ADAPTIVE PERMUTATION INVARIANT TRAINING WITH AUXILIARY INFORMATION FOR MONAURAL MULTI-TALKER SPEECH RECOGNITION

Xuankai Chang[1], *Yanmin Qian*[1†], *Dong Yu*[2]

[1]SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]Tencent AI Lab, Tencent, Bellevue, WA, USA
{xuank@sjtu.edu.cn, yanminqian@tencent.com, dyu@tencent.com}

ABSTRACT

In this paper, we extend our previous work on direct recognition of single-channel multi-talker mixed speech using permutation invariant training (PIT). We propose to adapt the PIT models with auxiliary features such as pitch and i-vector, and to exploit the gender information with multi-task learning which jointly optimizes for the speech recognition and speaker-pair prediction. We also compare CNN-BLSTMs against BLSTM-RNNs used in our previous PIT-ASR model. The experimental results on the artificially mixed two-talker AMI data indicate that our proposed model improvements can reduce word error rate (WER) by $\sim$10.0% relative to our previous work for both speakers in the mixed speech. Our results also confirm that PIT can be easily combined with advanced techniques to improve the performance on multi-talker speech recognition.

Index Terms— permutation invariant training, multi-talker speech recognition, speaker adaptation, auxiliary features

1. INTRODUCTION

Over the past few years, due to the advances in deep learning technology, the performance on single-talker speech recognition has been significantly improved and has even reached human parity in some scenarios once considered very difficult [1]. However, we still suffer from obvious degradations on automatic speech recognition (ASR) performance when the interfering signals, such as background noise, reverberation and speech from other talkers, cannot be ignored.

In this paper, we focus on the scenario where multiple talkers speak at the same time and only a single channel of mixed speech is available. Many attempts have been made to attack this problem, but the results so far are still far from satisfaction [2, 3]. The main difficulty comes from the "label ambiguity" or "label permutation" problem. In recent years, many works have been conducted to address this problem. Weng et al. [4] adopted a deep learning model to recognize the mixed speech directly by assigning the senone labels of the talkers according to the energy of the speech. To deal with the speaker switch problem, a two-talker joint-decoder with a speaker switching penalty was used to trace speakers. Hershey et al. [5, 6] proposed a technique called deep clustering (DPCL) to separate the speech streams by mapping a speaker's time-frequency bins into an embedding space where the bins belong to the same speakers are close and that of different speakers are far away from each other. Chen et al. [7] used a technique called deep attractor network (DANet) which learns a high-dimensional embedding of the acoustic signals and clustered embeddings with attractor points. Yu et al. [8, 9, 10, 11, 12, 13, 14] proposed a simple and effective technique named permutation invariant training (PIT) which trains a deep neural network by minimizing the average minimum error with the best output-target assignment at the utterance level.

Despite the progresses made in monaural multi-talker speech recognition, the word error rates (WER) reported in previous works are still much higher than that in single-talker cases [5, 9]. In single-talker speech recognition, speaker adaptation reduces the mismatch between the training and the test speakers and improves the WER for the test speakers. There are several categories of neural network adaptation techniques for single-talker speech recognition [15]. One approach is to adapt the acoustic features through the feature space transformation, such as CMLLR (or fMLLR) [16]. Another approach is adapting all or a subset of parameters of neural networks [15, 17, 18, 19]. The third approach uses auxiliary features in an adaptive training mode and makes the model aware of the variations. The auxiliary features can be i-vector for speaker [20], noise code for noise [21], and T60 for reverberation [22].

In this paper, we investigate how adaptation techniques perform on monaural multi-talker speech recognition. The auxiliary feature assisted adaptive training is developed for the PIT-ASR model [9]. The assumption is that the appropriate speaker-dependent feature and structure can make the speaker tracing easier in PIT and lead to better recognition accuracy. The auxiliary features explored here include pitch and i-vectors of the mixed utterance. We also propose to exploit the gender information in the mixed speech with a multi-task learning architecture that jointly optimizes for the speech recognition and gender-pair prediction. Significant WER reduction is observed on the artificially mixed AMI data with these model improvements.

The rest of the paper is organized as follows. In Section 2 we introduce permutation invariant training for monaural multi-talker speech recognition. In Section 3 we describe the convolutional neural network (CNN)-long short-term memory (LSTM) recurrent neural networks (RNNs). We describe the auxiliary feature based PIT adaptation technique and the multi-task learning framework in Section 4, and present experimental results in Section 5. We conclude the paper in Section 6.

