# SEQUENCE MODELING IN UNSUPERVISED SINGLE-CHANNEL OVERLAPPED SPEECH RECOGNITION

*Zhehuai Chen*[1,2] *and Jasha Droppo*[2]

[1]Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]Microsoft AI and Research, Redmond, WA, USA

## ABSTRACT

Unsupervised single-channel overlapped speech recognition is one of the hardest problems in automatic speech recognition (ASR). The problems can be modularized into three sub-problems: frame-wise interpreting, sequence level speaker tracing and speech recognition. Nevertheless, previous acoustic models formulate the correlation between sequential labels implicitly, which limit the modeling effect. In this work, we include explicit models for the sequential label correlation during training. This is relevant to models given by both the feature sequence and the output of the last frame. Moreover, we propose to integrate the linguistic information into the assignment decision of the permutation invariant training (PIT). Namely, a senone level neural network language model (NNLM) trained in the clean speech alignment is integrated, while the objective function is still cross-entropy. The proposed methods can be combined with an improved version of PIT and sequence discriminative training, which brings about further over 10% relative improvement of WER in the artificial overlapped Switchboard and hub5e-swb dataset.

***Index Terms***— unsupervised single channel overlapped speech recognition, permutation invariant training, temporal correlation modeling, language model

## 1. INTRODUCTION

The cocktail party problem [1, 2], referring to multi-talker overlapped speech recognition, is critical to enable automatic speech recognition (ASR) scenarios such as automatic meeting transcription, automatic captioning for audio/video recordings, and multi-party human-machine interactions, where overlapped speech is commonly observed and all streams need to be transcribed. However, the problem is still one of the hardest problems in ASR, despite encouraging progresses [3, 4, 5, 6].

In this paper, we aim to solve the speech recognition problem when multiple unseen talkers speak at the same time and only a single channel of overlapped speech is available. This is useful when only a single microphone is present, or when microphone array based algorithms fail to perfectly separate the speech. [7] divides the problem into three sub-problems: frame-wise interpreting, speaker tracing and speech recognition. These modules are independently pretrained and jointly fine-tuned, which improves the accuracy of the model. The paper follows and extends this technique.

As speaker tracing and speech recognition are both sequence level problems, sequence modeling is the key to success. Previous methods cope with the problem implicitly or explicitly. In computational auditory scene analysis (CASA) [3], there are two main stages: the segmentation and grouping. Segmentation stage is to decompose mixed speech into time-frequency segments assumed to be derived from the corresponding speakers based on perceptual grouping cues [8]. Grouping is to sequentially concatenate the segments to generate independent streams for each speaker. In the deep learning era, [9, 10] propose deep clustering (DPCL), in which a deep network is trained to produce spectrogram embeddings that are discriminative for partition labels given in training data. The model is optimized so that in the neural network embedding space the time-frequency bins belonging to the same speaker are closer and those of different speakers are farther away. The DPCL grouping state applies a clustering algorithm to these embeddings. A language model is employed to aid the grouping stage of a multi-stream joint decoder in [6]. Although this sequence-level information improves accuracy, it suffers from exponential growth in the search space and is therefore inappropriate for large vocabulary continuous speech recognition (LVCSR). Permutation invariant training (PIT) [11] jointly models the voice discrimination, speaker tracing and speech recognition with an unified sequence level criterion. After it determines the output-target assignment with the minimum error at utterance level based on the forward-pass result, it minimizes the error given the assignment. Bidirectional long short term memory (BLSTM) is employed to enhance sequence modeling effect. [7] improves PIT using sequence discriminative criterion for both the assignment and the error criterion. All of these previous works model the sequence level correlation implicitly, which may limit the modeling effect.

In this work, we include explicit models for the sequential label correlation during training, which are given by both the feature sequence and the output of the last frame. Besides, we propose to integrate the linguistic information into the assignment decision of the permutation invariant training (PIT). Namely, a senone level neural network language model (NNLM) trained in the clean speech alignment is integrated, while the objective function is still cross-entropy. The whole paper is arranged as follows. In Section 2, transfer learning based progressive joint training framework [7] is briefly reviewed. In Section 3, temporal correlation modeling and language model integration are proposed. Section 5 describes experiments and analysis, followed by the conclusion in section 6.

