the acoustic features. The speaker conditioning process allows the acoustic model to perform computation in the context of target-speaker auxiliary information. The proposed speaker conditioning method is a general approach and can be applied to any acoustic model architecture. Here, we employ speaker conditioning on a ResNet acoustic model. Experiments on the WSJ corpus show that the proposed speaker conditioning method is an effective solution to fuse speaker auxiliary information with acoustic features for multi-speaker speech recognition, achieving +9% and +20% relative WER reduction for clean and overlap speech scenarios, respectively, compared to the original ResNet acoustic model baseline.

Index Terms— Affine transformation, overlap speech recognition, feature-wise linear modulation, multi-speaker recognition, acoustic modeling

1. INTRODUCTION

Multi-talker speech recognition is focused on recognizing individual speech sources from overlap speech, and is one main challenge for current ASR systems [1, 2, 3, 4, 5, 6, 7, 8]. Current solutions for multi-speaker speech recognition can be categorized into two main approaches: *(i)* performing front-end speech processing based on separation on the overlap speech, then applying ASR to the separated speech signals [9, 10, 11, 12, 13, 14, 15]; or *(ii)* skipping the explicit separation step and developing a multi-speaker speech recognition system directly using either hybrid [16, 17, 18] or end-to-end [19, 20] ASR frameworks. Recently, an end-to-end multi-speaker speech recognition system was proposed based on Transformers [1]. This approach achieved considerable improvement at the expense of more computational cost for a reasonable temporal resolution. In another study [21], overlap speech was considered as a mismatch condition of the clean speech recognition scenario, and teacher-student training was employed for transfer learning from clean to overlap speech. The main drawback of this approach is requiring training sets with parallel clean and overlapped speech, which is difficult to collect in real-world applications [22]. Recently, several studies [16, 23, 24] have used speaker-specific embeddings to learn a frame-level mask for the target speaker which suppresses interfering speech. Although these approaches use the additional speaker-specific information to guide the ASR system, their main limitation is that they assume only one speaker is active in each Time-Frequency bin.

To address the challenges of single-channel multi-speaker speech recognition, in this study, we focus on speaker conditioning of the Acoustic Model (AM) by performing an affine transformation. In contrast to former approaches which employ speaker embedding to estimate speaker-specific masks, we propose to use speaker embedding to compute parameters of the affine transformation, allowing the acoustic model to conduct its computation in the context of the desired speaker auxiliary information. The proposed speaker conditioning method is a general approach and can be applied to any AM architecture. In this paper, we employ speaker conditioning on a ResNet acoustic model in hybrid DNN-HMM setup. Experiments are performed on WSJ corpus, achieving +9% and +20% relative WER reduction for clean and overlap speech scenarios, compared to the original ResNet acoustic model. The contributions of this paper are threefold:

- Proposing speaker conditioning of the ResNet acoustic model using an affine transformation.
- Comparing the proposed method with alternate feature-wise acoustic model transformations such as conditional biasing and middle feature-map fusion.
- Evaluating the performance of the proposed speaker conditioned ASR trained on an alternate input feature called Wav2Vec representation.

The remainder of this paper is organized as follows. In Sec.2, the problem is outlined and proposed method described. Sec.3, presents experiments and results. Finally the conclusions are discussed in Sec.4.

