Computer Speech and Language 102 (2027) 102028

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Talker-independent multi-pitch tracking in noisy and reverberant scenarios

Yixuan Zhang [a,b,*], Heming Wang [c], DeLiang Wang [d]

[a] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43221, USA
[b] Tencent HunYuan, Bellevue, WA, 98004, USA
[c] Microsoft, Mountain View, CA, 94043, USA
[d] School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, China

ARTICLE INFO

Dataset link: https://github.com/YIXUANZ/multipitch

Keywords:
Multi-pitch tracking
Speaker separation
TF-CrossNet
DC-CRN
Room reverberation

ABSTRACT

Talker-independent multi-pitch tracking is a persistent challenge in speech processing in real-world environments with room reverberation and background noise. This paper addresses this challenge by investigating the accuracy of ground truth pitch states, training strategy, and model architecture. We find that, while the use of synthetic data is advantageous in single-talker pitch tracking for improving the accuracy of pitch labels, it is unsuitable for multi-pitch tracking. Thus, we propose different ground truth pitch state generation methods (hereafter referred to as "label generation") depending on what ground truth pitch state should be in reverberant speech. The proposed label generation leverages a single-talker pitch tracking model trained with synthetic data and a laryngograph-based voicing detection model. We propose a multi-pitch tracking model by utilizing a talker-independent speaker separation module and a pitch-tracking module, trained sequentially. Systematic evaluations demonstrate that the proposed model obtains superior performance compared to baseline methods across various scenarios and under different ground truth references. For instance, our model achieves a relative total error reduction of 32% over the best baseline in clean same-gender scenarios; in noisy-reverberant conditions, we reduce the total error rate by 50% for same-gender mixtures compared to conventional multi-pitch tracking approaches. To facilitate replication and future multi-pitch tracking research, we make pretrained models and a test set of manually-corrected ground truth pitch tracks of reverberant LibriSpeech utterances publicly available.

1. Introduction

As a fundamental attribute of speech, pitch is utilized in many speech processing tasks, such as speech synthesis (Kawahara et al., 2008; Morise et al., 2016; Takagi et al., 2025) and emotion recognition (Busso et al., 2009; Stasiak and Rychlicki-Kicior, 2012). Recently, pitch has also been shown to be effective in addressing the permutation ambiguity problem in talker-independent speaker separation (Taherian et al., 2024). Pitch tracking in multi-talker scenarios is challenging due to the difficulty of accurately estimating speech pitch in the presence of interfering speech and correctly associating each pitch point with its corresponding speaker. These difficulties become more significant in noisy or reverberant environments.

Previous studies on multi-pitch tracking can be mainly categorized into two types. The first type, which is commonly employed in signal processing and computational auditory scene analysis (CASA) algorithms, involves acoustic modeling followed by talker-independent tracking. While most of these algorithms are built to track two pitch contours without assigning them to specific

* Corresponding author.
E-mail address: zhang.7388@buckeyemail.osu.edu (Y. Zhang).

https://doi.org/10.1016/j.csl.2026.102028
Received 23 November 2025; Received in revised form 5 June 2026; Accepted 15 July 2026
Available online 21 July 2026

speakers (Bach and Jordan, 2005; Christensen and Jakobsson, 2009; Jin and Wang, 2010), a few methods consider speaker assignment (Duan et al., 2013) but achieve only limited success. Methods that consider speaker assignment typically attempt to cluster continuous pitch contours into two speakers, which is found challenging because individual pitch contours are often too short for effective clustering.

The second type employs deep learning methods, where speaker separation serves as an initial step followed by pitch detection. These methods demonstrate significantly improved performance compared to signal processing and CASA methods. For instance, Liu and Wang (2018) apply utterance-level permutation invariant training (uPIT) for multi-pitch tracking by leveraging permutation information from the speaker separation stage for pitch tracking. A factorial hidden Markov model (FHMM) is then used to enhance pitch continuity. The performance of such multi-pitch tracking models is closely tied to the performance of speaker separation. Substantial progress has been made in speaker separation since the introduction of deep clustering (Hershey et al., 2016) and PIT (Kolbæk et al., 2017), which effectively address the permutation ambiguity issue in talker-independent training. Subsequent studies such as DeepCASA (Liu and Wang, 2019), ConvTasNet (Luo and Mesgarani, 2019), DPRNN (Luo et al., 2020), TF-GridNet (Wang et al., 2023) and TF-CrossNet (Kalkhorani and Wang, 2024) have made large advances. It is found in Li et al. (2020) that speaker separation and pitch tracking are interrelated and mutually beneficial tasks. Beyond these approaches, self-supervised learning methods like EMSSL (Li et al., 2022), have also been explored, enabling pitch tracking to be trained on unlabeled data with reported performance comparable to supervised methods.