†Yanmin Qian and Dong Yu are the corresponding authors and now Yanmin Qian is with Tencent AI Lab, Tencent, Bellevue, WA, USA.

This work was partly supported by the Tencent-Shanghai Jiao Tong University joint project, the China NSFC projects (No. U1736202 and No. 61603252), and the Shanghai Sailing Program No. 16YF1405300. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

2. PERMUTATION INVARIANT TRAINING FOR MULTI-TALKER SPEECH RECOGNITION

We assume that a linearly mixed single microphone signal $\mathbf{y}[n] = \sum_{s=1}^{S} \mathbf{x}_s[n]$ is given, where $\mathbf{x}_s[n], s = 1, \cdots, S$ are S streams of speech sources from different speakers. The goal is to separate and recognize these streams. In the case of $S >= 2$, assigning the correct target to the corresponding output layer could be difficult because speech sources are symmetric given the mixture (i.e., $\mathbf{x}_1 + \mathbf{x}_2$ equals to $\mathbf{x}_2 + \mathbf{x}_1$ when $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same characteristics), which is referred to as the "label permutation problem".

In our previous work [9], a deep bidirectional LSTM takes features $\mathbf{Y}$ of the mixed speech $\mathbf{y}$ as inputs, and outputs S individual speech streams $\mathbf{O}^s, s = 1, \cdots, S$, which is the output segment for stream s. In the training process, we adopt PIT and minimize the objective function

$$J = \frac{1}{S} \min_{s' \in permu(S)} \sum_{s} \sum_{t} CE(\ell_t^{s'}, \mathbf{O}_t^s), s = 1, \cdots, S \quad (1)$$

where $permu(S)$ is a permutation of $1, \cdots, S$. The architecture of PIT-ASR model is shown in Figure 1. Note that PIT automatically finds the appropriate assignment no matter how the labels are ordered, and solves the label permutation problem and speaker tracing problem by computing the cross entropy (CE) over the whole sequence for each assignment. Compared to DPCL [5] or DANet [7], this structure is much simpler and more compact since it allows direct multi-talker mixed speech recognition without explicit separation. After the PIT model training, the individual output posterior streams can be used for decoding as normal to obtain the final recognition result.

3. CONVOLUTIONAL NEURAL NETWORK - LONG SHORT-TERM MEMORY NEURAL NETWORK WITH PIT

Considering that the recurrent neural networks (RNNs) can take advantage of the long-range dependency and improve speaker tracing, we used a pure deep bidirectional LSTM-RNN (BLSTM-RNN) in our previous work [9, 11]. The convolutional neural network (CNN), as an alternative neural network structure, has shown promising results in some single-talker speech recognition tasks [23, 24, 25, 26]. The speech signals have structures along both time and frequency axes. However, conventional RNNs only model correlation along the time axis and ignore the structure along the frequency axis, which contains useful information for speaker tracing. In this work we introduce the convolutional operation into the PIT-ASR model. The model uses convolutional operations to extract shift-invariant features from speech signals and BLSTMs to perform speaker tracing and speech separation and recognition, as shown in Figure 2. This CNN-BLSTM architecture computes

$$\mathbf{H}_0 = \mathbf{Y} \quad (2)$$

$$\mathbf{H}_i = CNN_i(\mathbf{H}_{i-1}), i = 1, \cdots, N^C \quad (3)$$

$$\mathbf{H}_i^f = LSTM_i^f(\mathbf{H}_{i-1}), i = N^C + 1, \cdots, N^R \quad (4)$$

$$\mathbf{H}_i^b = LSTM_i^b(\mathbf{H}_{i-1}), i = N^C + 1, \cdots, N^R \quad (5)$$

$$\mathbf{H}_i = Stack(\mathbf{H}_i^f, \mathbf{H}_i^b), i = N^C + 1, \cdots, N^R \quad (6)$$

$$\mathbf{H}_o^s = Linear(\mathbf{H}_{N^R}), s = 1, \cdots, S \quad (7)$$

$$\mathbf{O}^s = Softmax(\mathbf{H}_O^s), s = 1, \cdots, S \quad (8)$$

where $\mathbf{H}_0$ is the input, N^C and N^R are the layer indices of the last CNN and LSTM layers. $LSTM_i^f$ and $LSTM_i^b$ are the forward and backward LSTMs at hidden layer i respectively. $\mathbf{H}_o^s, s = 1, \cdots, S$ is the excitation at output layer for each speech stream s. Note that, each output layer represents an estimate of the senone posterior probability for a speech stream. No additional clustering or speaker tracing is needed. The acoustic model is trained by minimizing the objective function as in Eq (1).

