Technical Report OSU-CISRC-8/15-TR12
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2015**
File: **TR12.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Speaker-dependent multipitch tracking using deep neural networks

Yuzhou Liu
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
liuyuz@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive and Brain Sciences
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Abstract – Multipitch tracking is important for speech and signal processing. However, it is challenging to design an algorithm that achieves accurate pitch estimation and correct speaker assignment at the same time. In this paper, we use deep neural networks (DNNs) to model the probabilistic pitch states of two simultaneous speakers. To capture speaker-dependent information, we propose two types of DNN with different training strategies. The first is trained for each speaker enrolled in the system (speaker-dependent DNN), and the second is trained for each speaker pair (speaker-pair-dependent DNN). Several extensions, including gender-pair-dependent DNNs, speaker adaptation of gender-pair-dependent DNNs and multi-ratio training, are introduced later to relax constraints. A factorial hidden Markov model (FHMM) then integrates pitch probabilities and generates the most likely pitch tracks with a junction tree algorithm. Experiments show that the proposed methods substantially outperform other speaker-independent and speaker-dependent multipitch trackers on two-speaker mixtures. With multi-ratio training, our methods achieve consistent performance at various energies ratios of the two speakers in a mixture.

Index Terms – Multipitch tracking, deep neural networks, speaker-dependent modeling, speaker adaptation, multi-condition training, factorial hidden Markov model.

1 Introduction

There is a long-standing interest in estimating the pitch, or the fundamental frequency (F0) of speech. A reliable estimation of pitch is critical for many speech processing applications, including automatic speech recognition [5], speaker identification [33] and speech separation [27]. Over the last few decades, various algorithms have been designed for tracking the pitch of a single speaker [4] [6] [25], and they achieved good performance under clean or modestly noisy conditions. However, pitch tracking when speech is severely corrupted by interfering speakers is still a challenging problem.

This paper is concerned with multipitch tracking when two speakers are talking simultaneously. A number of studies have investigated this problem. Wu *et al.* [31] proposed a probabilistic representation of pitch and tracked continuous pitch contours with a hidden Markov model (HMM). Sha and Saul [24] modeled the instantaneous frequency spectrogram with nonnegative matrix factorization and used the inferred weight coefficients to determine pitch candidates. Bach and Jordan [2] proposed direct probabilistic modeling of the spectrogram and tracked several pitches with a factorial HMM (FHMM). Hu and Wang [16] proposed a tandem algorithm that performs pitch estimation and voiced speech segregation jointly, producing a set of pitch contours and their associated binary masks. Jin and Wang [18] improved Wu *et al.*'s system by designing new techniques for channel selection and pitch score estimation in the context of reverberant and noisy signals. The abovementioned studies built a general system without modeling the characteristics of any specific speaker, and can thus be denoted as speaker-independent models. Although most speaker-independent models perform well for estimating pitch periods, they can not assign pitch estimates to the underlying speakers for multipitch tracking. To alleviate this problem, Hu and Wang [17] built their system on the tandem algorithm [16] and grouped simultaneous pitch contours into two speakers using a constrained clustering algorithm. Similarly, Duan *et al.* [9] took the pitch estimates of speaker-independent multipitch trackers as input and streamed pitch points by clustering. However, both approaches achieved limited improvement as individual pitch contours and points are usually too short to contain enough speaker information for clustering. On the other hand, speaker-dependent models have been investigated recently. Wohlmayr, Stark and Pernkopf [30] modeled the probability of pitch periods using speaker-dependent Gaussian mixture models (GMMs), and then used a speaker-dependent FHMM to track pitches of two simultaneous speakers. They have shown significant improvement over a speaker-independent approach [31].

In this paper, we propose a speaker-dependent and discriminative technique to model the pitch probability at each time frame. Specifically, we use deep neural networks (DNNs) to model the posterior probability that a pair of frequency bins (pitch states) is pitched given frame-level observations. A DNN is a feedforward neural network that contains more than one hidden layer [15]. Recently, Han and Wang [13] used DNNs to model the posterior probability

of pitch states for single-pitch tracking in noisy conditions, which motivates the use of DNNs for multipitch tracking in this study. To leverage individual speaker characteristics, we train a DNN for each speaker enrolled in the system, denoted as speaker-dependent DNNs or SD-DNNs. We also train DNNs for different pairs of speakers, denoted as speaker-pair-dependent DNNs or SPD-DNNs. We then extend the DNN based models to relax practical constraints. To deal with unseen speakers, we train three gender-pair-dependent DNNs (male-male, male-female and female-female, denoted as GPD-DNNs) as a generalization of SPD-DNNs. GPD-DNNs only require gender information during testing. With insufficient training data, direct training of SD-DNNs or SPD-DNNs may result in overfitting. To examine this issue, we conduct a fast adaptation of GPD-DNNs for each speaker pair with limited training data. Also, the utterances of the two speakers in a mixture usually have different energy ratios, leading to a ratio mismatch between training and test. We address this problem by including various speaker energy ratios in training, denoted as the multi-ratio training.

After estimating the posterior probability of pitch states, we use an FHMM for pitch tracking. Under the framework of the FHMM, the pitch state of each speaker evolves within its own Markov chain, while the emission probability is derived by the posterior probability estimated by DNNs. We then use the junction tree algorithm [19] to track the most likely pitch tracks.

The rest of the paper is organized as follows. The next section gives an overview of the system architecture. Feature extraction is discussed in Section 3. The details of DNN based posterior probability estimation are introduced in Section 4. Section 5 describes the FHMM for multipitch tracking. Experimental results and comparisons are presented in Section 6. Finally, we conclude the paper and discuss related issues in Section 7. A preliminary version of this paper is presented at Interspeech 2015 [21].

2 System Overview

A diagram of our proposed multipitch tracker is illustrated in Fig. 1. The input to the system is a speech mixture v_t sampled at 1.6 KHz:

$$v_t = u_t^1 + u_t^2 \tag{1}$$

where u_t^1 and u_t^2 are utterances of two speakers. Given the mixture, our system first extracts frame-level features $\mathbf{y}_m$, which corresponds to the first module in the diagram.

In the second stage, features are fed into DNNs to derive the posterior probability of pitches at frame m, i. e., $p(x_m^1, x_m^2|\mathbf{y}_m)$, where x_m^1 and x_m^2 denote pitch states of two speakers at frame m. Both x_m^1 and x_m^2 have 68 states $(s^1, s^2, s^3, \ldots, s^{68})$, where s^1 refers to an unvoiced or silent state, and s^2 to s^{68} encode different pitch frequencies ranging from 60 to 404 Hz [13]. Specifically, we quantize the pitch frequency range 60 to 404 Hz using 24 bins per octave

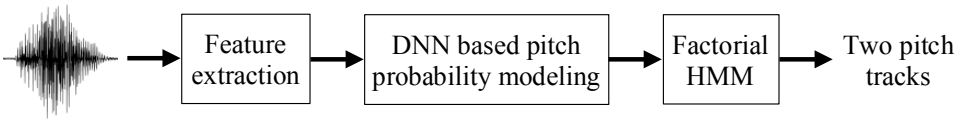

Figure 1: Diagram of the proposed multipitch tracker.

on a logarithmic scale, resulting in a total of 67 bins. $p(x_m^1 = s^i, x_m^2 = s^j|\mathbf{y}_m)$ equals one if groundtruth pitches fall into the i^{th} and j^{th} frequency bins respectively. We propose two types of DNN to estimate the posterior probability, which are the speaker-dependent DNNs and the speaker-pair-dependent DNNs. We also explore several extensions. The detailed settings of DNNs can be found in Section 4.

The final module converts the posterior probability $p(x_m^1, x_m^2|\mathbf{y}_m)$ to the emission probability of an FHMM $p(\mathbf{y}_m|x_m^1, x_m^2)$. The junction tree algorithm is then applied to infer the most likely pitch tracks. Note that in the following sections, a pitch contour refers to a continuous pitch trajectory from the same speaker, and a pitch track refers to a set of pitch contours from the same speaker.