## 2. TRANSFER LEARNING BASED PROGRESSIVE JOINT MODELING

In the original formulation, a PIT-ASR model consists of a single monolithic structure that predicts independent targets for each speak-

The work of Zhehuai (Tom) Chen was performed during his internship at Microsoft Research. The authors would like to thank Jinyu Li and Wayne Xiong for many helpful conversations. Zhehuai would like to further thank Jerry and Juncheng Gu for their supports during the internship. Zhehuai's travel to the ICASSP 2018 conference was supported by the Shanghai International Science and Technology Cooperation Fund (No. 16550720300).

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

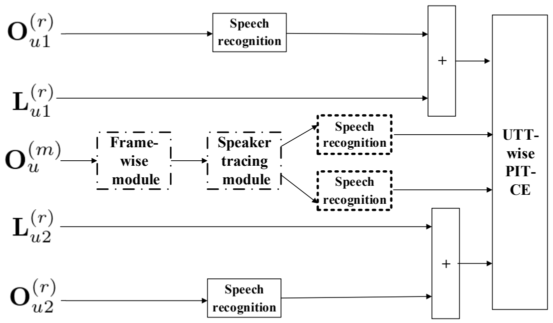


**Fig. 1**. *Transfer Learning Based Progressive Joint Training. The dash-dot blocks indicate the learnable model parameters. The dot-dot blocks indicate the learnable and shared model parameters.*

er [11]. [7] improves this by replacing the main network structure with a modular structure: frame-wise interpreting, speaker tracing, and speech recognition modules. These modules are independently pretrained and jointly fine-tuned, which improves the accuracy of the resulting model.

Firstly, the frame-wise module is designed to extract the local time-frequency information necessary to separate the overlapped speech into individual acoustic representations. Second, the speaker tracing module accepts frame-wise acoustic representations from the frame-wise module and traces the speaker information. Third, the speech recognition modules accept the sequences of recovered acoustic features from each speaker, and produce a sequence of label scores suitable for use in an automatic speech recognition system.

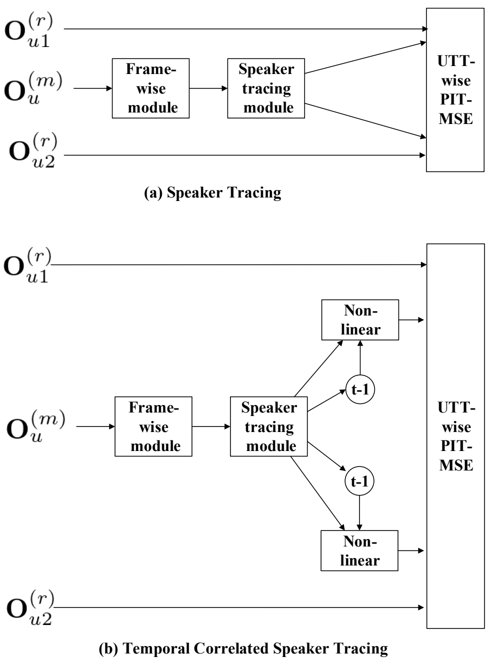
Although it is possible to train the modularized network from random initialization, it is better to use a progressive training strategy. The strategy is motivated by the curriculum learning theory [12], which integrates both modularization and joint training. We train a simple model first, and then use it as a pre-trained block for a more complicated model and task. Thus the model becomes progressively more complex while solving more difficult problems from frame-wise mean squared error to whole utterance cross entropy (CE).

Transfer learning (teacher-student) based domain adaptation can be used to further improve the joint training. Here, the student is the multi-channel speech recognition system. It operates in the target domain of mixed speech acoustic data, and must produce separate outputs for each speaker in the mixture. The teacher also must produce separate outputs for each speaker, but has access to the source domain: un-mixed clean speech. The teacher model is a set of clean speech acoustic models operating independently on the separate channels of clean speech. The transfer learning method then minimizes the Kullback-Leibler divergence (KLD) between the output distribution of the mixed speech model and the set of clean speech models. It is notable that when this method is applied to the modular structure proposed in this work, as in Figure 1, the speech recognition modules can be initialized with an exact copy of the teacher model, called self-transfer learning in [7] and recently [13]. More details of this framework can be referred to [7].

