PERMUTATION INVARIANT TRAINING FOR SPEAKER-INDEPENDENT MULTI-PITCH TRACKING

Yuzhou Liu[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{liuyuz,dwang}@cse.ohio-state.edu

ABSTRACT

Speaker-independent multi-pitch tracking has been a long-standing problem in speech processing. In this study, we extend a recurrent neural network - factorial hidden Markov model (RNN-FHMM) framework, and use the utterance-level permutation invariant training (uPIT) criterion for multi-pitch tracking. Separated speech and label permutations from a speech separation uPIT-RNN have been further incorporated to improve pitch tracking performance. We evaluate our methods on the GRID database. Results indicate that the proposed speech separation - pitch tracking system with matched uPIT label permutations outperforms all other gender-dependent and speaker-independent multi-pitch trackers. The improvement is more significant for challenging same-gender mixtures.

Index Terms— Multi-pitch tracking, recurrent neural network, permutation invariant training, speech separation

1. INTRODUCTION

Conventional pitch tracking algorithms [2, 13, 15, 17] fail to produce consistent results when the target speech is interfered by a competing speaker. In other words, they can not track two pitch candidates simultaneously. The task is known as multi-pitch tracking in speech processing, and this paper is mainly concerned with multi-pitch tracking of two concurrent speakers.

During the last couple decades, lots of approaches have been proposed for multi-pitch tracking. Based on the speaker dependencies and speaker assignments of approaches, we can broadly categorize them into three groups: speaker-independent (SI) approaches without speaker assignment (SA), SI approaches with SA, and speaker-dependent approaches (SD) with SA. Many studies fall into the first category [1, 3, 10, 19]: SI without SA, where a speaker-independent model is built to track two pitch contours without assigning them to any specific speaker. Most of them consist of two stages: harmonic modeling and speaker-independent tracking. SI approaches without SA can reliably estimate pitch periods at the frame level, but they do not perform any speaker assignment, which makes them uninformative for applications like co-channel speech separation and SD emotion analysis. SI approaches with SA [5, 9] are introduced to address this problem, which cluster short-term pitch contours to sequentially group them into two speakers. However, they achieve limited success as individual pitch contours are usually too short to contain enough information for clustering.

On the other hand, SD approaches try to address the pitch assignment problem of by utilizing speaker information. For example, Wohlmayr et al. [18] train SD Gaussian mixture models (GMMs) to estimate the frame-level probability of pitch periods, and connect them with an SD factorial hidden Markov model (FHMM) for continuous pitch tracking. Our previous work [14] uses speaker-pair-dependent (SPD) deep neural networks (DNNs) to predict the pitch states of both speakers at each frame. An SPD-DNN has two output layers, each for a specific speaker, and it is trained only for a specific speaker pair. An FHMM is then adopted to track pitch of two speakers through time. We have shown that our approach significantly outperforms other existing SI and SD multi-pitch tracking algorithms. However, speaker-dependent information is usually unavailable for most real applications, therefore, in [14] we further propose gender-pair-dependent (GPD) DNNs to relax this constraint. Specifically, DNNs for three gender pairs, i.e., male-female, male-male and female-female, are trained using GPD data. For the male-female gender pair, we associate the first output layer with training labels of male speakers, and second output layer with female speakers. For same-gender pairs, we divide the training speakers into two groups, with the group A having lower average pitch, and group B having higher average pitch. The first output is assigned to speakers in group A, and the second output to group B. The GPD-DNN based approach works extremely well for male-female speaker pairs. For same-gender pairs, our solution is still suboptimal. The grouping criterion are based on the global pitch properties of speakers. However, because same-gender pairs have very close pitch ranges, in certain utterances the order of average pitch may flip for the two groups, which leads to conflicting gradients during the training process.