3 Feature Extraction

Features should encode the information of pitch and speaker identity at the same time. We investigate three features: cochleagram, log spectrogram and mel-frequency cepstral coefficients in our study.

3.1 Cochleagram

To get the cochleagram feature, we first decompose the input signal into the time-frequency domain by using a bank of 64 gammatone filters whose center frequencies range from 50 Hz to 8000 Hz. Gammatone filters model the impulse responses of auditory filters and are widely used [27]. We divide each subband signal into 20 ms frames with a 10 ms frame shift. The cochleagram is derived by computing the energy of each subband signal at each frame. We then loudness compress the cochleagram with a cubic root operation to get the final cochleagram feature.

3.2 Log Spectrogram

To get the spectrogram feature, the signal is first divided into 32 ms frames with a 10 ms frame shift. We then apply a Hamming window to each frame and derive the spectrogram using 1024-point FFT. Lastly, we compute the logarithm of the amplitude spectrum, and pick bins 2-65 (corresponding to a frequency range up to 1000 Hz) as our frame-level feature vector. This feature is adopted by Wohlmayr *et al.* in their GMM-FHMM based approach [30].

3.3 Mel-frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCC) are widely used in automatic speech recognition and speaker recognition. To compute MFCC, we divide the input signal into 20 ms frames with a 10 ms frame shift. The power spectrogram is derived using short-time Fourier transform filtered by a Hamming window. Next we use a bank of 64 mel scale filters to convert the power spectrogram into mel scale. Lastly, logarithm compression and discrete cosine transform are applied to compute 31-dimensional MFCC.

To make use of the temporal context, we concatenate neighboring frames into one feature vector. Denoting the feature vector extracted within frame m as $\hat{\mathbf{y}}_m$, we have:

$$\mathbf{y}_m = [\hat{\mathbf{y}}_{m-d}, \ldots, \hat{\mathbf{y}}_m, \ldots, \hat{\mathbf{y}}_{m+d}] \tag{2}$$

where d is chosen to be 5 using cross validation.

4 DNN based Pitch Probability Modeling

DNNs have been successfully applied in various speech processing applications. In this section, we first introduce two types of DNN for posterior probability estimation. Next we extend the models to relax practical constraints.

4.1 Speaker-dependent DNNs

The goal of DNNs is to model the posterior probability that a pair of pitch states occurs at frame m, i. e., $p(x_m^1, x_m^2|\mathbf{y}_m)$. However, this would be difficult without the prior knowledge of the underlying speakers. We first focus on training speaker-dependent DNNs to model the posterior probability.

According to the chain rule in probability theory:

$$p(x_m^1, x_m^2|\mathbf{y}_m) = p(x_m^1|\mathbf{y}_m)p(x_m^2|x_m^1, \mathbf{y}_m) \tag{3}$$

we can estimate $p(x_m^1|\mathbf{y}_m)$ and $p(x_m^2|x_m^1, \mathbf{y}_m)$ in turn to get $p(x_m^1, x_m^2|\mathbf{y}_m)$. In this study, we estimate the pitch-state probability of speaker one $p(x_m^1|\mathbf{y}_m)$ by training a DNN. The input layer of the DNN corresponds to the frame-level feature vector of the mixture. There are four hidden layers in the DNN, and each hidden layer has 1024 hidden units with the ReLU activation function [11]. The output layer has 68 softmax output units, denoted as $(O_1^1, O_1^2, \ldots, O_1^{68})$, where O_1^j estimates $p(x_m^1 = s^j|\mathbf{y}_m)$. Hence there are 67 '0's and a '1' in the desired output. The value '1' corresponds to the frequency bin of the groundtruth pitch. We use cross-entropy as the cost function. The standard backpropagation algorithm and dropout regularization (dropout rate 0.2) are used to train the network, with no pretraining [14]. We adopt mini-batch stochastic gradient descent along with a momentum term (0.9) for the

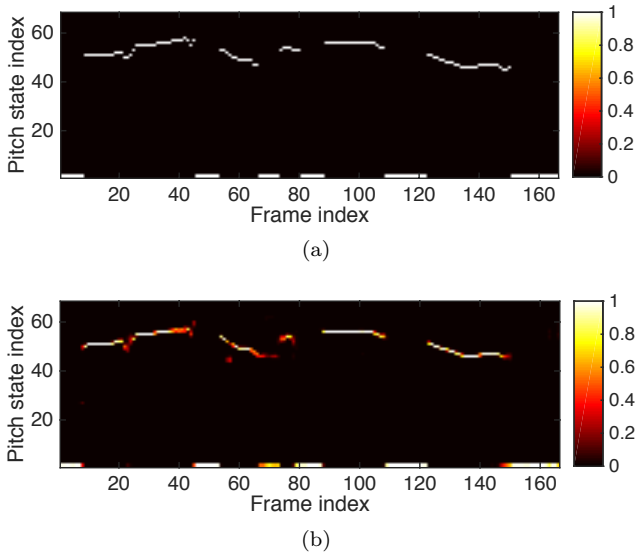

Figure 2: Pitch probability modeling of the first speaker in a female-female mixture at 0 dB. (a) Groundtruth probabilities of pitch states. (b) Probabilities of pitch states estimated by a DNN.

optimization. The choice of DNN parameters is justified in Section 6.2. The training data contain mixtures of speaker one and a set of interfering speakers.

Fig. 2 compares the groundtruth and estimated pitch-state probabilities of speaker one in a female-female test mixture. As shown in the figure, the DNN rather accurately models the conditional probability of x_m^1, even without knowing x_m^2. Therefore we further assume the conditional independence between x_m^1 and x_m^2:

$$p(x_m^2|x_m^1, \mathbf{y}_m) = p(x_m^2|\mathbf{y}_m) \tag{4}$$

In the next step, we train another DNN to model $p(x_m^2|\mathbf{y}_m)$ using exactly the same structure and training methodology as for the first DNN. After estimating $p(x_m^2|\mathbf{y}_m)$, the original posterior probability can be obtained by:

$$p(x_m^1, x_m^2|\mathbf{y}_m) = p(x_m^1|\mathbf{y}_m)p(x_m^2|\mathbf{y}_m) \tag{5}$$

Because we train a DNN for each enrolled speaker, we denote this model as the speaker-dependent DNN (SD-DNN).

4.2 Speaker-pair-dependent DNNs

A speaker-pair-dependent DNN (SPD-DNN) is a DNN trained on a specific pair of speakers. The structure of a SPD-DNN is quite similar to that of a SD-DNN. The input layer corresponds to the frame-level feature vector. There are four hidden layers with 1024 ReLU units. Instead of estimating the probability for only one speaker, we concatenate the pitch-state probabilities of the other speaker into the DNN output. The resulting output layer has 136

units, denoted as $(O_1^1, \ldots, O_1^{68}, O_2^1, \ldots, O_2^{68})$, where O_i^j estimates $p(x_m^i = s^j|\mathbf{y}_m)$. To correctly model the probability distribution, the activation function for the output layer is a softmax function. Assuming that output units before applying the activation function have values $(v_1^1, \ldots, v_1^{68}, v_2^1, \ldots, v_2^{68})$, we have:

$$O_i^j = \frac{exp(v_i^j)}{\sum_{k=1}^{68} exp(v_i^k)}, \quad \text{for } i = 1 \text{ or } 2,\ 1 \leq j \leq 68 \tag{6}$$

Other training details exactly follow SD-DNNs. The posterior probability of pitch states is estimated by:

$$p(x_m^1 = s^i, x_m^2 = s^j|\mathbf{y}_m) = O_1^i O_2^j \tag{7}$$

Because SPD-DNNs are trained on speaker pairs, they should accurately capture the underlying speaker information. On the other hand, for a system with N speakers enrolled, we need to train N SD-DNNs, but $\frac{N(N-1)}{2}$ SPD-DNNs.