4. AUXILIARY FEATURE ASSISTED ADAPTATION

4.1. Speaker Characterizing Features

We conjecture that making the multi-talker model speaker-aware can help speaker tracing and improve speech separation and recognition. For this reason, we explored speaker adaptation techniques for the PIT-ASR model with speaker characterizing features such as pitch, i-vector, and gender-pair.

Pitch is important information to differentiate speakers. For example, the F0 of female is usually higher than that of male. In this work, pitch is used as an auxiliary feature. Various pitch estimators have been developed[27, 28, 29, 30]. In our experiment, we used Kaldi [31] to extract pitch features from the mixed speech. The basic idea is to find the lag values that maximize the Normalized Cross Correlation Function (NCCF). The output of the pitch extraction tool is the pitch and NCCF at each frame.

i-vector is considered good representation of speaker identity and is widely used to recognize speakers [32]. Recently, it has been used in speaker adaptation of single-talker speech recognition [33]. To extract i-vectors, we first derive a super-vector M from the universal background model (UBM) to represent the combination of speaker and session. The probability model of the super-vector is

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \quad (9)$$

where $\mathbf{m}$ is a speaker- and session-independent super-vector, $\mathbf{T}$ is a low rank matrix which captures the speaker and session variability, and i-vector is the posterior mean of $\mathbf{w}$. In this work, we integrate the i-vector estimated from the mixed-speech into PIT-ASR model, and make the multi-talker model speaker-aware.

4.2. Adaptive Training with Auxiliary Information

To adapt the acoustic model to a certain speaker-pair, we can provide the acoustic model with speaker characterizing features as auxiliary features. For BLSTM-RNNs, this is done by simply augmenting the speech features with auxiliary features. For CNN-BLSTMs, however, things are a little bit more complicated. This is because directly applying convolutional operations to i-vectors is not effective [34]. To solve this problem, we add a transformation layer, shown as the bottom-right dashed rectangular in Figure 2, to convert the auxiliary features $\mathbf{Y}^{Aux}$ to an intermediate representation. The transformed representation is combined with the feature maps from the convolutional layers and fed into the following BLSTM layers.

4.3. Exploit Gender-pair Information with Multi-Task Learning

The vocal tract lengths for males and females are notably different. However, they are very close for same-gender speakers. It has been reported that multi-talker speech separation [35] and recognition [11] are much harder on same-gender mixed speech than on opposite-gender mixed speech. One solution, which would make the system very complex, is to train a separate model for each