## 3. SEQUENCE LEVEL CORRELATION MODELING

### 3.1. Temporal correlation modeling

For blind source separation, time-frequency sparsity and morphological diversity are assumed. Thus, the frequency bins between adjacent frames of the same speaker are correlated and used as the key hint in blind source separation.



**(a) Speaker Tracing**

**(b) Temporal Correlated Speaker Tracing**

**Fig. 2**. *Temporal Correlated Speaker Tracing Module.*

Nevertheless, previous deep learning based methods model the temporal correlation implicitly. Namely the temporal correlation modeling is only operated in feature level, and the label independence is introduced in both modeling and the inference. For $N$ speakers, given the mixed data $\mathbf{O}_u^{(m)}$, the model infers an acoustic representation $o_{utn}$ for each speaker $n$ at frame $t$ of utterance $u$.

$$o_{utn} = \mathcal{F}_{utn}(\mathbf{O}_u^{(m)}) \tag{1}$$

where $\mathcal{F}_{utn}(\cdot)$ is the neural network model output of speaker $n$ at frame $t$ in utterance $u$, and in [11], it's BLSTMs to model the whole feature sequence $\mathbf{O}_u^{(m)}$.

We believe there should be some output patterns in ideal binary or ratio masks of the same speaker. These output patterns can be potentially utilized to improve mask estimation, as the output patterns represent some kind of regularization that the estimated masks should keep in accordance with.

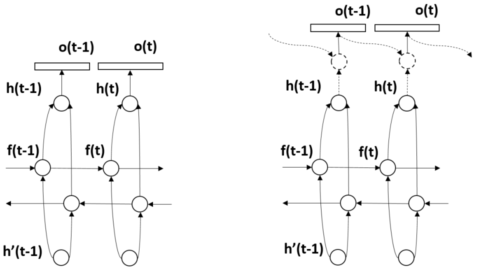
In this work, we propose to predict $o_{utn}$ not only from the feature sequence, but also from the predicted result of the last frame of the same output stream $o_{u(t-1)n}$,

$$o_{utn} = \mathcal{F}'_{utn}(\mathbf{O}_u^{(m)}, o_{u(t-1)n}) \tag{2}$$

where $\mathcal{F}'_{utn}(\cdot)$ is the proposed neural network architecture, which can be implemented by a recurrent connection between the output of the same stream at the last frame, $o_{u(t-1)n}$, and the stream-dependent hidden state before the current output. Figure 2 shows the new recurrent connections in the speaker tracing module [1].

Comparing the temporal correlated structure in the last layer with the original BLSTM implementation in [11], the advantages are two-fold: Firstly, the correlation between adjacent outputs is enhanced, although BLSTM already has strong temporal modeling effect in the whole sequence. Figure 3 shows the example of the temporal correlated structure in BLSTM implementation. For the

[1] Different from the idea proposed in Section 3.2, this structure is retained both in training and inference.



(a) BLSTM (b) Temporal Correlated BLSTM

**Fig. 3**. *Temporal Correlated Structure in BLSTM. The dot-dot lines and nodes are the structure inserted in the last layer of the original BLSTM implementation. One layer of BLSTM with one forward layer and one backward layer is taken as an example.*

original BLSTM, the error signal follows back-propagation through time (BPTT) paths, e.g. $o(t) \rightharpoonup h(t) \rightharpoonup f(t) \rightharpoonup f(t-1) \rightharpoonup h'(t-1) \rightharpoonup \dots$ . Thus the error signal from the output of frame $t$, $o(t)$, cannot directly communicate with the last hidden layer $h(t-1)$ of the last output $o(t-1)$. In the proposed structure, one of the BPTT paths is, $o(t) \rightharpoonup o(t-1) \rightharpoonup h(t-1) \rightharpoonup \dots$ . Thus the correlation between adjacent outputs is enhanced. Especially, the temporal correlation is essential in the problem as discussed before. Moreover, the enhanced correlation helps classifiers to keep decorrelation between output streams, which alleviates the cross talk errors, i.e. one person says a word, but it appears in both streams. Secondly, making final layer with a single direction also meets the monotonic characteristic in speech and helps the model to converge.