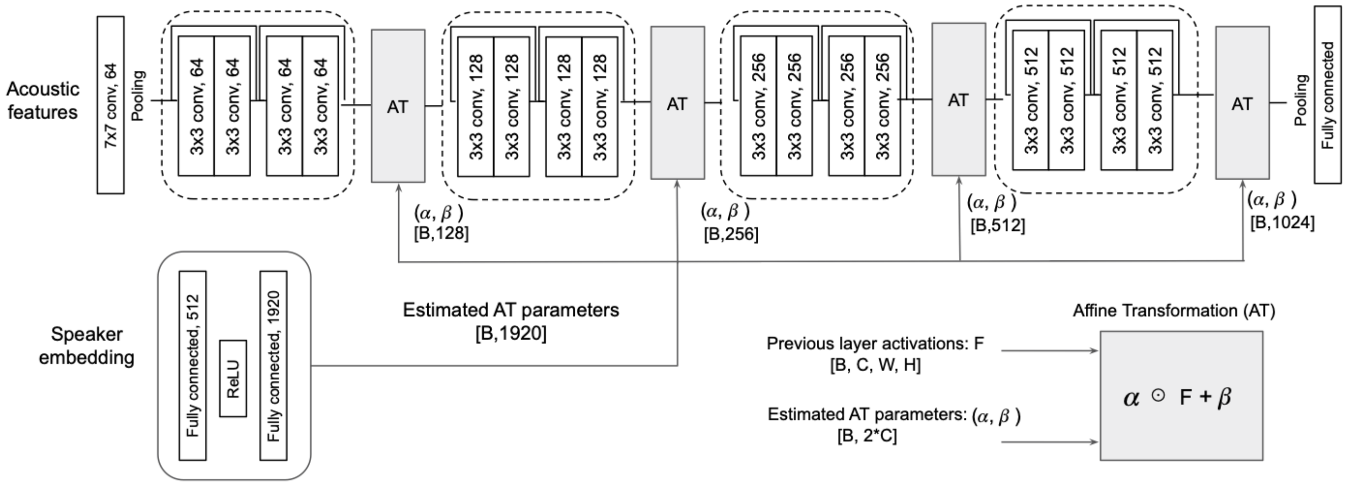

Fig. 1. The proposed speaker conditioned ResNet18 acoustic model using Affine Transformation (AT) blocks.

2. SINGLE-CHANNEL MULTI-SPEAKER ASR

Multi-speaker speech recognition can gain substantial improvement by deploying other sources of information such as speaker identity in addition to acoustic features [22, 23]. However, designing an efficient method to fuse a combination of multiple sources (i.e., acoustic features and speaker embeddings) to obtain higher quality and improved information is still a challenging task. Additionally, capturing complex interactions between multiple sources should maintain a balanced compromise between model/network computational cost and performance. A popular method to address this problem is to use a feature-wise transformation [25] which can model the complex relation between speaker-specific characteristics and acoustic features in a multi-speaker speech scenario to identify and recognize the desired speaker in the mixed speech recording. This transformation can be performed in several manners such as *conditional biasing*, *conditional scaling*, and *conditional affine transformation*. In this section, we focus on a conditional affine transformation which is a more general approach. The aforementioned conditional biasing and scaling are two specific examples of conditional affine transformation.

2.1. Conditional affine transformation

Affine Transformation (AT) influences the output of the acoustic model network by applying a linear modulation to the network's intermediate features. This modulation is parameterized by shifts and coefficients obtained based on speaker-specific embedding. Let x be a context-expanded window of acoustic features for overlap speech, and y_s be a phoneme label or a senone alignment (i.e., from GMM-HMM) for the target-speaker speech signal. DNN acoustic models are used estimate the posterior probability as:

$$p(y_s|x, s) = DNN(x, z_s), \tag{1}$$

where DNN is typically trained to maximize the log probability of the phoneme alignment or minimize the cross-entropy error, and s is the target speaker with an x-vector [26] embedding z_s. In this study, the original ResNet18 model is considered as our baseline. Next, affine transformation layers are inserted into ResNet18 network to build the speaker conditioned acoustic model. The scale and bias factors of AT are estimated by a two-layer fully connected network h based on x-vector z_s as:

$$(\alpha_{i,c},\ \beta_{i,c}) = h(z_s) \tag{2}$$

where i and c refer to the i-th data sample in the minibatch, and the c-th channel feature map. Once $\alpha_{i,c}$ and $\beta_{i,c}$ are estimated, they are used to modulate the ResNet's intermediate activations $F_{i,c}$ as:

$$AT(F_{i,c}^{l}|\alpha_{i,c}, \beta_{i,c}, F_{i,c}^{l-1}) = \alpha_{i,c} \odot F_{i,c}^{l-1} + \beta_{i,c} \tag{3}$$

where AT and l represent the Affine Transformation, and network's layer. The proposed speaker conditioned ResNet18 is shown in Fig.1. The speaker embedding x-vector is submitted to the network h to estimate a $[B, 1920]$ matrix which is $(\alpha_{i,c}, \beta_{i,c})$ pairs of AT layers. Each AT layer receives two inputs: the previous layer output, and the $(\alpha_{i,c}, \beta_{i,c})$ pair. The dimension of $\alpha_{i,c}$ and $\beta_{i,c}$ is $[B, C]$ each. In the AT layer, each channel of the extracted feature map is scaled by $\alpha_{i,c}$ and shifted by $\beta_{i,c}$ to modulate the feature-map distribution of activations based on the target-speaker embedding.