Recently, an utterance-level permutation invariant training (uPIT) [12] algorithm is proposed to deal with the output permutation problem. Instead of using a predefined ordering of the two labels, uPIT uses a bi-directional long short-term memory recurrent neural network (BLSTM-RNN) to jointly optimize the label assignment and training error end-to-end. When training the BLSTM-RNN, uPIT allows two possible label permutations for the two speakers within an utterance, and only use the one with the lowest loss to update the network. In this way, it finds a locally optimized solution to the output assignment. uPIT leads to substantial improvement in speaker-independent co-channel speech separation. In this study, we apply the uPIT method to train an SI BLSTM-RNN for multi-pitch estimation. We follow our framework in [14], and compare the SI uPIT-BLSTM with SPD-BLSTM and GPD-BLSTM for pitch states estimation. Later, we find that by performing multi-pitch tracking alone, uPIT does not generate promising results for same-gender speaker pairs. Thus we further extend our system by using speech separation (SS) uPIT-BLSTM as the front-end. Three new structures are explored. In the first structure, we try to directly apply the RAPT [17] single-pitch tracking algorithm after uPIT

This research was supported in pspectraart by an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

based speech separation. The second structure concatenates SS with multi-pitch tracking, by using the outputs of uPIT-SS as additional features for uPIT based multi-pitch tracking. Lastly, we modify the second structure by using the label permutation in uPIT-SS for the multi-pitch tracking network. Consistent improvement is achieved by using the third structure.

In the remainder of this paper, after reviewing the SPD and GPD systems in Section 2, the uPIT based systems are introduced in Section 3. Section 4 describes the FHMM for multipitch tracking. Experimental results and comparisons are presented in Section 5. A conclusion is given in Section 6.

2. SPEAKER-PAIR AND GENDER-PAIR DEPENDENT PITCH PROBABILITY ESTIMATION

2.1. Overview

Our previous pitch tracking algorithm [14] consists of two stages: pitch probability estimation and FHMM. We introduce the first stage in this section. The input to the system is a speech mixture v_t: $v_t = u_t^1 + u_t^2$, where u_t^1 and u_t^2 are utterances of two speakers. Given the mixture, our system first extracts frame-level log magnitude short-time discrete Fourier transform (STFT) features $\mathbf{y}_m$. We then feed $\mathbf{y}_m$ into neural networks to estimate the posterior pitch probabilities at frame m, i. e., $p(x_m^1, x_m^2|\mathbf{y}_m)$. x_m^1 and x_m^2 denote pitch states of the two speakers at frame m. We quantize the frequency range 60 to 404 Hz into 67 bins using 24 bins per octave in a logarithmic scale. Each bin corresponds to one pitch state. An additional pitch state represents silence or unvoiced speech. $p(x_m^1(s_1), x_m^2(s_2)|\mathbf{y}_m)$ equals one if groundtruth pitches fall into the s_1^{th} and s_2^{th} frequency bins respectively. Since BLSTM-RNNs [8] can make better use of the temporal context, we use BLSTM-RNNs instead of DNNs in this study.

2.2. SPD-BLSTM for Pitch Probability Estimation

An SPD-BLSTM is a BLSTM-RNN trained on a specific pair of speakers. There are two 68-unit softmax output layers in SPD-BLSTM, with each one estimating pitch state of the i^{th} speaker $p(x_m^i|\mathbf{y}_m)$. Denoting the i^{th} output layer at frame m by O_m^i, we can write the frame-level cross-entropy loss as:

$$J_m = -\sum_{i=1}^{2}\sum_{s=1}^{68} p(x_m^i(s)|\mathbf{y}_m)\log(O_m^i(s)) \tag{1}$$

where s is the index for 68 pitch states.

The final frame-level pitch probability is estimated by:

$$p(x_m^1, x_m^2|\mathbf{y}_m) = O_m^1 O_m^2 \tag{2}$$

This model is denoted by SPD-PITCH in the rest of the paper.

2.3. GPD-BLSTM for Pitch Probability Estimation

SPD-PITCH is not applicable to untrained speakers, thus we introduce GPD models in [14] to relax this constraint, denoted by GPD-PITCH. GPD-PITCH adopts the same structure as SPD-PITCH. Three sets of GPD data, including male-female, male-male and female-female, are generated to train different GPD-BLSTM-RNNs. For the male-female GPD-BLSTM-RNN, we associate O_m^1 with male speakers, O_m^2 with female speakers. For same-gender GPD-BLSTM-RNNs, we divide the speakers in the training set in to two groups w.r.t. the overall average pitch. O_m^1 is paired with the group of with lower average pitch, and O_m^2 is paired with speakers with higher average pitch. During testing, a gender-pair detection algorithm can be used to assign test samples to their corresponding gender pairs. GPD-PITCH is then used for pitch probability estimation.