### 3.2. Language Model Integration

In the original PIT, the output-target assignment is decided by the minimum error at the utterance level based on the forward-pass result as below,

$$\mathcal{J}_{\text{U-PIT-CE}} = \sum_u \min_{s' \in \mathbf{S}} \sum_t \frac{1}{N} \sum_{n \in [1,N]} CE(l_{utn}^{(s')}, l_{utn}^{(r)}) \tag{3}$$

where, $\mathcal{J}_{\text{U-PIT-CE}}$ is the objective function of PIT-based speech recognition, PIT-ASR [11]. $\mathbf{S}$ is the permutation set of the reference label and the inference output. $l_{utn}^{(s')}$ is the $n$-th inference label of permutation $s'$ at frame $t$ in utterance $u$ and $l_{utn}^{(r)}$ is the corresponding transcription label obtained by clean speech forced-alignment [14].

To improve the permutation assignment, we propose to utilize both acoustic knowledge, PIT-trained model, and linguistic knowledge, prior probability from language model. Namely, in permutation assignment stage, the $CE(\cdot)$ is replaced with maximum a posteriori (MAP) decision process, $MAP(\cdot)$.

$$MAP(l_{utn}^{(s')}, l_{utn}^{(r)}) = \frac{P(l_{utn}^{(r)}|\mathbf{O}_u^{(m)})/P(l) \cdot P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')})}{P(\mathbf{O}_u^{(m)})} \tag{4}$$

$$\approx \frac{P(l_{utn}^{(r)}|\mathbf{O}_u^{(m)})}{P(l)} \cdot (\ P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')})\ )^{\lambda} \tag{5}$$

where $P(l_{utn}^{(r)}|\mathbf{O}_u^{(m)})$ is the acoustic model probability of the reference label $l_{utn}^{(r)}$ given the overlapped feature input $\mathbf{O}_u^{(m)}$. $P(l)$ is the prior probability of the output label. $\mathbf{L}_{u(t-1)n}^{(s')}$ is the inference label sequence up to frame $t-1$ of speaker $n$ in permutation $s'$ at utterance $u$. $P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')})$ is the language model probability of the current reference label $l_{utn}^{(r)}$ given the history sequence $\mathbf{L}_{u(t-1)n}^{(s')}$. As the output label is senones, a senone level neural network language model (NNLM) is proposed to model $P(l_{utn}^{(r)}|\mathbf{L}_{u(t-1)n}^{(s')})$. The NNLM is trained on the transcription alignment of the training data and its parameters are fixed in PIT joint training. $\lambda$ is the language model weight added in Equation (5).

The normalization term $P(\mathbf{O}_u^{(m)})$ in Equation (4) is ignored for simplicity as the proposed method uses NNLM, the history can not be truncated, which results in hardness in the search space modeling. Notably, the optimization stage does not take this approximation. Given the assignment obtained from $MAP(\cdot)$, the parameter of PIT-trained acoustic model is still updated by $CE(\cdot)$ for respective stream. Namely, the proposed method aims to improve the permutation assignment but not the optimization in training stage, whereas the pure acoustic model is combined with a more powerful word level language model in decoding stage.

## 4. RELATION TO PRIOR WORK

In this work, the blind source separation acoustic model is given by both the feature sequence and the output of the last frame. There are several works introducing recurrent connections between outputs of adjacent frames. A similar structure and the corresponding sequence criterion are proposed in ASR [15], to alleviate conditional independence assumption in CTC. [16] proposes to use the recurrent connections with specific training strategy to replace recurrent neural network (RNN) in supervised speech separation. It also shows the importance of recurrent connections in the speech separation. Nevertheless, the proposed method differs in two aspects. Firstly, the task addressed in this paper is un-supervised single-channel overlapped speech recognition. PIT is used to solve the further assignment decision problem, which replaces the supervised multi-speakers objective function. Secondly, the proposed method applies the structure in BLSTMs and shows that even based on RNNs, the recurrent connection between outputs of adjacent frames is important.