Table 1. Comparing our ResNet18 acoustic model baseline with other approaches on WSJ (WER in %).

System	Dev-93	Eval-92
Lee et al. 2021 [27]	12.0	9.9
Higuchi et al. 2020 [28]	15.4	12.1
Chi et al. 2020 [29]	13.7	11.4
Rouhe et al. 2020 [30]	13.2	9.3
Sabour et al. 2018 [31]	-	9.3
Borgholt et al. 2020 [32]	-	9.3
Park et al. 2019 [33]	-	7.8
Our baseline	12.1	7.9

Table 2. WER of the proposed speaker conditioned ResNet18 acoustic model with Affine Transformation (AT) in different settings. Each experiment is repeated three times and the average WER is reported.

	Simulated overlap speech based on Dev-93						Clean speech		
Acoustic model	0dB	5dB	10dB	15dB	20dB	25dB	Dev-93	Eval-92	Eval-93
ResNet18 (baseline)	65.06	58.29	47.03	34.72	24.36	17.62	12.14	7.92	10.81
ResNet18 + AT (proposed)	63.83	55.83	43.30	29.90	20.34	15.15	11.50	7.66	9.64
ResNet18 + AT (bias=0)	63.69	55.23	43.57	30.15	19.95	15.19	11.75	7.66	9.57
ResNet18 + AT (scale=1)	65.10	57.66	46.39	32.79	22.29	16.65	12.25	7.91	10.57
ResNet18 + AT (sigmoid(scale))	64.63	56.66	45.33	31.78	21.680	16.250	11.993	7.803	10.263
ResNet18 + AT (tanh(scale))	64.53	56.82	45.27	31.64	21.19	16.02	11.99	7.72	9.97
ResNet18 + AT (Block1)	63.68	55.33	**43.31**	29.44	**19.51**	**14.67**	11.56	**7.49**	**9.85**
ResNet18 + AT (Block 1-2)	**63.50**	**55.19**	43.60	**29.33**	19.79	14.85	**11.33**	7.50	9.88
ResNet18 + AT (Block 1-3)	63.51	55.29	43.35	29.65	19.87	14.88	11.49	7.55	9.93
ResNet18 + AT (Block 4)	64.66	57.66	47.28	33.86	23.27	16.54	11.64	8.19	10.47

3. EXPERIMENTS AND RESULTS

In this section, we investigate the performance of the proposed speaker conditioning method presented in Fig.1 on WSJ corpus. In order to conduct the experiments, clean *tr-si284* is used in the training phase for all acoustic models. We partitioned *tr-si284* into a training set (90%) and a held-out cross-validation set (10%). ASR performance for different acoustic models are reported in terms of Word-Error-Rate (WER) on clean *dev-93*, *eval-93*, and *eval-92*. Additionally, overlap speech is generated based on *dev-93* by selecting random utterances from random speakers and adding them with Signal-to-Interference Ratio (SIR) ranging from 0 to 25dB with increments of 5dB. The baseline acoustic model is ResNet18 with 3400 output senones. The network parameters are updated by the gradients of the cross entropy loss using Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and initial learning rate 0.01. The training process is completed by performing early stopping [34]. The maximum number of epochs is set to 100, batch size 1024 context-expanded frames; learning rate is decreased by 50% if the cv loss improvement is less than 0.01 for 3 successive epochs. The early stopping is performed if no improvement is observed on the cv loss once the learning rate has decayed 6 times. 13-dim MFCC computed over a $25ms$ window with $10ms$ shift with a 20 frame context (10 frames on each side) is used for training the acoustic model. Consistent with the standard Kaldi recipe for WSJ, we use the trigram language models provided by LDC for WSJ data. In order to minimize the effect of parameter initialization on the acoustic model and final WER, we repeat each experiment three times with different initial parameters.