3. UPIT FOR SI PITCH PROBABILITY ESTIMATION

3.1. uPIT based SI Multi-pitch Tracking

The label layout in GPD-PITCH provides a reasonable way to differentiate two speakers with little information available. It leads to promising results for different-gender pairs since speech of male and female is intrinsically different in terms of pitch range, timbre, etc. However, for same-gender pairs, GPD-PITCH's ordering criterion of training labels is suboptimal. The order of average pitch of two same-gender speakers are usually different across utterances. Moreover, some other important characteristics of speech, including timbre, unvoiced speech, etc., can not be reflected by average pitch.

Utterance-level permutation invariant training (uPIT) has been proposed to replace rule-based label permutation. In uPIT, the two training labels are provided as a whole set instead of an ordered list, and the output-label pairing $i \leftrightarrow \theta^i$, for a given utterance, is defined as the pairing that minimizes the utterance-level training loss over all possible speaker permutations P. Taking the cross-entropy loss as an example, the optimal permutation is presented as:

$$\theta_* = \underset{\theta\in P}{\operatorname{argmin}} -\sum_m \sum_{i=1}^{2}\sum_{s=1}^{68} p(x_m^{\theta^i}(s)|\mathbf{y}_m)\log(O_m^i(s)) \tag{3}$$

θ_* is used for all frames within the current training utterance. The frame-level loss can then be represented by:

$$J_m^{uPIT} = -\sum_{i=1}^{2}\sum_{s=1}^{68} p(x_m^{\theta_*^i}(s)|\mathbf{y}_m)\log(O_m^i(s)) \tag{4}$$

In this study, we train an SI-BLSTM-RNN with uPIT to predict the pitch states of two speakers, denoted by uPIT-PITCH. As the training goes, we expect uPIT-PITCH to learn the correct output permutation for both different-gender and same-gender pairs.

3.2. uPIT based Speech Separation followed by Single Pitch Tracking

uPIT is originally proposed for two-talker speech separation (SS) in single-microphone recordings. Therefore, an alternative way to apply uPIT for multi-pitch tracking is to first perform uPIT based SS, and then track the separated signals of the two speakers using conventional single pitch tracking algorithms.

In this study, we follow the uPIT based speech separation framework in [12], and train a BLSTM-RNN to predict the spectra of two speakers. Magnitude STFT of the mixture is used as the input feature. Two time-frequency masks are then predicted and multiplied with the mixture STFT, to reconstruct the phase sensitive approximation (PSA) [12] of the two speakers' spectra. The uPIT training criterion are used for the two spectrum outputs. During inference, the estimated outputs are coupled with the noisy phase of the mixture to resynthesize two time-domain signals. In the end, the RAPT [17] algorithm is applied to the resulting signals for pitch tracking. This approach is denoted by uPIT-SS-RAPT.

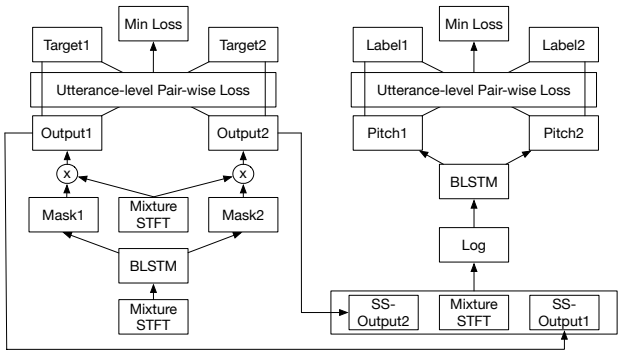

Fig. 1: Diagram of uPIT-SS-PITCH.