Despite these advances, existing methods have notable limitations. First, they often rely on conventional pitch tracking algorithms such as RAPT (Talkin and Kleijn, 1995) and PRAAT (Boersma, 2001) for ground-truth pitch extraction, which limits pitch label accuracy and, consequently, the quality of trained models. Note that, throughout this paper, we use pitch label (or simply label) to refer to ground-truth pitch state for model training. Second, deep neural network (DNN) architectures of current pitch tracking models do not incorporate recent designs such as self-attention mechanisms, likely restricting their performance. Third, optimal training strategies for the dual modules of speaker separation and pitch tracking remain under-explored. Finally, most existing studies are evaluated in clean conditions, and they are untested in challenging conditions such as reverberant and noisy environments.

The goal of this study is to enhance multi-pitch tracking performance across clean, reverberant, and noisy reverberant conditions. While synthetic data has been proven effective for single-talker pitch tracking, we find it inadequate for multi-pitch tracking due to poor generalization of speaker separation models trained on synthetic mixtures. Previous models often rely on simplistic DNN architectures and lack robust training strategies, resulting in suboptimal multi-pitch tracking performance, particularly in challenging reverberant and noisy environments. Additionally, what should be considered the ground-truth pitch for reverberant speech remains unclear, limiting the development of supervised pitch tracking methods where training labels must be provided. Our work tackles these challenges through enhanced label generation, a two-stage DNN architecture, and optimized training approaches to achieve superior multi-pitch tracking performance.

Our main contributions are summarized below.

- **Improved Pitch Label Accuracy**: We propose two methods for improving the accuracy of ground truth pitch extraction. For anechoic speech, we combine a single-talker pitch tracking model trained on synthetic data and a laryngograph-based voicing detection model. For reverberant speech, we introduce a combined estimator that leverages both PRAAT and data-driven methods.
- **Two-Stage Architecture for Multi-pitch Tracking**: We propose a two-stage, talker-independent multi-pitch tracking model, adapting TF-CrossNet (Kalkhorani and Wang, 2024) for speaker separation and DC-CRN (Zhang et al., 2023) for pitch tracking, substantially outperforming existing data-driven methods.
- **Optimized Training**: We investigate sequential and cascade training strategies and find that sequential training shows slightly better results.
- **Ground-Truth Pitch Analysis for Reverberant Speech**: We address the fundamental question of what the ground-truth pitch label should be for reverberant speech. After examining a definition based on anechoic speech reflecting a speech-production perspective, and another based on reverberant speech from a speech-perception viewpoint, we suggest that anechoic speech pitch should be the preferred pitch label.
- **A New Public Dataset and Pretrained Models**: Finally, we provide a test set with manually corrected labels for reverberant speech pitch to support evaluation. The test set and pretrained models are publicly available at Github to facilitate future research (https://github.com/YIXUANZ/multipitch).

The remainder of this paper is structured as follows. Section 2 presents the proposed multi-pitch tracking algorithm, where the two-stage architecture and optimized training are described. Section 3 describes the experimental design. Section 4 provides the evaluation and comparison results, and the label generation methods for anechoic speech and reverberant speech are presented in Sections 4.1 and 4.5, respectively; Section 4.5 also provides a ground-truth pitch analysis for reverberant speech. Section 5 concludes the paper.

2. Algorithm description

As illustrated in Fig. 1, our proposed multi-pitch tracking model adopts a framework similar to that in Liu and Wang (2018) and Li et al. (2020), utilizing an architecture with two sequential modules. The first module performs speaker separation, while the second handles pitch tracking using DC-CRN (Densely-Connected Convolutional Recurrent Network) (Zhang et al., 2023). In the figure, $\mathbf{X}$ denotes the complex short-time Fourier transform (STFT) of the input mixture, $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$ those of the estimated speech signals,

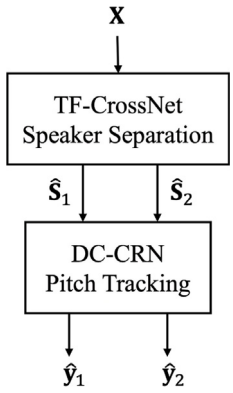

Fig. 1. Diagram of the proposed multi-pitch tracking model framework.