Moreover, a senone level neural network language model (NNLM) trained in the clean speech alignment is integrated in the permutation assignment, while the objective function is CE. In end-to-end system, NNLM can also be integrated with the acoustic model and jointly trained [17]. Nevertheless, to train a pure acoustic model and combine it with more powerful word level language model, the proposed method does not combine acoustic model and language model together. Although multi-output sequence discriminative training [7] also uses linguistic information to solve the single-channel overlapped speech recognition problem, there are three fundamental differences. Firstly, [7] uses a MAP formulation for the assignment decision and the objective function, whereas the proposed method uses a MAP formulation for the assignment decision only. Secondly, the proposed method uses NNLM for better modeling effect in language model while n-gram language model is used in sequence discriminative training for tractable search space modeling. Thirdly, [7] specifically designs a search space of the multi-outputs to calculate $P(\mathbf{O}_u^{(m)})$, while the proposed method need to ignore $P(\mathbf{O}_u^{(m)})$ because of the hardness in modeling discussed previously. With $P(\mathbf{O}_u^{(m)})$ modeling, [7] does discriminative training with competing hypotheses modeling. The proposed method uses NNLM to model the sequential labels correlation better. Thus totally speaking, two methods are operated in different levels, and can be combined together, shown in Section 5.4.

## 5. EXPERIMENT

### 5.1. Experimental Setup and Baseline Performance

The artificially overlapped Switchboard corpus and Switchboard (SWB) subset of the NIST 2000 CTS test set is used as in [7].After

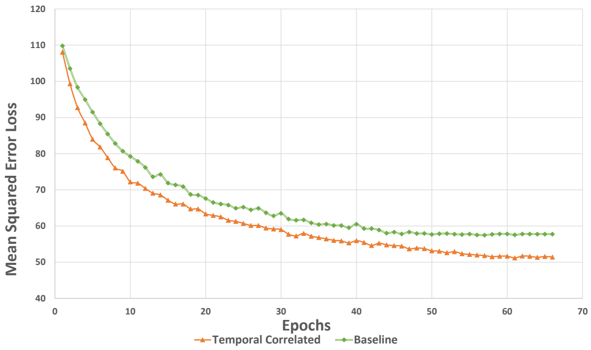


**Fig. 4**. *Validation Curves of Temporal Correlated Structure and the original BLSTM in Speaker Tracing. Each epoch contains 24 hours of data.*

overlapping, there's 150 hours training data, and 915 utterances in the testset. After decoding, there are 1830 utterances for evaluation, and the shortest utterance in the hub5e-swb dataset is discarded. Additionally, we define a small training set, the *50 hours dataset*, as a random 50 hour subset of the *150 hours dataset*. Without specific notation, experiments are reported on the 50 hours dataset [2].

In the training stage, 80-dimensional log-filterbank features [3] were extracted every 10 milliseconds, using a 25-millisecond analysis window. All neural networks were trained with the Microsoft Cognitive Toolkit (CNTK) [18]. Models use three state left-to-right triphone models with 9000 tied states (senones) [19]. The baseline model is trained by the transfer learning based progressive joint training method as in [7] with the same model setup. The speaker tracing module consists of 6 bidirectional LSTM layers with 768 memory cells in each layer and directly outputs multiple channels of the 80 dimensional log Mel-frequency features the speech recognition module expects. The speech recognition module, pretrained as a clean speech model, is composed of 4 bidirectional LSTM layers with 768 memory cells in each layer.

The evaluation was performed as in [7]. The baseline performances in this corpus are listed in Table 1. The PIT-ASR system proposed in [11] is in the first row and the transfer learning based progressive joint training system [7] is in the second row. The performance gap between them comes from better model generalization discussed in [7]. The second row is taken as the baseline for the latter comparison.

**Table 1**. *Baseline Performance of Transfer Learning Based Progressive Joint Training.*

| Neural network | Model | WER |
|---|---|---|
| 6 BLSTM + 4 BLSTM | PIT-ASR | 57.5 |
| | progressive joint training + clean teacher | **38.9** |

### 5.2. Temporal correlation modeling

Figure 4 shows the training curve of speaker tracing module with and without temporal correlated structure. A moderate gap can be observed in the figure, while the curve without temporal correlated structure also converges earlier, which shows more powerful modeling effect in temporal correlated structure.

Table 2 shows the performance of temporal correlation modeling combined with transfer learning based progressive joint training. The baseline is in the first row without temporal correlated structure. The second to fourth rows show the temporal correlated structure with different numbers of non-linear layers between the output layer of the last frame and that of the current frame, namely the number of non-linear layers in the dot-dot nodes in Figure 3(b). As a single non-linear layer denoted in the third row is the best, we use this configuration for the remainder of the experiments.