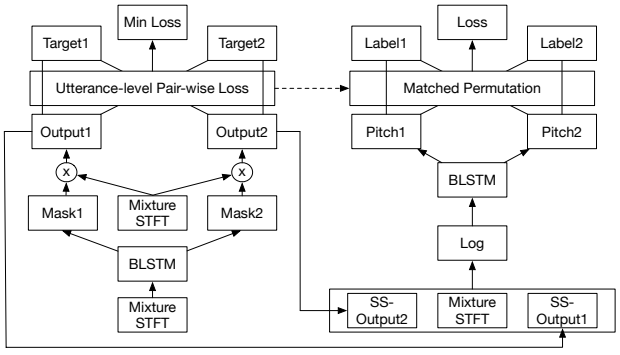

Fig. 2: Diagram of uPIT-SS-PERM-PITCH.

3.3. uPIT based Speech Separation followed by uPIT based Multi-pitch Tracking

uPIT-SS-RAPT gives excellent performance if the speech separation module works properly. However, for some challenging same-gender mixtures, where uPIT-SS struggles with, directly applying RAPT on top of uPIT-SS may be error-prone. There two main reasons. First, speaker assignment errors in uPIT-SS directly affect the pitch assignment in RAPT. Second, the masked signals may contain some artifacts which degrades the performance of RAPT.

To overcome these issues, we directly concatenate the two outputs form uPIT based speech separation with the magnitude STFT of the mixture to form a new input for uPIT-PITCH. A diagram of the network is shown in Fig. 1. The left module corresponds to uPIT-SS, which is trained first as the basis. The right module corresponds to uPIT-PITCH. We do not feed the two outputs of uPIT-SS to separate networks for single-pitch tracking, because that may compound the assignment errors generated by uPIT-SS. Input features in the right module are pre-processed using logarithm compression before feeding into the BLSTM-RNN. Lastly, it should be noted that the label permutation of the two modules are optimized independently. We expect uPIT-PITCH takes useful information from uPIT-SS to help pitch tracking. We denote the network by uPIT-SS-PITCH in this study.

3.4. uPIT based Speech Separation followed by Multi-pitch Tracking with Matched Permutation

One observation in uPIT-PITCH inference is that for some same-gender speaker pairs, the speaker assignment of pitch swaps very often across time. One possible reason is that the pitch label itself is not informative enough to correctly assign same-gender speakers to different labels. On the other hand, much richer information (unvoiced speech, timbre) is contained in SS's training targets, which may lead to better optimized label permutation during training.

To take advantage uPIT-SS's label permutation, we modify uPIT-SS-PITCH, and use the label permutation in the SS module for that in the pitch module, which is denoted by uPIT-SS-PERM-PITCH. A diagram of the system is shown in Fig. 2. There is a dash line connecting the label permutations of the two modules, meaning that the permutations are matched. By using this structure, better utterance-level label permutation for pitch tracking might be achieved.

4. FHMM INFERENCE

After BLSTM based pitch probability estimation, we use a factorial HMM to infer the most likely pitch tracks. The hidden variables are the pitch states of two speakers (x_m^1, x_m^2), and the observation variable is the feature vector $\mathbf{y}_m$. Prior probabilities and transition matrices of the hidden variables are computed from single-speaker training data either speaker-dependently for SPD-PITCH, or speaker-independently for all other models. Laplace smoothing is applied during the computation. We then compute the emission probability of the FHMM from the estimated posterior probability, and apply the junction tree algorithm to infer the most likely sequence of pitch states. In the end, frame-level pitch states are converted back to the centers of frequency bins, and smoothing is applied to get a continuous pitch track.