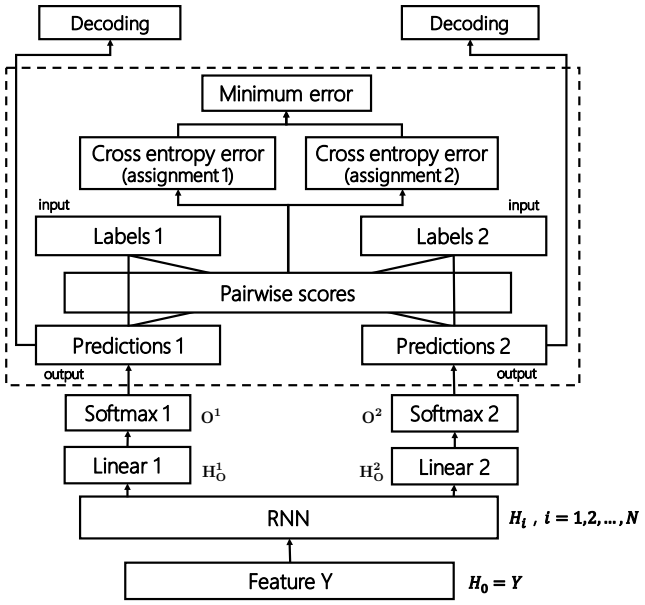

Fig. 1. Two-talker speech recognition with permutation invariant training

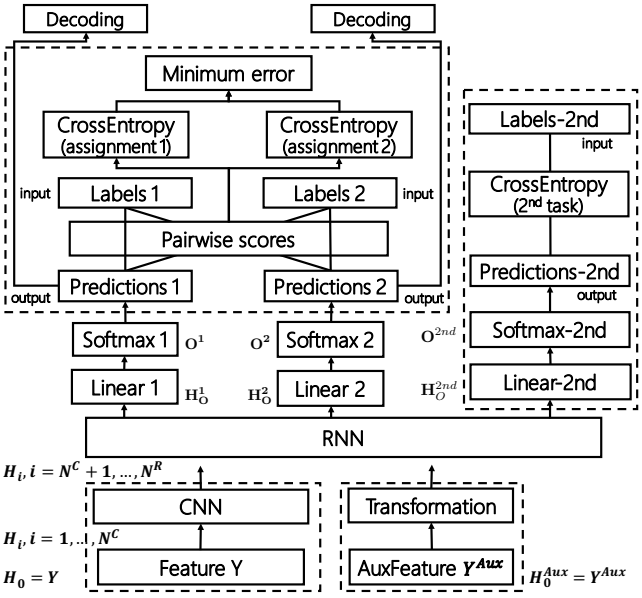

Fig. 2. Auxiliary-feature assisted adaptive PIT-ASR model with CNN-BLSTMs

gender combination. In this work, we take a different yet simpler solution by providing gender-pair information to the model at training time. More specifically, the gender-pair is encoded as a 3-dimensional one-hot vector to represent the conditions of *Male+Male*, *Female+Female*, *Opposite-Gender*.

In the adaptation technique described in Section 4.2, we need to estimate the auxiliary feature in both training and testing. However, it is extremely hard, if not impossible, to reliably estimate features, such as the gender-pair vector, given only the mixed speech during testing.

To address this problem, we propose a multi-task learning framework to exploit the gender-pair information without estimating them during testing. More specifically, we train the PIT-ASR model with an additional gender classification task. The cross entropy criterion is used for this additional task and the final objective function is

$$J^{MTL} = J + \lambda \sum_t CE(\ell_t^{2nd}, \mathbf{O}_t^{2nd}) \tag{10}$$

where J^{MTL} is the criterion for multi-task learning, ℓ^{2nd} and $\mathbf{O}^{2nd}$ are labels and model predictions for the second task, and J is the major task using the PIT criterion (Eq (1)). The hyper-parameter λ can be tuned on a development set. The structure of the multi-task training component is shown as the top-right dashed rectangular in Figure 2. After model training, the parameters and components related only to the secondary task can be removed and the model can be evaluated without estimating the auxiliary information (such as the gender-pair vector) related to the secondary task.

5. EXPERIMENTS

To evaluate the proposed methods, we train and test the models on our released artificially mixed two-talker AMI IHM corpus [11]. The 80hr training subset is used in this work for fast evaluation, and the 8hr two-talker mixed speech with all SNR levels is used in evaluation. Additional details on the multi-talker speech corpus can be found in [11].

Kaldi [36] is used for pitch and i-vector extraction. 3-dimensional pitch features are extracted from the mixed speech with all the parameters set to default values. The i-vector extractor is trained using 13-dimensional MFCC of all the data in the training set of over 100 speaker pairs. The UBM has 2048 diagonal Gaussians, and the dimension of i-vectors is 10. The basic acoustic features used in training all the PIT models are 40-dimensional log filter bank. All neural network acoustic models are trained using the Microsoft Cognitive Toolkit (CNTK) [37] using 1-bit SGD [38] on 2 GPUs. The learning rate is $2e-4$ and gradients are clipped with the threshold $3e-4$.

In the decoding, we use a 50K-word dictionary and a trigram language model interpolated from the ones created using the AMI transcripts and the Fisher English corpus. For scoring, we evaluate the hypotheses pairwisely against the two references, and make the assignment with better WER as the final recognition results for each utterance.