**Table 2**. *Effect of Temporal Correlation Modeling*

| Temporal Correlated | # of non-linear | WER | Rel. (%) |
|---|---|---|---|
| × | 0 | 38.9 | 0 |
| √ | 0 | 37.5 | -3.6 |
| | 1 | **35.8** | **-8.0** |
| | 2 | 36.7 | -5.7 |

### 5.3. Language Model Integration

Table 3 shows the performance of the proposed language model integration method, namely deciding assignment by $MAP(\cdot)$ and optimizing by $CE(\cdot)$ discussed in Section 3.2. $\lambda = 10$ and the perplexity (PPL) of NNLM in the training and development sets are 12.6 and 13.0 respectively.

The proposed method achieves improvement versus the baseline using $CE(\cdot)$ in both assignment decision and optimization stages. Besides, with more data, the relative improvement becomes larger. We believe as the acoustic model becomes stronger, the assignment decision is no longer prone to over-fit to the language model.

**Table 3**. *Language Model Integration in Transfer Learning Based Progressive Joint Training*

| | | 50 hours | | 150 hours | |
|---|---|---|---|---|---|
| Assign. | Opt. | WER | Rel. (%) | WER | Rel. (%) |
| CE | CE | 38.9 | 0 | 32.8 | 0 |
| MAP | CE | 37.3 | **-4.1** | 30.9 | **-5.8** |

### 5.4. Combination

Table 4 firstly combines the proposed methods together. Besides, the proposed methods can be further combined with multi-outputs sequence discriminative training proposed in [7]. It shows that the proposed language model integration method and sequence discriminative training are operated in different levels and can be combined together.

**Table 4**. *Combining the Proposed Methods and Sequence Discriminative Training. All systems are based on transfer learning based progressive joint training.*

| Method | WER | Rel. (%) |
|---|---|---|
| baseline | 38.9 | 0 |
| + Temporal Correlated | 35.8 | -8.0 |
| + LM Integration | 34.4 | -11.5 |
| + LF-DC-bMMI | 31.6 | -18.8 |

## 6. CONCLUSION

In this paper, the sequence modeling is improved by two strategies: temporal correlation modeling and language model integration. The proposed methods are combined with an improved version of PIT and sequence discriminative training, which brings about further over 10% improvement. Future works include the combination of the acoustic and language models joint training [17], adaptive training [20] and the end-to-end sequence modeling [21, 22, 23].

[2] [7] reveals that the transfer learning based progressive joint training method, baseline, works well in 50 hours dataset, which consistently shows 10-20% relative performance gaps compared with 150 hours dataset.

[3] Preliminary experiments show it is better in this task.

## 7. REFERENCES

[1] E Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] Albert S Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[3] DeLiang Wang and Guy J Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[4] Martin Cooke, John R Hershey, and Steven J Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[5] Jun Du, Yanhui Tu, Yong Xu, Lirong Dai, and Chin-Hui Lee, "Speech separation of a target speaker based on deep neural networks," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 473–477.

[6] 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 (TASLP)*, vol. 23, no. 10, pp. 1670–1679, 2015.

[7] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[8] Max Wertheimer, "Laws of organization in perceptual forms.," 1938.

[9] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[10] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[11] Dong Yu, Xuankai Chang, and Yanmin Qian, "Recognizing multi-talker speech with permutation invariant training," *CoRR*, vol. abs/1704.01985, 2017.

[12] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[13] Tian Tan, Yanmin Qian, and Yu Dong, "Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition," in *ICASSP*, April 2018.

[14] Philip C Woodland, Julian J Odell, Valtcho Valtchev, and Steve J Young, "Large vocabulary continuous speech recognition using htk," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. Ieee, 1994, vol. 2, pp. II–125.

[15] Sak Hasim, Shannon Matt, Rao Kanishka, and Francoise Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping," in *INTERSPEECH*, 2017.

[16] Zhong-Qiu Wang and DeLiang Wang, "Recurrent deep stacking networks for supervised speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 71–75.

[17] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[18] Frank Seide and Amit Agarwal, "Cntk: Microsoft's open-source deep-learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2135–2135.

[19] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[20] Xuankai Chang, Yanmin Qian, and Dong Yu, "Adaptive permutation invariant training with auxiliary information for monaural multi-talker speech recognition," in *ICASSP*, April 2018.

[21] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[22] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[23] Zhehuai Chen, Qi Liu, Hao Li, and Kai Yu, "On modular training of neural acoustics-to-word model for lvcsr," in *ICASSP*, April 2018.