4.3 Extensions

SD-DNNs and SPD-DNNs utilize detailed speaker information to estimate the posterior probability of pitch states. In this section, we introduce extensions to relax their practical constraints.

4.3.1 Gender-pair-dependent DNN

SD-DNNs and SPD-DNNs are not applicable to unseen speakers. To deal with this constraint, we extend our speaker-dependent models to gender-dependent ones. In this way, only the genders of the two underlying speakers are needed during testing.

A straightforward way to design a gender-dependent model is to follow the structure of SD-DNNs and train two DNNs for male and female speakers, respectively. This idea works well for male-female mixtures, but can not distinguish the two speakers of the same gender. Therefore we build our gender-dependent model by extending SPD-DNNs to gender-pair-dependent DNNs or GPD-DNNs. We train three GPD-DNNs for different gender pairs: male-female, male-male and female-female. The structure of a GPD-DNN is chosen to be the same as a SPD-DNN for simplicity. For the male-female GPD-DNN, the pitch-state probabilities of the male speaker correspond to the first 68 output units, and the female speaker the remaining output units. For same-gender GPD-DNNs, the first 68 output units correspond to the speaker with lower average pitch, and the other output units correspond to the speaker with higher average pitch. Although this layout may lead to incorrect speaker assignment at some frames, it provides a reasonable way to distinguish two speakers with the little information available. Other training aspects exactly follow SPD-DNNs.

4.3.2 Adaptation of GPD-DNNs with Limited Training Data

SD-DNNs and SPD-DNNs would overfit if we could not collect enough training data. One way to address this problem is to perform speaker adaptation of GPD-DNNs with limited data. Speaker adaptation of DNNs has been studied in automatic speech recognition. Two typical approaches include incorporating speaker-dependent information into DNN's input [1] [23] and regularized retraining [20] [32]. In the first approach, speaker dependent information, like i-vectors and speaker codes, is incorporated into the input of DNNs and the original features are projected into a speaker-normalized space. In regularized retraining, the weights of DNNs are modified using the adaptation data. To ensure that the adapted model does not deviate too much from the original model, a regularization term is added to the cost function. Both approaches substantially improve the performance of unadapted DNNs.

We use regularized retraining to perform speaker adaptation. For each new speaker pair, we retrain all the weights of the corresponding GPD-DNN on limited adaptation data with a relatively small learning rate (0.001) and a weight decay (L_2 regularization) of 0.0001. Other training aspects follow those for training SPD-DNNs.

4.3.3 Multi-ratio Training

Utterances of the two speakers in a mixture usually have different energy ratios. A ratio mismatch between training and test may result in performance degradation for supervised algorithms. Under the framework of GMM-FHMM, Wohlmayr *et al.* alleviated this problem by adding a gain parameter to the mean vectors of each GMM [29]. An expectation-maximization based algorithm was then performed to estimate the gains for each test mixture. In addition, they included gain-robust NMF-based pitch detection [22]. However, it is unclear how to apply these techniques to DNNs.

Generally speaking, the performance of supervised learning is sensitive to the information contained in the training set. Therefore a simple and effective way for improving generalization is to enlarge the training set by including various acoustic conditions [28]. In this study, we perform multi-condition training by creating mixtures at different speaker energy ratios, denoted as multi-ratio training. The resulting DNNs are denoted as ratio-adapted DNNs. The details of multi-ratio training are given in Section 6.

5 Factorial HMM Inference

Once all posterior probabilities are estimated by DNNs, we use a factorial HMM to infer the most likely pitch tracks. A factorial HMM is a graphical model that contains several Markov chains [10]. In this study, we only discuss the case of two Markov chains, as shown in Fig. 3. The hidden variables (x_m^1, x_m^2) are the pitch states of two speakers, and the observation variable is the feature vector $\mathbf{y}_m$. The Markov assumption implies that $\mathbf{y}_m$ is independent of

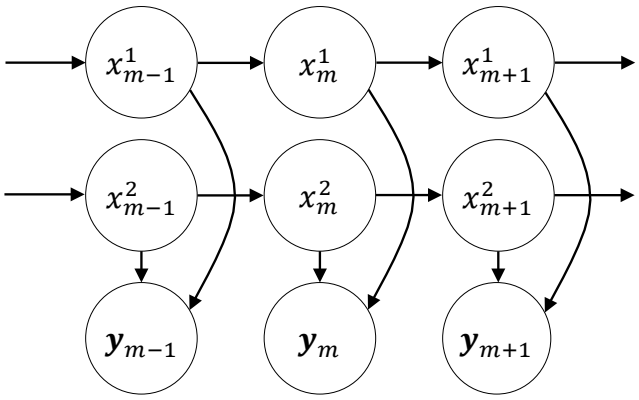

Figure 3: A Factorial HMM with two Markov chains.

all variables given x_m^1 and x_m^2. Assuming the total number of frames is M, we denote the sequence of variables in boldface: $\mathbf{X} = \bigcup_{m=1}^{M}\{x_m^1, x_m^2\}$, $\mathbf{Y} = \bigcup_{m=1}^{M}\{\mathbf{y}_m\}$. The overall joint probability of the model is given by:

$$p(\mathbf{X}, \mathbf{Y}) = p(x_1^1)p(x_1^2)p(\mathbf{y}_1|x_1^1, x_1^2) \prod_{m=2}^{M} p(x_m^1|x_{m-1}^1)p(x_m^2|x_{m-1}^2)p(\mathbf{y}_m|x_m^1, x_m^2) \tag{8}$$

Prior probabilities and transition matrices of the hidden variables are computed from single-speaker recordings in the training set either speaker-dependently (for SD-DNNs and SPD-DNNs) or gender-dependently (for GPD-DNNs). To avoid a probability of zero, Laplace smoothing is applied during the computation, where we add one to each possible observation. The emission probability can be computed using the estimated posterior probability and Bayes rule:

$$p(\mathbf{y}_m|x_m^1, x_m^2) = \frac{p(x_m^1, x_m^2|\mathbf{y}_m)p(\mathbf{y}_m)}{p(x_m^1)p(x_m^2)} \tag{9}$$

where $p(\mathbf{y}_m)$ is a constant for all feature vectors.

Once all probabilities are derived, we apply the junction tree algorithm to infer the most likely sequence of pitch states. The first step of this algorithm is to convert the directed graphical model to an undirected graphical model. In the next step, the nodes in the undirected graph are arranged to form a junction tree, where belief propagation is performed. For more details on the junction tree algorithm, we refer the interested reader to [19] [30]. The time complexity of the junction tree algorithm is $O(2 \times 68^3 \times M)$ in our study. We then convert derived pitch states to the mean frequencies of the corresponding frequency bins. In the end, we use a moving average window of length three to smooth frequencies and get final pitch estimates.

6 Evaluations and Comparisons

6.1 Corpus and Error Measurement

For evaluations, we first use the GRID database [7], which is also used in [30] hence facilitating our comparisons. The corpus consists of 1000 sentences spoken by each of 34 speakers (18 male, 16 female). Two male and two female speakers (No. 1, 2, 18, 20, same as [30]), denoted as MA1, MA2, FE1 and FE2, are selected to train and test the proposed methods, except for GPD-DNNs which are tested on the same four speakers but trained on another set of speakers. We denote these four speakers as Set One. For each speaker in Set One, 950 sentences are selected for training, 40 sentences are used for choosing the best DNN weights during training, and the remaining ten sentences are used for testing. Note that all test sentences used in [30] are also included in our test set. Another ten male and ten female speakers (No. 3, 5, 6, 9, 10, 12, 13, 14, 17, 19; 4, 7, 8, 11, 15, 16, 21, 22, 23, 24) are used in the training of SD-DNNs and GPD-DNNs, where again for each speaker we select 950 sentences for training, and 40 sentences for selecting the best DNN weights. We denote these twenty speakers as Set Two. Groundtruth pitches are extracted from single-speaker sentences using RAPT [25], which outperforms other pitch trackers on clean speech signals [8].