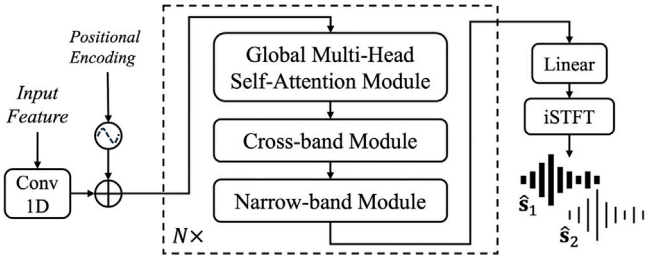

Fig. 2. Diagram of the TF-CrossNet architecture consisting of N TF-CrossNet blocks. $\hat{\mathbf{s}}_j$ represents the estimated signal for speaker j.

and $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ the pitch estimates. The DC-CRN module processes complex STFTs through seven convolutional densely-connected (Conv-DC) blocks, which promote feature reuse and gradient flow. This is followed by a bidirectional long short-term memory (BLSTM) layer to capture temporal dependencies and fully connected layers that output probabilistic pitch state vectors. We explore several strategies to train the model: (1) multi-stage independent training, where the speaker separation and pitch tracking modules are trained separately; (2) sequential training, where the speaker separation model is trained independently, and the pitch tracking model is subsequently trained using the output produced by the trained separation model; and (3) cascade training, where the two models are trained simultaneously with a combined loss $L_{cascade}$,

$$L_{cascade} = L_{sep} + \alpha * L_{pitch}, \tag{1}$$

where L_{sep} and L_{pitch} are described in Section 2.3, and α is a coefficient to balance the two tasks. During cascade training, the separation module is optimized by both the gradients from L_{sep} to maintain reconstruction quality, and the back-propagated gradients from L_{pitch} that flows through the pitch tracking module. Simultaneously, the parameters of the pitch tracking module are updated directly by the gradients originating from L_{pitch}. This joint gradient flow ensures that the separation stage produces features that are conducive to the downstream pitch estimation task.

2.1. Speaker separation module

We employ the TF-CrossNet architecture (Kalkhorani and Wang, 2024) for our speaker separation module. TF-CrossNet achieves the state-of-the-art performance in speaker separation, including under noisy and reverberant conditions. As depicted in Fig. 2, the model consists of an encoder layer, N TF-CrossNet blocks, and a decoder layer, with each TF-CrossNet block containing a global multi-head self-attention (GMHSA) module, a cross-band module, and a narrow-band module.

The encoder layer extracts acoustic features from the input $X_{t,f}$, a complex-domain feature formed by concatenating the real and imaginary parts of the STFT of the input signal. A frame length of 32 ms with 50% overlap between adjacent frames is used in calculating STFTs. Before applying the STFT, the input signal is variance-normalized. The encoder layer converts the input feature with dimensions of $2 \times F \times T$ into an acoustic feature with dimensions of $C \times F \times T$, where C, F, and T represent the number of hidden channels, frequency bins, and frames, respectively. Following the encoder layer, random-chunk positional encoding (RCPE) is employed (Kalkhorani and Wang, 2024), to address the out-of-distribution problem in existing positional encoding schemes, demonstrating enhanced generalization capabilities for handling longer sequences. In RCPE, a contiguous chunk of positional embedding vectors is selected from a pre-computed positional encoding matrix during training. The selection is performed by,

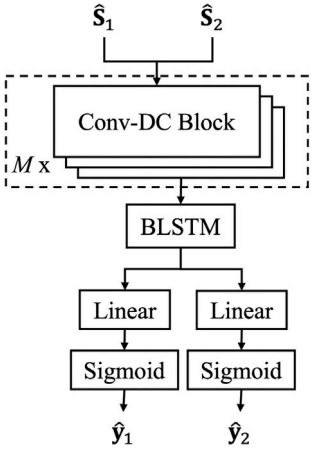

Fig. 3. Diagram of the DC-CRN architecture consisting of M convolutional densely-connected (Conv-DC) blocks.