5.1. PIT-ASR with CNN-BLSTMs

Firstly we construct the normal PIT-ASR model, that introduced in our previous work [9], as the baseline. It is a 6-layer BLSTM-RNN with 768 memory cells in each layer. The output of the last LSTM layer is sent directly to two output layers with softmax activation, representing two recognition streams. These two outputs are then used in decoding to obtain the hypotheses for two talkers. The baseline results are shown at the top of Table 1. Our proposed PIT-ASR model can recognize multi-talker speech directly. However, compared to the normal single-talker ASR on AMI IHM (around 27.0% WER without adaptation), there is still a large gap for both speakers. We also show WERs under different gender combinations. It is observed that the recognition accuracy is significantly worse when the two speakers are of the same gender.

The CNN-BLSTM contains 2 convolutional layers followed by

4 bidirectional LSTM layers. The input feature map is 11×40. The two convolutional layers apply a 9×9 kernel with stride 1×1 and a 3×3 kernel with stride 2×2, respectively. There are 32 and 64 feature maps in the two convolutional layers respectively. The number of memory cells is 768 in each LSTM layer. The results are reported at the bottom of Table 1. We can observe that the CNN-BLSTM model outperforms the BLSTM-RNN model by $\sim$6.0% relatively.

Table 1. WER (%) of the PIT-ASR model with different model structures and gender combinations

Model	Gender Combination	WER 1	WER 2
BLSTM	All	**55.21**	**64.23**
	opposite	52.41	61.61
	same	58.48	67.27
CNN-BLSTM	All	**51.93**	**60.13**
	opposite	49.40	57.90
	same	54.89	62.72

5.2. PIT with auxiliary feature based adaptation

The auxiliary feature based adaptation is evaluated. Pitch and i-vectors are used as auxiliary features in this architecture. Their integration with the PIT-ASR model is illustrated in Figure 2. For the BLSTM-RNN model, the auxiliary features are stacked with the FBank feature directly. In the CNN-BLSTM model, however, the auxiliary features are transferred through a 256-cell hidden layer and then stacked with the outputs of the CNN layers. From Table 2 we can observe that both pitch and i-vector can improve the recognition accuracy although they are both estimated from the mixed speech, and i-vector is more effective than pitch. We conjecture that the auxiliary features estimated from the mixed speech helps because these features provide adaptive bias to the model as discussed in [39]. As long as these features are consistent and provide information better performance can be expected.

Moreover, the combination of multiple auxiliary features leads to slight additional gain. Overall, the auxiliary feature based adaptation achieved relative 8.0% WER reduction on both speakers against the baseline.

Table 2. WER (%) of PIT-ASR with auxiliary feature based adaptation

Model	Adapt on	WER 1	WER 2
BLSTM	—	55.21	64.23
	pitch	51.88	60.54
	i-vector	51.61	59.99
	pitch + i-vector	51.29	59.78
CNN-BLSTM	pitch + i-vector	50.64	58.78

5.3. Exploit gender-pair information with multi-task Learning

As described in Section 4.3, the gender-pair prediction can be used as a second task to improve the system performance. In our experiments the λ in Eq (10) is set to 0.3. The results, shown in Table 3, indicate that multi-task learning with gender-pair estimation significantly outperforms the baseline accuracy reported in Table 1.

We further combine the auxiliary-feature based adaptation and the multi-task learning in an integrated framework as illustrated in Figure 2. The results reported in Table 3 show that the combined architecture further improves the performance, and the best system reduces WER by $\sim$10.0% relative to the baseline.

Table 3. WER (%) of PIT-ASR with multi-task learning

Model	2nd Task	Adapt on	WER 1	WER 2
BLSTM	—	—	55.21	64.23
	gender	—	52.47	60.31
		pitch+i-vector	51.11	59.35
CNN-BLSTM	—	—	51.93	60.13
	gender	—	51.10	58.76
		pitch+i-vector	**50.21**	**58.17**

6. CONCLUSION

In this paper, we extended our previous work on permutation invariant training for monaural multi-talker speech recognition. We showed that CNN-BLSTMs outperform BLSTM-RNNs with a big margin for both speakers in this task. We developed and evaluated the auxiliary feature assisted adaptation technique for the PIT-ASR model. Two types of auxiliary features, namely pitch and i-vector, were explored and evaluated. We further proposed to exploit the gender-pair information in the multi-task learning framework to improve the recognition accuracy. The results on the artificially mixed two-talker AMI corpus show that all the auxiliary features and the adaptation architectures help to boost recognition accuracy. The final framework with all ingredients integrated achieved the best performance. Our results also confirm that PIT can be easily combined with advanced techniques, such as the adaptation and multi-task learning evaluated in this work, to improve the performance on multi-talker speech recognition. This is an attractive property that makes PIT a nice modeling technique for multi-talker speech recognition.