5. EXPERIMENTAL RESULTS AND COMPARISONS

5.1. Experimental Setup

We conduct experiments on the GRID database [4], which consists of 1,000 sentences spoken by each of 34 speakers. Two male and two female speakers ((No. 1, 2, 18, 20) are selected for testing, denoted by Set One. For each speaker in Set One, we randomly select 10 utterances, and mix them with every other speaker in Set One at -9, -6, -3, 0, 3, 6, and 9 dB. In total, $10\times10\times7$ test mixtures are generated for each of the 6 speaker pairs. We report results w.r.t. different gender combinations, and absolute energy differences between two test speakers. SPD-PITCH is trained within Set One, where 60,000 training mixtures are generated for each speaker pair by randomly mixing 900 training utterances of both speakers at a random energy ratio between -5 and 5 dB. Set Two is used to train all speaker-independent models, where another 10 male and 10 female speakers (No. 3, 5, 6, 9, 10, 12, 13, 14, 17, 19; 4, 7, 11, 15, 16, 21, 22, 23, 24, 27) with 900 training utterances each, are selected. For GPD-PITCH, the 60,000-mixture male-female training set is generated by randomly mixing male utterances with female utterances in Set Two at between -5 and 5 dB. We then divide same-gender speakers in Set Two into two groups based on their average pitch. For each same-gender pair, 60,000 mixtures are produced by mixing utterances from different groups at between -5 and 5 dB. Lastly, an SI training set used for all uPIT based models are generated by randomly mixing different-speaker utterances within Set Two at between -5 and 5 dB.

Reference pitch is extracted from single-speaker utterances using the RAPT algorithm [17]. All mixtures are sampled at 16 kHz. We extract STFT features using a frame length of 32ms, a frame shift of 10 ms, and the square root of hanning window.

Results are reported using the error measure E_{Total} proposed in [18], which jointly evaluates the performance in terms of pitch accuracy and speaker assignment. E_{Total} combines the percentile

Table 1: E_{Total} (%) of different multi-pitch tracking approaches w.r.t. different gender combinations, and absolute energy differences.

Methods	Same-Gender				Different-Gender			
	0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
Wohlmayr et al. SD	31.0	31.5	32.6	34.1	26.0	26.6	27.1	28.3
SPD-PITCH	12.4	12.4	12.8	13.9	11.5	11.6	11.8	12.5
GPD-PITCH	25.7	26.0	27.3	29.6	**14.3**	**14.6**	**15.2**	**16.3**
uPIT-PITCH	**25.1**	**24.9**	**24.4**	**25.2**	14.8	14.8	15.4	16.7

Table 2: E_{Total} (%) of joint speech separation - multi-pitch tracking w.r.t. different gender combinations, and absolute energy differences.

Methods	Same-Gender				Different-Gender			
	0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
uPIT-SS-RAPT	25.4	25.6	26.0	28.8	12.4	12.7	13.4	15.4
uPIT-SS-PITCH	24.6	24.6	24.2	26.2	14.5	14.6	15.0	16.3
uPIT-SS-PERM-PITCH	**23.8**	**23.4**	**23.5**	**25.3**	14.3	14.5	15.1	16.5

representation of voicing decision errors, permutation errors, gross errors and fine errors. The lower, the better. The details of E_{Total} can be found in [14, 18].

5.2. Models

All pitch estimation BLSTM-RNNs in this study share the same structure. There are three 500-unit BLSTM layers in the model. Two output layers with the softmax activation function are then used to predict pitch states. The networks are trained with the Adam optimization algorithm [11] and dropout regularization [7]. The initial learning rate is set to 0.001, and we decrease the learning rate by a ratio of 0.8 when the cross-validation loss stops decreasing for over 4 epochs. The maximum number of epochs is 30 for SPD-PITCH, and 100 for all other models.

The uPIT based speech separation network contains 3 BLSTM layers, each with 896 units. Two 257-unit ReLU [6] output layers are used to predict phase sensitive masks. The initial learning rate is set to 0.0002. All other training recipes follows the pitch BLSTM.

For reference, we compare all our methods with Wohlmayr et al.'s speaker-dependent GMM-FHMM model with gain adaptation [16, 18], which represents the state-of-the-art for SD multi-pitch tracking. The SD models in [16, 18] are trained within Set One, with the same RAPT based reference pitch. We would like to thank M. Wohlmayr, M. Stark, and F. Pernkopf for providing their pitch tracking code to us.