$$\mathrm{RCPE}(t,i) = \begin{cases} \mathrm{PE}(t+\tau,i) & \text{if training,} \\ \mathrm{PE}(t,i) & \text{else,} \end{cases} \tag{2}$$

where, during training, a random index τ is selected from $[1, T^{max} - T + 1]$, and T^{max} is the maximum desired sequence length during inference. A chunk from index τ to $\tau + T - 1$ is then selected. For validation or testing, the first T embedding vectors are chosen. The pre-computed positional encoding matrix PE is defined as,

$$\mathrm{PE}(t,2i) = \sin\left(\frac{t}{10\,000^{\frac{2i}{F\times C}}}\right), \tag{3a}$$

$$\mathrm{PE}(t,2i+1) = \cos\left(\frac{t}{10\,000^{\frac{2i}{F\times C}}}\right). \tag{3b}$$

where $t \in [1, T]$ and $i \in [1, F \times C]$ represent the indices for the time and feature dimensions, respectively. The RCPE feature is added to the input features and fed to the TF-CrossNet blocks (see Fig. 2). The block outputs are sent to a linear decoder, mapping the processed features to the predicted real and imaginary parts of each speaker. An inverse STFT is finally applied to reconstruct the time-domain speech signals $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{s}}_2$.

In a TF-CrossNet block (Kalkhorani and Wang, 2024), the global multi-head self-attention (GMHSA) module begins by extracting frame-level features using a point-wise two-dimensional convolutional layer, followed by a parametric rectified linear unit (PReLU) activation and layer normalization. The frame-level features have $L(2E+C/L)$ channels, which are split into L sets of queries Q, keys K, and values V, serving as inputs to the multi-head self-attention mechanism that captures global correlations, with L representing the number of heads. The dimensions of the queries, keys, and values are $E \times F \times T$, $E \times F \times T$, and $C/L \times F \times T$, respectively. The outputs from the multi-head self-attention mechanism are concatenated and sent to another point-wise Conv2D layer with D output channels, followed by PReLU activation and layer normalization. The final output of the GMHSA module is formed by adding this processed output to the original input to the module. The cross-band module consists of a frequency-convolutional (Freq-Conv) module, a full-band linear module, and another Freq-Conv module. Each Freq-Conv module includes a layer normalization step, a grouped convolution layer along the frequency axis (F-GConv1D), and a PReLU activation function. The full-band linear module comprises a layer normalization step, a linear layer followed by a sigmoid-weighted linear unit (SiLU) activation function, several linear layers applied along the frequency axis, and a final linear layer followed by a SiLU activation function. The first linear layer reduces the number of hidden channels from C to C'. The parameters of the intermediate linear layers are shared across repeated TF-CrossNet blocks to improve parameter efficiency. The final linear layer restores the number of hidden channels back to C. The output of the cross-band module is the sum of the output from the final linear layer and the input to the final linear layer. In the narrow-band module, the first linear layer increases the number of hidden channels from C to C'', while the final layer restores the number of channels back to C. The T-Conv layer contains three grouped one-dimensional convolution (T-GConv1D) layers, each followed by a SiLU activation function. The second T-GConv1D layer is additionally followed by a grouped normalization layer. In terms of network configuration details, we set the kernel size of the encoder layer, F-GConv1D, and T-GConv1D to 5, 3, and 5, respectively. We set the number of groups for F-GConv1D, T-GConv1D, and group normalization to 8. The model has $N = 12$ TF-CrossNet blocks, with hidden channel sizes set to $C = 192$, $C' = 16$, and $C'' = 384$. The GMHSA module has 4 self-attention heads with an embedding dimension of $D = 64$ and $E = \lceil 512/F \rceil$, where $\lceil \cdot \rceil$ represents the ceiling operation.