Performance of the ResNet18 baseline is compared with recent studies in Table 1. The main purpose of this comparison is to ensure that our baseline achieves a competitive performance compared to recent studies, and it is seen we have a strong starting point for further developing our proposed speaker conditioning technique. There are other approaches that leverage transfer learning, semi supervised learning, or more advanced language models to achieve further improvement. However, since we focus on speaker conditioning of acoustic model, we train our ResNet18 acoustic model only on *tr-si284*, and use Kaldi for training the language model. The first row of Table 2 presents performance of the baseline on overlap speech, which is severely degraded. Therefore, we build on the ResNet18 acoustic model baseline and apply our Affine Transformation (AT) layers as depicted in Fig.1. The results for speaker conditioned ResNet using AT are presented in the second row of Table 2 which shows $+2\%$ relative improvement for severe overlap speech recordings (i.e., 0dB) and an average of $+5\%$ relative improvement on clean test sets. Since AT effectively performs speaker-adaptation, the trained acoustic model is tuned to the target speaker, therefore, it achieves better performance even on the clean test sets. The maximum relative improvement is achieved for input SIR 20% in which the level of overlap speech is neither too severe nor too easy for the acoustic model; therefore, the target-speaker auxiliary information can be very helpful in improving performance.

Moreover, the effect of α and β is studied separately by setting $\alpha = 1$ and $\beta = 0$. The result in Table 2 manifest that the effectiveness of the conditional Affine Transformation can be mainly attributed to the scale coefficient rather than the shift parameter. Therefore, we further investigate the effect of α by restricting its value to $(0, 1)$ using Sigmoid, and $(-1, 1)$ using $\tanh$ function. Nevertheless, the results reported in Table 2 reveal that unrestricted α achieves better performance which may be due to the flexibility it provides for the network to learn the range that best suits the data. So far, the AT layers have been applied to all ResNet18 blocks

Table 3. WER of the proposed speaker conditioned ResNet18 based on Affine Transformation (AT) compared to other fusion techniques. Each experiment is repeated three times and the average WER is reported.

	Simulated overlap speech based on Dev-93						Clean speech		
Signal-to-Interference Ratio	0dB	5dB	10dB	15dB	20dB	25dB	Dev-93	Eval-92	Eval-93
ResNet18 + Conditional biasing	64.66	57.20	45.74	32.75	23.54	17.87	13.10	8.44	11.22
ResNet18 + Middle fusion	63.94	57.51	47.81	34.02	23.19	16.68	11.81	8.24	10.64
ResNet18 + AT (proposed)	**63.68**	**55.33**	**43.31**	**29.44**	**19.51**	**14.67**	**11.56**	**7.49**	**9.85**

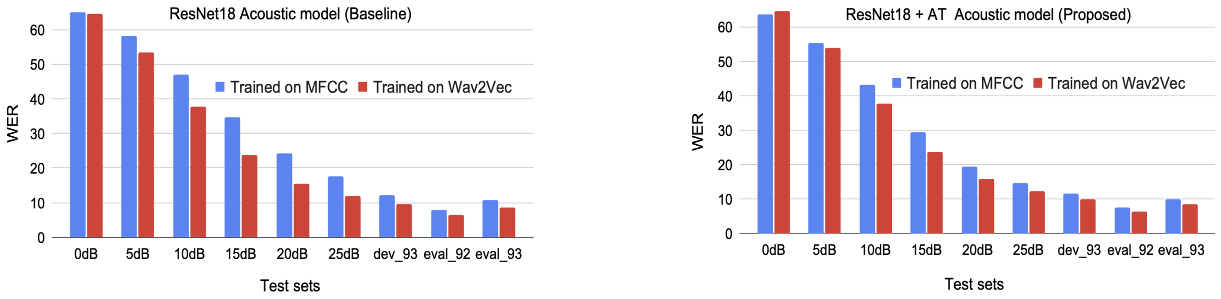

Fig. 2. WER of ResNet18 baseline and proposed ResNet18 + AT trained on MFCC and Wav2Vec input features.