To mix two sentences u_t^1 and u_t^2, we first select a speaker ratio R in dB, and amplify one of the speakers by R dB. A mixture with a speaker ratio of R dB is created by combining the resulting sentences using: $v_t = 10^{R/20}u_t^1 + u_t^2$ or $v_t = u_t^1 + 10^{R/20}u_t^2$. Note that if we choose a speaker ratio of 0 dB, the two equations to derive v_t are the same. For comparison reasons, we use a matched speaker ratio of 0 dB in the training and test of SD-DNNs, SPD-DNNs, GPD-DNNs and adaptation of GPD-DNNs. Unmatched speaker ratios are used to test multi-ratio training. The details of the training and test set are as follows:

- SD-DNNs: training mixtures are created by mixing each sentence of the target speaker in Set One with 60 random sentences in Set Two at 0 dB. Thus there are 57000 training mixtures created for every target speaker. The test is conducted within Set One. We exhaustively mix test sentences for each speaker pair in Set One at 0 dB, resulting in a total of $10\times10\times6 = 600$ test mixtures.

- SPD-DNNs: for each speaker pair in Set One, we build the training set by mixing sentences of the two speakers at 0 dB. We make sure that each sentence of one speaker is randomly mixed with 60 sentences of the other speaker. Therefore 57000 mixtures are created to train each speaker pair. We use the same test set as for SD-DNNs.

- GPD-DNNs: the training is conducted within Set Two. We randomly create 57000 mixtures for each gender pair at 0 dB. The same test set is used as for SD-DNNs.

- Adaptation of GPD-DNNs: For each speaker pair in Set One, we randomly select 100 mixtures from the SPD-DNN's training set as the adaptation data. The same test set is

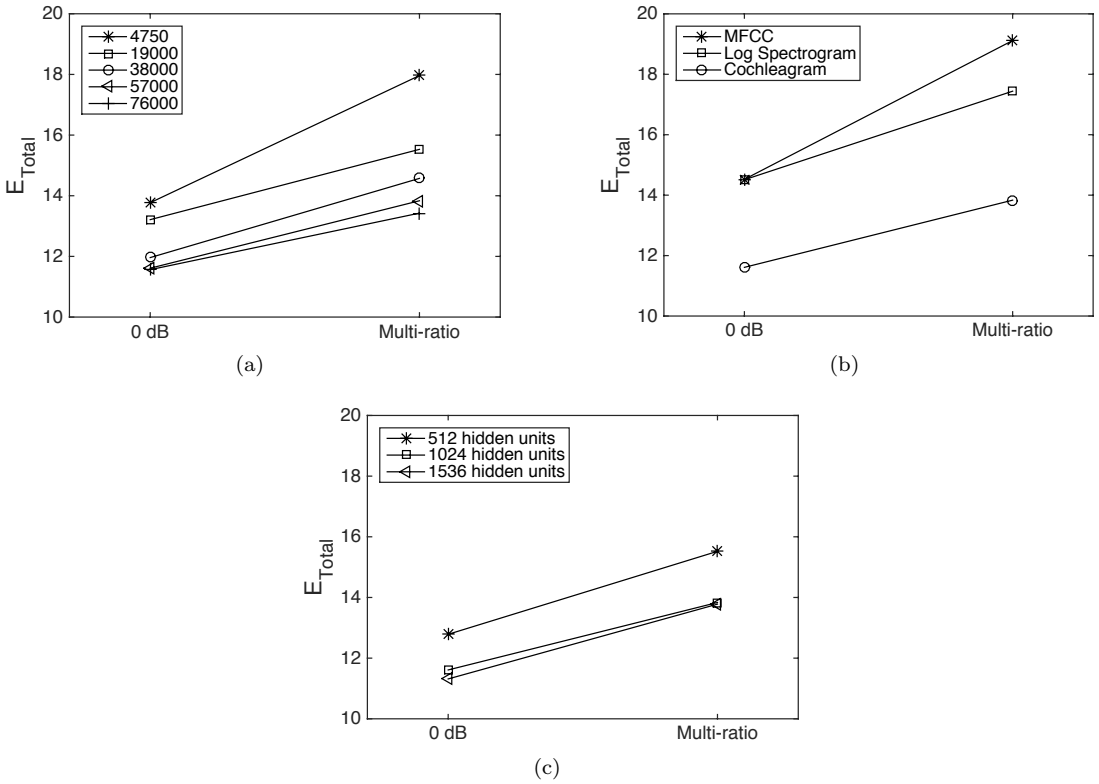

Figure 4: Average E_{total} of SPD-DNNs with different (a) sizes of training set, (b) features, (c) numbers of hidden units.

used as for SD-DNNs.

- Multi-ratio training: the training is conducted for both SD-DNNs and SPD-DNNs. Mixtures are no longer created at 0 dB in this experiment. Instead, we randomly amplify one of the two speakers with a random ratio of $R = \{-12, 6, 0, 6, 12\}$ dB for each training mixture. As for the test set, we alternately amplified one of the two sentences with a ratio out of $R = \{-15, -12, -9, -6, -3, 0, 3, 6, 9, 12, 15\}$ dB, which gives $10{\times}10{\times}6{\times}2 = 1200$ mixtures at each speaker energy ratio, and 13200 mixtures in total; note that each mixture at 0 dB appears twice in test.

In addition, we test our proposed methods using the FDA database [3] where the groundtruth pitches are derived from laryngograph data.

We evaluate pitch tracking results using the error measure proposed in [30], which jointly evaluates the performance in terms of pitch accuracy and speaker assignment. Assuming that the ground truth pitch tracks are F_m^1 and F_m^2, we globally assign each estimated pitch track to a groundtruth pitch track based on the minimum mean square error and denote the assigned estimated pitch tracks as f_m^1 and f_m^2. The pitch frequency deviation of speaker i, $i{\in}\{1, 2\}$, is:

$$\Delta f_m^i = \frac{|f_m^i - F_m^i|}{F_m^i} \tag{10}$$

Table 1: E_{Total} for different multipitch trackers tested on 600 test mixtures of the GRID corpus

		E_{01}	E_{02}	E_{10}	E_{12}	E_{20}	E_{21}	E_{Gross}	E_{Fine}	E_{Perm}	E_{Total}
Jin and Wang	Mean	4.54	1.25	6.97	5.51	1.94	**12.81**	4.80	6.93	6.47	**51.21**
	Std	2.34	1.38	3.55	3.33	2.16	5.54	4.65	3.17	5.34	11.71
Hu and Wang	Mean	3.88	0.23	6.77	4.71	1.76	**14.66**	2.17	6.00	2.49	**42.67**
	Std	2.83	0.66	3.73	3.73	2.47	6.33	3.71	2.06	4.29	10.92
Wohlmayr *et al.* SD	Mean	1.81	0.06	5.89	2.68	1.39	**10.81**	0.93	2.79	0.37	**26.73**
	Std	1.64	0.26	3.42	2.18	2.06	5.26	1.14	0.73	0.79	9.49
SD-DNN	Mean	2.03	0.13	1.96	5.65	0.07	**2.77**	0.72	2.31	1.00	**16.65**
	Std	1.66	0.43	1.94	5.40	0.28	2.04	1.25	0.83	2.23	7.86
SPD-DNN	Mean	1.72	0.08	1.53	3.25	0.05	**2.55**	0.52	1.96	0.15	**11.82**
	Std	1.42	0.29	1.54	2.14	0.24	1.91	0.91	0.35	0.54	3.34