7. REFERENCES

[1] Xiong et al., "The Microsoft 2016 conversational speech recognition system," in *ICASSP*, 2017, pp. 5255–5259.

[2] Zoubin Ghahramani and Michael I Jordan, "Factorial hidden markov models," in *Advances in Neural Information Processing Systems*, 1996, pp. 472–478.

[3] 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.

[4] Chao Weng, Dong Yu, Michael L. Seltzer, and Jasha Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *TASLP*, vol. 23, no. 10, pp. 1670–1679, 2015.

[5] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, 2016, pp. 31–35.

[6] Isik et al., "Single-channel multi-speaker separation using deep clustering," in *INTERSPEECH*, 2016, pp. 545–549.

[7] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *ICASSP*, 2017, pp. 246–250.

[8] Dong Yu, Morten Kolbk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245.

[9] Dong Yu, Xuankai Chang, and Yanmin Qian, "Recognizing multi-talker speech with permutation invariant training," in *INTERSPEECH*, 2017, pp. 2456–2460.

[10] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *TASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

[11] Yanmin Qian, Xuankai Chang, and Dong Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *submitted to Speech Communication, CoRR*, vol. abs/1707.06527, 2017.

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

[13] Zhehuai Chen and Jasha Droppo, "Sequence modeling in unsupervised single-channel overlapped speech recognition," in *ICASSP*, 2018.

[14] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *TASLP*, vol. 26, no. 1, pp. 184–196, Jan 2018.

[15] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *SLT*. IEEE, 2014, pp. 171–176.

[16] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[17] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *SLT*. IEEE, 2012, pp. 366–369.

[18] Sabato Marco Siniscalchi, Jinyu Li, and Chin-Hui Lee, "Hermitian polynomial for speaker adaptation of connectionist speech recognition systems," *TASLP*, vol. 21, no. 10, pp. 2152–2161, 2013.

[19] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*. IEEE, 2013, pp. 7893–7897.

[20] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors.," in *ASRU*, 2013, pp. 55–59.

[21] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *ICASSP*. IEEE, 2013, pp. 7398–7402.

[22] Ritwik Giri, Michael L Seltzer, Jasha Droppo, and Dong Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *ICASSP*. IEEE, 2015, pp. 5014–5018.

[23] Abdel-Hamid et al., "Convolutional neural networks for speech recognition," *TASLP*, vol. 22, no. 10, pp. 1533–1545, 2014.

[24] Sainath et al., "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[25] Yu et al., "Deep convolutional neural networks with layer-wise context expansion and attention.," in *INTERSPEECH*, 2016, pp. 17–21.

[26] Yanmin Qian, Mengxiao Bi, Tian Tan, and Kai Yu, "Very deep convolutional neural networks for noise robust speech recognition," *TASLP*, vol. 24, no. 12, pp. 2263–2276, 2016.

[27] Alain De Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[28] David Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[29] Mingyang Wu, DeLiang Wang, and Guy J Brown, "A multipitch tracking algorithm for noisy speech," *TSAP*, vol. 11, no. 3, pp. 229–241, 2003.

[30] Arturo Camacho and John G Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[31] Ghahremani et al., "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*, 2014, pp. 2494–2498.

[32] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *TSAP*, vol. 13, no. 3, pp. 345–354, 2005.

[33] Vishwa Gupta, Patrick Kenny, Pierre Ouellet, and Themos Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *ICASSP*. IEEE, 2014, pp. 6334–6338.

[34] Yanmin Qian and Philip C Woodland, "Very deep convolutional neural networks for robust speech recognition," in *SLT*. IEEE, 2016, pp. 481–488.

[35] Yannan Wang, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *TASLP*, vol. 25, no. 7, pp. 1535–1546, 2017.

[36] Daniel et al., "The kaldi speech recognition toolkit," in *ASRU*, 2011, number EPFL-CONF-192584.

[37] Amit et al., "An introduction to computational networks and the computational network toolkit," Tech. Rep., Microsoft Technical Report MSR-TR-2014-112, 2014.

[38] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs.," in *INTERSPEECH*, 2014, pp. 1058–1062.

[39] Dong Yu and Li Deng, *Automatic speech recognition: A deep learning approach*, Springer, 2014.