(each dashed rectangular in Fig. 1 is considered as a block). To find the best network depth in which AT layers are most effective, several experiments are conducted with AT only applied to specific individual blocks. Based on these experiments, the AT layers are most effective when applied only to the first block (block1), and least effective when only applied to the last block (block4). However, applying AT layers to the first two blocks (block1-2) and the first three blocks (block 1-3) did not improve ASR performance, while it differently adds computational cost. To summarize our findings based on the experiments, unrestricted-scale Affine Transformation applied to the initial blocks of the ResNet18 acoustic model achieves the best overall results while simultaneously maintaining the lowest computational cost.

Next, the proposed method is compared with other speaker conditioning techniques in Table 3. Conditional biasing refers to adding speaker information (x-vector) as a bias to the acoustic features in the first hidden layer. Middle fusion refers to adding the x-vector to the intermediate extracted feature map after the second block. Therefore, the intermediate feature map is conditioned before entering block 3 for extracting further higher-level features adapted to the target speaker. As shown in Table 3, the proposed speaker conditioning based on Affine Transformation outperforms all other approaches in both clean and overlap speech scenarios.

So far, the focus of this study has been on designing the acoustic model. However, performance of the acoustic model can further improve by deploying more robust input features other than MFCC. In the final section, we evaluate the proposed method trained on noise-invariant Wav2Vec features [35]. Wav2Vec representation has been trained on large amounts of unlabeled audio data in an unsupervised manner. Fig. 2 (left) shows the WER of the baseline ResNet18 trained on MFCC and Wav2Vec features, which manifests the effectiveness of Wav2Vec in reducing the WER across all test sets in the absence of speaker auxiliary information. The highest improvement is achieved for overlap speech with SIR 15dB, which is $+11\%$ absolute improvement in WER. Fig.2 (right) depicts the WER of the proposed speaker conditioned ResNet18 trained on MFCC and Wav2Vec. Similar to the baseline, the speaker conditioned acoustic model benefits from the Wav2Vec features by achieving $+6\%$ absolute improvement in WER for SIR 15dB. However, due to the availability of speaker information, the acoustic model is less sensitive to the robustness of the input acoustic features, and thus, the amount of improvement from Wav2Vec is less in the proposed speaker conditioned ResNet18 compared to the baseline. In conclusion, the WER across all test sets is improved by using the proposed speaker conditioned acoustic model trained on wav2Vec. For example, on the overlap speech test set with SIR 15dB, the proposed ResNet18 with Affine Transformation trained on Wav2Vec gains +33% relative (+11% absolute) improvement in WER compared to the original ResNet trained on MFCC.

4. CONCLUSION

In this study, we proposed a speaker conditioning method for acoustic modeling in multi-speaker speech recognition. In the proposed method, Affine Transformation layers are inserted into the acoustic model architecture to fuse speaker-specific information with the acoustic model. The proposed speaker conditioned acoustic model was compared with other fusion techniques such as early fusion of speaker embedding and middle feature-map fusion. Additionally, the performance of the proposed method was evaluated on alternate input features called Wav2Vec. The results on WSJ corpus clearly demonstrate that the proposed speaker conditioned acoustic model based on affine transformation achieves consistent WER improvement for clean and overlap speech scenarios.

5. REFERENCES

[1] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6134–6138.

[2] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al., "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[3] Midia Yousefi and John H.L. Hansen, "Real-Time Speaker Counting in a Cocktail Party Scenario Using Attention-Guided Convolutional Neural Network," in *Proc. Interspeech 2021*, 2021, pp. 1484–1488.

[4] Midia Yousefi, Navid Shokouhi, and John HL Hansen, "Assessing speaker engagement in 2-person debates: Overlap detection in united states presidential debates.," in *Interspeech*, 2018, pp. 2117–2121.