The voicing decision error E_{ij}, $i \neq j$, denotes the percentage of time frames where i pitch points are wrongly detected as j pitch points. For each speaker i, the permutation error E^i_{Perm} is set to one at time frames where the voicing decision for both estimates is correct, but Δf^i_m exceeds 20%, and f^i_m is within the 20% error bound of the other reference pitch, i. e., the error is due to incorrect speaker assignment. The overall permutation error E_{Perm} is the percentage of time frames where either E^1_{Perm} or E^2_{Perm} is one. Next, for each speaker i, the gross error E^i_{Gross} is set to one at time frames where the voicing decision for both estimates is correct, but Δf^i_m exceeds 20% with no permutation error. The overall gross error E_{Gross} is the percentage of time frames where either E^1_{Gross} or E^2_{Gross} is one. The fine detection error E^i_{Fine} is defined as the average of Δf^i_m in percent at time frames where Δf^i_m is smaller than 20%. $E_{Fine} = E^1_{Fine} + E^2_{Fine}$. The total error is used as the overall performance measure:

$$E_{total} = E_{01} + E_{02} + E_{10} + E_{12} + E_{20} + E_{21} + E_{Perm} + E_{Gross} + E_{Fine} \tag{11}$$

6.2 Parameter Selection

Because all proposed DNNs have similar structure, we conduct parameter selection for SPD-DNNs only. The resulting parameters are used in other models. We use a new pair of male speakers (No. 26 and 28 in the GRID corpus) as the development set. For each speaker, 950 sentences are used for training, 40 sentences are used for choosing the best DNN weights during training and 10 sentences are used for test. Besides the matched 0 dB training and test condition, we also train the SPD-DNN with multi-ratio training. The details of the training and test set follow Section 6.1. The results of multi-ratio training are averaged across all speaker ratios.

The size of the training set has strong impact on DNN's performance. We create five training sets by randomly mixing each sentence of one speaker with 5, 20, 40, 60 and 80 sentences of the other speaker, resulting in 4750, 19000, 38000, 57000, 76000 mixtures. A SPD-DNN is trained for each training set. The results are given in Fig. 4(a). In general,

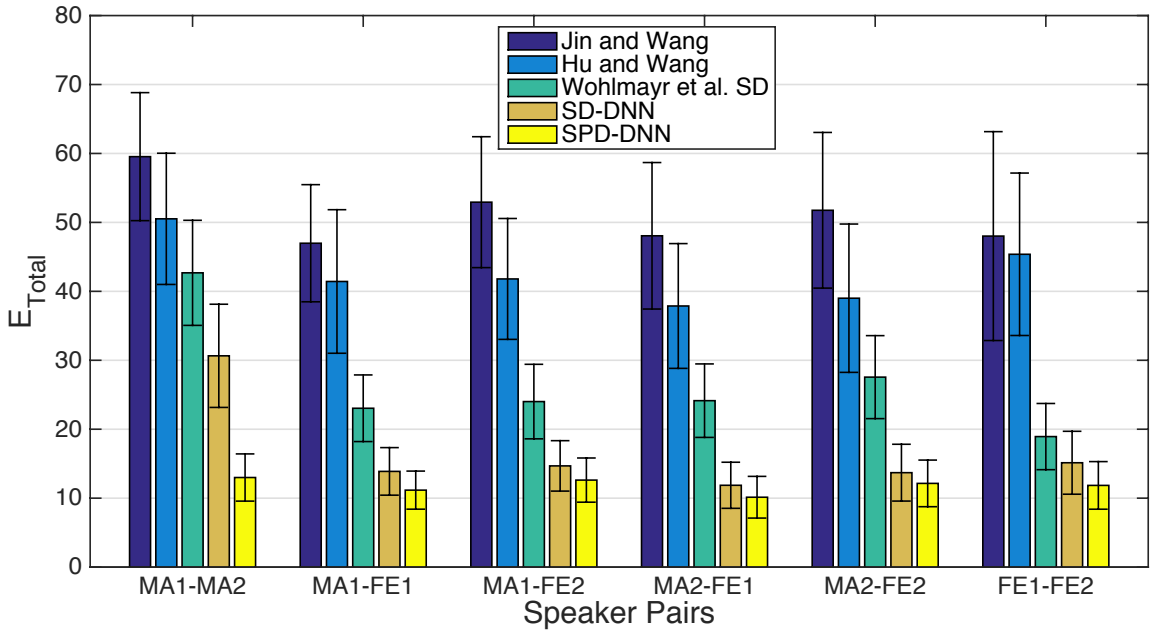

Figure 5: E_{total} of different approaches tested on six pairs of speakers. Error bars depict the mean and standard deviation of a method on the test mixtures of a given speaker pair.

the total error decreases with the increase of the training size, and the improvement becomes small when the training size reaches 57000. Taking the computational cost into consideration, we choose 57000 training mixtures in the following experiments.

Features are also important to the system. We compare three features: cochleagram, log spectrogram and MFCC in this study. As shown in Fig. 4(b), the cochleagram feature outperforms other two features under both experimental conditions.

Next, we investigate the number of hidden units used in SPD-DNNs. Three numbers are compared: 512, 1024 and 1536. As shown in Fig. 4(c), the total error is reduced by more than 1.1% when the number is increased from 512 to 1024. However, further increasing the number of hidden units does not significantly boost the performance. Other parameters, including the type of activation functions, the number of hidden layers, learning rate, mini-batch size and the number of neighboring frames, are also chosen from the same development set.

6.3 Results and Comparisons

We present our results, and compare with three state-of-the-art multipitch trackers: Jin and Wang's [18] (denoted as Jin and Wang), Hu and Wang's [17] (denoted as Hu and Wang) and Wohlmayr *et al.*'s [30] (denoted as Wohlmayr *et al.*). Jin and Wang's approach was designed for noisy and reverberant signals. They used correlogram to select reliable channels and tracked continuous pitch contours with an HMM. Hu and Wang built their system on top of the tandem algorithm [16]. They grouped pitch contours into two speakers using a constrained clustering algorithm. Both multipitch trackers are speaker-independent and unsupervised.

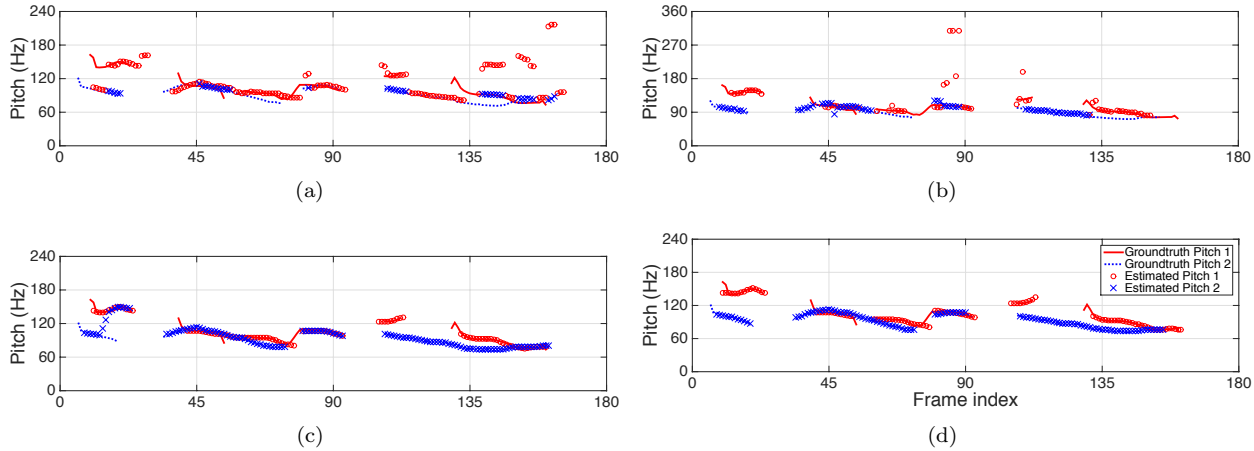

Figure 6: Multipitch tracking results on a test mixture (pbbv6n and priv3n) for the MA1-MA2 speaker pair. (a) Groundtruth pitch (lines and dotted lines) and estimated pitch (circles and crosses) by Jin and Wang. (b) By Wohlmayr *et al.* SD. (c) By SD-DNN. (d) By SPD-DNN.