5.3. Results and Comparisons

Results of multi-pitch trackers without speech separation modules are reported in Table 1. All proposed SPD/GPD/SI systems significantly outperform Wohlmayr et al.'s SD models, which reflects the excellent modeling capacity of neural networks. Due to the usage of speaker-dependent information, SPD-PITCH achieves the best results among all systems. GPD-PITCH yields slightly worse E_{Total} than SPD-PITCH on different-gender pairs, and far worse E_{Total} on same-gender pairs. This result is expected since same-gender pairs are a lot more challenging for speaker-independent approaches, and the ad-hoc label assignment in GPD-PITCH definitely exacerbates this problem. uPIT-PITCH matches GPD-PITCH's performance on different-gender pairs, and outperforms GPD-PITCH on same-gender pairs by a small margin, which shows that the label permutation optimized by uPIT leads to better generalization for neural networks. However, the improvement is still relative small, thus we further introduce speech separation to help multi-pitch tracking.

Table 2 reports results of all joint SS-PITCH systems. uPIT-SS-RAPT achieves exceptionally good results on different-gender pairs, primarily due to the fact that the same single pitch tracking algorithm, RAPT, is shared between uPIT-SS-RAPT and the reference pitch. To be more specific, when the separation module works well, uPIT-SS-RAPT tends to generate exactly the same pitch as the reference pitch, which also includes consistent pitch errors, and voicing decision errors in the reference pitch. These random errors by RAPT are regularized by BLSTM-RNNs during training, and thus are regarded as incorrect estimations for BLSTM based models. However, for uPIT-SS-RAPT, since the errors are consistent with the reference pitch, it would be recognized as correct estimation. On the other hand, uPIT-SS-RAPT works relatively poorly on challenging same-gender pairs, which implies that inaccurate estimation in speech separation introduces errors for pitch tracking. With the help of the additional input feature, uPIT-SS-PITCH consistently outperforms uPIT-PITCH. There is only one exception, which is for same-gender pairs at the level-difference of 9 dB. The reason is that severe mismatch happens in this condition, so that the additional input may be too noisy and to provide any useful information. Lastly, with matched label permutation, uPIT-SS-PERM-PITCH generates the best results among speaker-independent models for same-gender pairs. For different-gender pairs, uPIT-SS-PERM-PITCH also matches the male-female GPD-PITCH, which is specifically trained for male-female pairs, and has optimally assigned label permutations.

6. CONCLUSION

In this study, we have introduced utterance-level permutation invariant training for multi-pitch tracking. BLSTM-RNNs with two probabilistic pitch outputs are used as the base model. SPD, GPD and uPIT-SI training are compared. For uPIT based pitch estimation models, several extensions have been proposed, including incorporating outputs and label permutations from uPIT based speech separation. Experimental results show that our final model, uPIT-SS-PERM-PITCH, achieves the best results among all GPD and SI models, especially for same-gender speaker pairs. In the future, we will explore multi-target training and joint optimization for speech separation and multi-pitch tracking

7. REFERENCES

[1] F. Bach and M. Jordan, "Discriminative training of hidden Markov models for multiple pitch tracking," in *Proc. ICASSP*, 2005, pp. 489–492.

[2] A. D. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1917–1930, 2002.

[3] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, B. H. Juang, Ed. San Rafael, CA, USA: Morgan & Claypool, 2009.

[4] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, pp. 2421–2424, 2006.

[5] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 138–150, 2014.

[6] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, 2011, pp. 315–323.

[7] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[9] K. Hu and D. L. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 122–131, 2013.

[10] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1091–1102, 2011.

[11] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[12] M. Kolbæk, 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, Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[13] Y. Liu and D. L. Wang, "Robust pitch tracking in noisy speech using speaker-dependent deep neural networks," in *Proc. ICASSP*, 2016, pp. 5255–5259.

[14] ——, "Speaker-dependent multipitch tracking using deep neural networks," *J. Acoust. Soc. Amer.*, vol. 141, pp. 710–721, 2017.

[15] ——, "Time and frequency domain long short-term memory for noise robust pitch tracking," in *Proc. ICASSP*, 2017, pp. 5600–5604.

[16] R. Peharz, M. Wohlmayr, and F. Pernkopf, "Gain-robust multipitch tracking using sparse nonnegative matrix factorization," in *Proc. ICASSP*, 2011, pp. 5416–5419.

[17] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding Synth.*, pp. 495–518, 1995.

[18] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 799–810, 2011.

[19] M. Wu, D. L. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 229–241, 2003.