[5] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth'chime'speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[6] Yan-min Qian, Chao Weng, Xuan-kai Chang, Shuai Wang, and Dong Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[7] Seyedmahdad Mirsamadi and John HL Hansen, "Multi-domain adversarial training of neural network acoustic models for distant speech recognition," *Speech Communication*, vol. 106, pp. 21–30, 2019.

[8] Midia Yousefi and John HL Hansen, "Block-based high performance cnn architectures for frame-level overlapping speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 28–40, 2020.

[9] Christoph Boeddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[10] Midia Yousefi and Mohammad Hassan Savoji, "Supervised speech enhancement using online group-sparse convolutive nmf," in *2016 8th International Symposium on Telecommunications (IST)*. IEEE, 2016, pp. 494–499.

[11] Midia Yousefi and John HL Hansen, "Frame-based overlapping speech detection using convolutional neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6744–6748.

[12] Li Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data*, pp. 67–99. Springer, 2011.

[13] Arun Narayanan and DeLiang Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[14] Seyedmahdad Mirsamadi and John HL Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for asr applications," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[15] Midia Yousefi, Soheil Khorram, and John HL Hansen, "Probabilistic permutation invariant training for speech separation," *arXiv preprint arXiv:1908.01768*, 2019.

[16] Naoyuki Kanda, Yusuke Fujita, Shota Horiguchi, Rintaro Ikeshita, Kenji Nagamatsu, and Shinji Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6630–6634.

[17] Chao Weng, Dong Yu, Michael L Seltzer, and Jasha Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[18] Naoyuki Kanda, Shota Horiguchi, Yusuke Fujita, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 31–38.

[19] Hiroshi Seki, Takaaki Hori, Shinji Watanabe, Jonathan Le Roux, and John R Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2620–2630.

[20] Liang Lu, Naoyuki Kanda, Jinyu Li, and Yifan Gong, "Streaming end-to-end multi-talker speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 803–807, 2021.

[21] Zhehuai Chen, Jasha Droppo, Jinyu Li, and Wayne Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 184–196, 2017.

[22] Pavel Denisov and Ngoc Thang Vu, "End-to-end multi-speaker speech recognition using speaker embeddings and transfer learning," *arXiv preprint arXiv:1908.04737*, 2019.

[23] Yiming Wang, Xing Fan, I-Fan Chen, Yuzong Liu, Tongfei Chen, and Björn Hoffmeister, "End-to-end anchored speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7090–7094.

[24] Aswin Shanmugam Subramanian, Chao Weng, Meng Yu, Shi-Xiong Zhang, Yong Xu, Shinji Watanabe, and Dong Yu, "Far-field location guided target speech extraction using end-to-end speech recognition objectives," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7299–7303.

[25] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[26] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[27] Jaesong Lee and Shinji Watanabe, "Intermediate loss regularization for ctc-based speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6224–6228.

[28] Yosuke Higuchi, Shinji Watanabe, Nanxin Chen, Tetsuji Ogawa, and Tetsunori Kobayashi, "Mask ctc: Non-autoregressive end-to-end asr with ctc and mask predict," *arXiv preprint arXiv:2005.08700*, 2020.

[29] Ethan A Chi, Julian Salazar, and Katrin Kirchhoff, "Align-refine: Non-autoregressive speech recognition via iterative realignment," *arXiv preprint arXiv:2020.14233*, 2020.

[30] Aku Rouhe, Tuomas Kaseva, and Mikko Kurimo, "Speaker-aware training of attention-based end-to-end speech recognition using neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7064–7068.

[31] Sara Sabour, William Chan, and Mohammad Norouzi, "Optimal completion distillation for sequence learning," in *International Conference on Learning Representations*, 2018.

[32] Lasse Borgholt, Jakob D Havtorn, Željko Agić, Anders Søgaard, Lars Maaløe, and Christian Igel, "Do end-to-end speech recognition models care about context?," *Proc. Interspeech 2020*, pp. 4352–4356, 2020.

[33] Lukas Lee, Jinhwan Park, and Wonyong Sung, "Simple gated convnet for small footprint acoustic modeling," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 122–128.

[34] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization," *preprint arXiv:1611.03530*, 2016.

[35] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.