Wohlmayr *et al.* modeled speakers with GMMs, and used a mixture maximization model to obtain a probabilistic representation of pitch states. An FHMM was then applied to track pitch over time. The GMM-FHMM structure could also be extended to be gender-dependent. We denote Wohlmayr *et al.*'s speaker-dependent and gender-dependent models as Wohlmayr *et al.* SD and Wohlmayr *et al.* GD, respectively. Wohlmayr *et al.* trained their models on the GRID database with groudtruth pitches obtained also by RAPT. The test mixtures used in their study are included in our test set, and we directly adopt their trained models during evaluation.

We first evaluate the SD-DNN and SPD-DNN based methods. Table 1 compares the SD-DNN and SPD-DNN based methods with the other multipitch trackers on 600 test mixtures. Speaker-dependent approaches perform substantially better than speaker-independent approaches, and our SD-DNN and SPD-DNN based methods cut E_{Total} by more than 10% compared to Wohlmayer *et al.* SD. The major improvement in E_{Total} comes from E_{21}, which implies that our methods estimate pitch more accurately when the two speakers are both voiced. The SPD-DNN method performs better than the SD-DNN method, which is not surprising as SPD-DNNs are trained on individual speaker pairs. We further illustrate E_{Total} for each of the six speaker pairs in Fig. 5. As shown in the figure, our methods have lower errors across all pairs. SD-DNNs and SPD-DNNs perform comparably on five speaker pairs, and the latter achieve significantly lower E_{Total} on the most difficult pair of MA1-MA2. Fig. 6 illustrates pitch tracking results on one test mixture of MA1-MA2. Jin and Wang's approach fails to assign pitches to the underlying speakers. Wohlmayr *et al.*'s approach works better in terms of speaker assignment, but performs poorly when two pitch tracks are close to each other. Moreover, their resulting pitch contours lack continuity. The SD-DNN produces much smoother pitch contours. However, it still has incorrect speaker assignment at a few

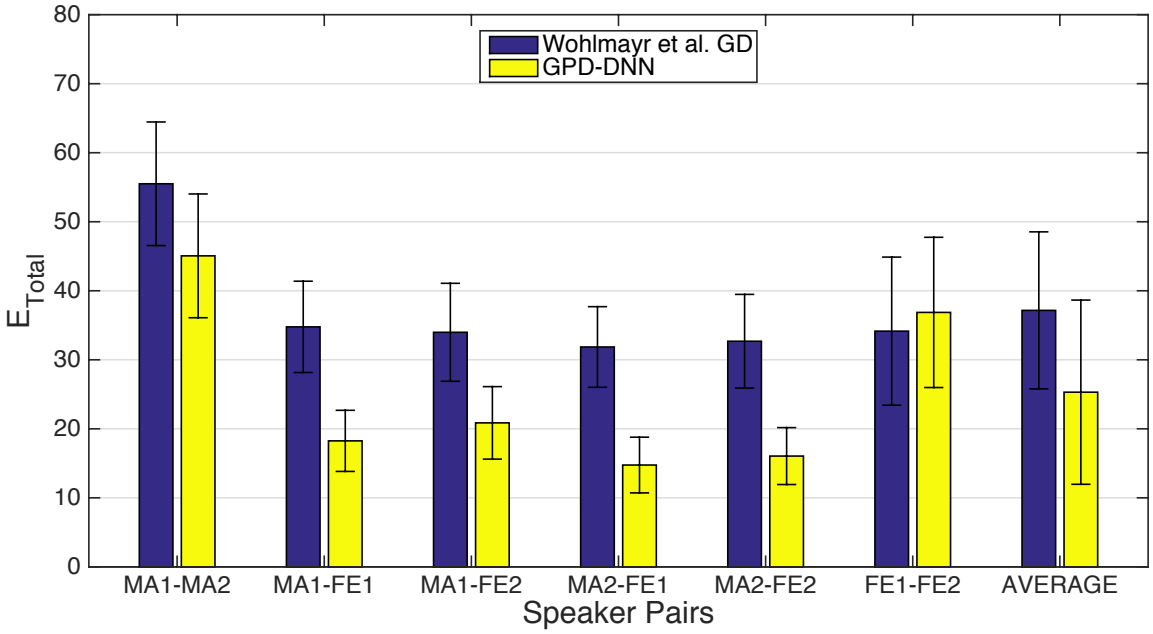

Figure 7: E_{total} of gender-dependent approaches. Error bars depict the mean and standard deviation of a method on the test mixtures of a given speaker pair.

frames. The SPD-DNN generates very good pitch tracks in both pitch accuracy and speaker assignment.

Next, we evaluate three extensions to the previous models. Fig. 7 shows the performance of the GPD-DNN based method. It produces comparable results to Wohlmayer *et al.*'s gender dependent model on same-gender mixtures but significantly outperforms it on male-female mixtures. The average E_{Total} of GPD-DNN is 11.85% lower than Wohlmayr *et al.*'s gender-dependent model, and even 1.42% lower than Wohlmayr *et al.*'s speaker-dependent model. However, the performance gap between GPD-DNNs and SD-DNNs/SPD-DNNs is larger than 8%. Therefore one should use SD-DNN/SPD-DNN based methods when speaker-dependent information is available.

Fig. 8 shows the performance of GPD-DNN adaptation. Four models are compared across all speaker pairs: (1) GPD-DNNs, (2) SPD-DNNs directly trained with 100 mixtures per speaker pair, (3) GPD-DNNs adapted with 100 mixtures per speaker pair, (4) SPD-DNNs trained with 57000 mixtures per speaker pair. As shown in the figure, SPD-DNNs trained with limited data perform better than GPD-DNNs on same-gender mixtures, but worse than GPD-DNNs on different-gender mixtures. GPD-DNN adaptation consistently outperforms the first two methods, resulting in more than 6% reduction in average E_{Total}. The results indicate the superiority of GPD-DNN adaptation for small training sizes.

Generalization to different speaker energy ratios is crucial to supervised multipitch trackers. Fig. 9 shows the performance of SD-DNN, SPD-DNN, and Wohlmayr *et al.*'s speaker-dependent models at various speaker ratios. All models are trained at 0 dB, and results are

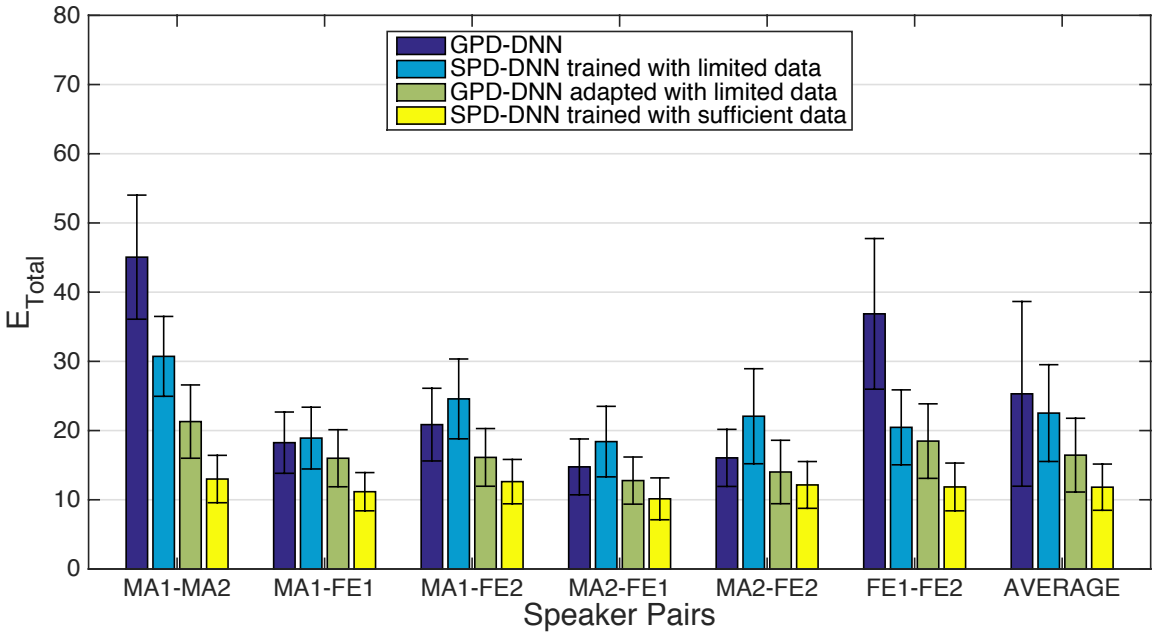

Figure 8: Performance of GPD-DNN adaptation. Error bars depict the mean and standard deviation of a method on the test mixtures of a given speaker pair.

averaged across all speaker pairs at each speaker ratio. As shown in the figure, the total error increases significantly when the speaker ratio deviates from 0 dB. Errors are not symmetric with respect to 0 dB, as we only scale the level of one speaker in order to create a specified ratio. For Wohlmayr *et al.*'s speaker-dependent model, when the speaker ratio is positive, the mixture becomes dominated by the amplified speaker, misleading the GMM of the weak speaker. For the SD-DNN and SPD-DNN based methods, it is hard for DNNs to recognize the weak speaker when the speaker ratio is too low. We then apply multi-ratio training for SD-DNNs and SPD-DNNs, and compare them with two unsupervised multipitch trackers, i. e., Jin and Wang, Hu and Wang, as well as the gain-adapted version of Wohlmayr *et al.*'s speaker-dependent models [29]. The results are given in Fig. 10. The performance of multi-ratio trained DNNs remains high across all speaker ratios. At 0 dB, multi-ratio trained SD-DNNs and SPD-DNNs produce only 0.07% and 0.29% higher errors than SD-DNNs and SPD-DNNs trained in the matched 0 dB condition, indicating their strong generalization ability.

In the above experiments, we use RAPT to extract the groundtruth pitch from single speaker recordings, which is error-prone in some situations. We now evaluate our methods on the FDA database [3], where the groundtruth pitch is directly given by laryngograph data. The corpus consists of recordings of 50 sentences by each of two speakers (a male and a female). For each speaker, we choose 40 sentences for testing and 40 test mixtures are created by mixing the test sentences at 0 dB. Because the energy level of sentences in the FDA database may be different from that in the GRID database, we use a ratio-adapted GPD-DNN trained on the GRID database for pitch-state probability estimation. We also perform speaker adaptation

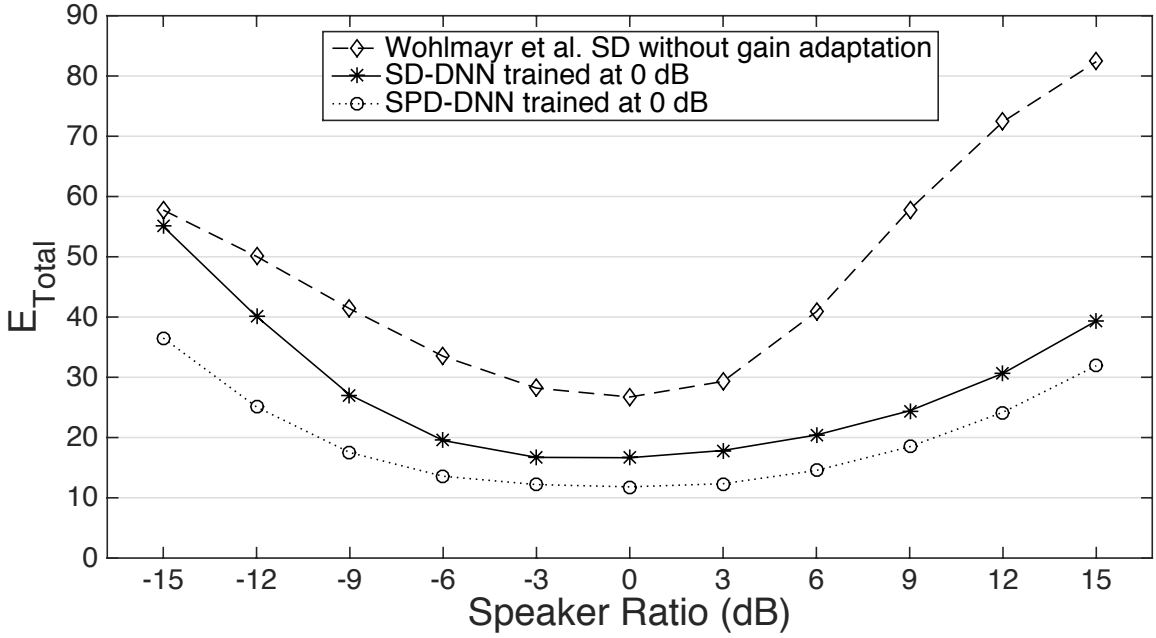

Figure 9: Results of different approaches tested on eleven speaker ratios. Each data point represents E_{total} averaged across 1200 test mixtures.

of the GPD-DNN with 10 adaptation sentences per speaker, i. e., 10×10 adaptation mixtures. We compare our methods with two unsupervised approaches, i. e., Jin and Wang, Hu and Wang, as well as the gain-adapted version of Wohlmayr *et al.*'s gender-dependent model. E_{Total} of different approaches is shown in Fig. 11. Results indicate that our GPD-DNN based method outperforms other approaches. The adaptation of the GPD-DNN further reduces the average total error by 8.42%.

In addition to E_{Total}, we use another metric to compare the performance in this experiment: overall multipitch accuracy used by Duan *et al.* [9]. To compute this accuracy, we first assign each estimated pitch track to a groundtruth pitch track. For each estimated pitch track, we call a pitch estimate at a frame correct if it deviates less than 10% from its corresponding groundtruth pitch. The overall multipitch accuracy is defined as:

$$Accuracy = \frac{TP}{TP + FP + FN} \tag{12}$$

where TP (true positive) is the total number of correctly estimated pitches, FP (false positive) is the total number of pitches that appear in some estimated pitch track but do not belong to the corresponding groundtruth pitch track, and FN (false negative) denotes the total number of pitches that appear in some groundtruth pitch track but do not belong to the corresponding estimated pitch track. Different assignments of estimated pitch tracks give us different accuracies, and we choose the highest value to represent the overall accuracy. Similar to Fig. 11, the GPD-DNN and GPD-DNN adaptation achieve accuracies of 69.78% and 82.58%. Hu and

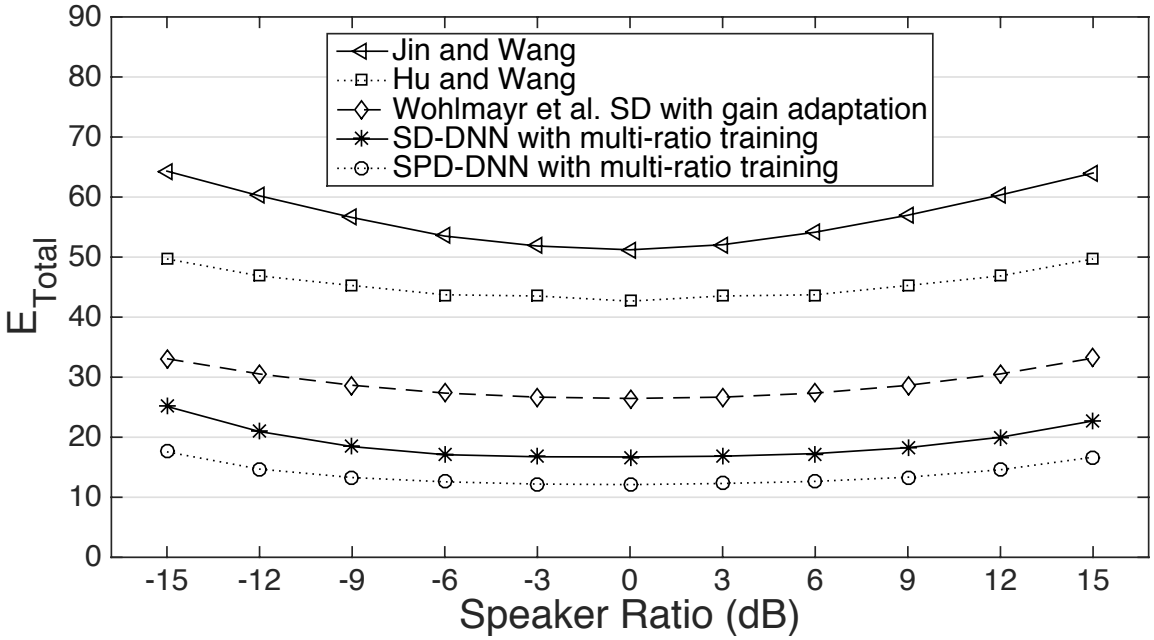

Figure 10: Results of different approaches tested on eleven speaker ratios. Each data point represents E_{total} averaged across 1200 test mixtures.

Table 2: Running time comparion for different approaches

	Jin and Wang	Hu and Wang	Wohlmayr *et al.* SD	SD-DNN	SPD-DNN
Time (s)	7.77	14.00	20.12	0.30	0.25

Wang's approach achieves an accuracy of 63.67%. The other two approaches have accuracies lower than 50%.

Lastly, we compare the computational complexity of different approaches. One hundred mixtures with the total length of 179.7 s are created for this evaluation. The test is performed on a machine with Intel i7-4770k CPU, 32 GB memory and NVIDIA GTX 780 GPU. Table 2 shows the average processing time per one second mixture. Results indicate that our methods are a lot more efficient.

7 Concluding Remarks

We have proposed speaker-dependent and speaker-pair-dependent DNNs to estimate the posterior probabilities of pitch states for two simultaneous speakers. Taking advantage of discriminative modeling and speaker-dependent information, our approach produces good pitch estimation in terms of both accuracy and speaker assignment, and significantly outperforms other state-of-the-art multipitch trackers. The SPD-DNN based method performs especially well when the two speakers have close pitch tracks. In order to relax constraints, we have introduced three extensions to SD-DNNs and SPD-DNNs. Gender-pair-dependent DNNs are designed for unseen speakers during testing, and they perform substantially better than other

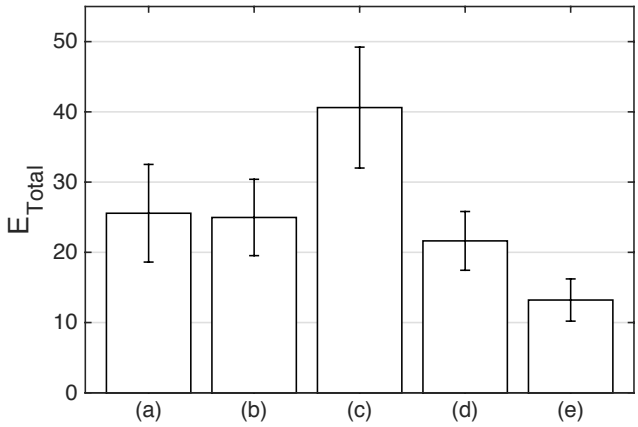

Figure 11: Results of different approaches tested on the FDA corpus. (a) Jin and Wang, (b) Hu and Wang, (c) Wohlmayr *et al.* GD with gain adaptation, (d) Ratio-adapted GPD-DNN and (e) Speaker adaptation of GPD-DNN. Error bars depict the mean and standard deviation of a method on the test mixtures.

speaker-independent and gender-dependent approaches on both GRID and FDA databases. Given limited speaker-dependent training data, speaker adaptation is effective for reducing tracking errors. Lastly, multi-ratio trained SD-DNNs and SPD-DNNs produce consistent results across various speaker ratios.

To apply our speaker-dependent models requires that the identities of the two speakers be known beforehand. Recently, Zhao *et al.* [34] proposed a DNN-based cochannel speaker identification algorithm, which can reliably identify the speakers in two-talker mixtures. Such an algorithm could be used to first identify the speakers in an input mixture, thus helping select trained SD-DNNs or SPD-DNNs.

Although the proposed models are designed for two-talker mixtures, they can be extended to mixtures with more than two speakers. In such cases, one would first estimate the pitch-state probabilities for each speaker using the corresponding SD-DNN. Then an FHMM with multiple Markov chains can be used to track several pitch tracks simultaneously. Many multipitch trackers [31] [18] deal with interfering speakers and additive noise at the same time. We can also extend our models to deal with background noise by training on a large set of noise corrupted mixtures.

To make use of the temporal context, we concatenate neighboring frames into a feature vector. Such a method can only capture temporal dynamics in a limited span. On the other hand, recurrent neural networks (RNNs) have self connections through time. Studies have shown that RNNs are good at modeling sequential data like handwriting [12] and speech [26]. We plan to explore RNNs in future work to better capture the temporal context.

Acknowledgments

We would like to thank M. Wohlmayr, M. Stark and F. Pernkopf for providing their pitch tracking code to us. This research was supported in part by an AFOSR grant (FA9550-12-1-0130) and the Ohio Supercomputer Center.

References

[1] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proceedings of ICASSP*, 2013, pp. 7942–7946.

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

[3] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," in *Proceedings of Eurospeech*, 1993, pp. 1003–1006.

[4] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," in *Glot Int.*, vol. 5, 2001, pp. 341–345.

[5] C. Chen, R. Gopinath, M. Monkowski, M. Picheny, and K. Shen, "New methods in continuous mandarin speech recognition," in *Proceedings of Eurospeech*, 1997, pp. 1543–1546.

[6] 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.

[7] 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.

[8] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of Interspeech*, 2011, pp. 1973–1976.

[9] 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.

[10] Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

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

[12] A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, and S. Fernández, "Unconstrained on-line handwriting recognition with recurrent neural networks," in *Proceedings of NIPS*, 2008, pp. 577–584.

[13] K. Han and D. L. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 2158–2168, 2014.

[14] 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.

[15] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[16] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 2067–2079, 2010.

[17] 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.

[18] 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.

[19] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1999.

[20] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proceedings of ICASSP*, 2013, pp. 7947–7951.

[21] Y. Liu and D. L. Wang, "Speaker-dependent multipitch tracking using deep neural networks," in *Proceedings of Interspeech*, in press, 2015.

[22] R. Peharz, M. Wohlmayr, and F. Pernkopf, "Gain-robust multi-pitch tracking using sparse nonnegative matrix factorization," in *Proceedings of ICASSP*, 2011, pp. 5416–5419.

[23] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proceedings of ASRU*, 2013, pp. 55–59.

[24] F. Sha and L. K. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Proceedings of NIPS*, 2005, pp. 1233–1240.

[25] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding Synth.*, vol. 495, p. 518, 1995.

[26] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," in *Proceedings of ICASSP*, 2012, pp. 4085–4088.

[27] D. L. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006.

[28] Y. Wang, J. Chen, and D. L. Wang, "Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training," *Technical Report OSU-CISRC-9/14-TR16*, Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA, 2014.

[29] M. Wohlmayr and F. Pernkopf, "EM-based gain adaptation for probabilistic multipitch tracking," in *Proceedings of Interspeech*, 2011, p. 196901972.

[30] 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.

[31] 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.

[32] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proceedings of ICASSP*, 2013, pp. 7893–7897.

[33] X. Zhao, Y. Shao, and D. L. Wang, "CASA-based robust speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1608–1616, 2012.

[34] X. Zhao, Y. Wang, and D. L. Wang, "Cochannel speaker identification in anechoic and reverberant conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1727–1736, 2015.