2.2. Pitch tracking module

2.2.1. Problem formulation

Following recent data-driven approaches for pitch tracking (Han and Wang, 2014; Ardaillon and Roebel, 2019; Kim et al., 2018; Zhang et al., 2023; Laube and Panos, 2025), we formulate the task as a multi-class classification problem, where each class corresponds to a distinct pitch state with an associated fundamental frequency. As shown in Fig. 3, the network outputs two vectors, $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, one for each speaker. Each element of a vector represents the probability of the $F0$ belonging to the corresponding pitch state, with the first element indicating the voicing state. The target fundamental frequency range considered aligns with previous studies (Ardaillon and Roebel, 2019; Zhang et al., 2023), spanning from 30 Hz to 1000 Hz. This range covers all possible fundamental frequencies in human voices. With an interval of 12.5 cents, the pitch range is divided into 486 states, corresponding to $y_k^i, i = 2, 3, \ldots, 487$ in the target F_0 vector $\boldsymbol{y}_k, k = 1, 2$. The value of the ith state is calculated by,

$$y_k^i = \exp\left[-\frac{(p_i - p_{true})^2}{2 \cdot 25^2}\right], \tag{4}$$

where p_i and p_{true} represent the fundamental frequency in cents of the ith pitch state and the ground-truth pitch state, respectively. To reduce the penalty for near-correct estimates, the target vector is Gaussian-blurred instead of directly using a one-hot vector. The element y_k^1 represents voicing. A frame is classified as voiced only if it contains a periodic or quasi-periodic speech signal, and y_k^1 is set to 1 for voiced speech and 0 for unvoiced speech, noise, or silence.

With the estimated output vectors $\hat{\mathbf{y}}_k$, the final pitch estimate $\widehat{F0}_k$ is calculated below,

$$I = \arg\max_{i \in \{2,3,\ldots,487\}} \hat{y}_k^i, \qquad \hat{p} = \frac{\sum_{i=I-4}^{I+4} \hat{y}_k^i p_i}{\sum_{i=I-4}^{I+4} \hat{y}_k^i}, \tag{5}$$

$$\hat{f}_k = \hat{p} \cdot \mathbb{I}(\hat{y}_k^1 > 0.5). \tag{6}$$

In Eq. (5), the index I denotes the index of the maximum element in $\hat{\mathbf{y}}_k$. Following the pitch refinement technique proposed in Kim et al. (2018), the pitch estimate $\hat{p}$ is then calculated as the weighted average of the Ith pitch state and its neighboring pitch states, $p_{I-4}, \ldots, p_{I+4}$. When the indices of certain neighboring pitch states fall outside the valid range (i.e., less than 2 or greater than 487), the summation in both the numerator and the denominator is dynamically restricted to those states whose indices remain within the valid range. The voicing indicator function $\mathbb{I}(\hat{y}_k^1 > 0.5)$ equals 1 if $\hat{y}_k^1$ exceeds the threshold of 0.5, and 0 otherwise. For evaluation, the calculated $\widehat{F0}_k$ is converted to Hz.

2.2.2. Model architecture

The pitch tracking module is adapted from the single-talker pitch tracking model for noisy speech presented in a previous study (Zhang et al., 2023). We extend this $F0$ estimation model to the two-speaker setting. The input to the network is a concatenation of the real and imaginary parts of the complex STFT of the estimated speech signals. To calculate STFT, we use a 128 ms Hamming window with 10 ms frame shift. The network consists of seven densely-connected convolutional (Conv-DC) blocks, a two-layer bidirectional long short-term memory (BLSTM) block, and two fully connected layers with sigmoidal activation functions, which produce $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, respectively. Each Conv-DC block comprises four composite layers followed by a gated convolutional layer. Each composite layer includes a 2D convolutional layer, batch normalization, and the exponential linear unit. The input to each composite or gated convolutional layer is formed by concatenating the outputs from all preceding layers, which reuses features computed in earlier layers, improving the information flow between layers. The final layer is a gated convolutional layer that integrates gated linear units (Dauphin et al., 2017). A grouping strategy (Gao et al., 2018) is adopted in the two-layer BLSTM, which reduces computational complexity without significantly degrading performance. The input features and hidden states of both BLSTM layers are divided into four disjoint groups. In the first BLSTM layer, intra-group features are learned within each group. The second BLSTM layer receives the rearranged outputs from the first layer and models inter-group dependencies. Layer normalization is applied following each recurrent layer. The final output is obtained by concatenating the outputs of the second BLSTM layer.

2.3. Loss functions

2.3.1. Speaker separation

For the speaker separation module, we adopt a loss function similar to one in Wang et al. (2023). Specifically, it is a scale-invariant signal-to-distortion ratio (SI-SDR) loss with a mixture constraint, which is applied as a loss term measured from the sum of the target sources and the sum of the scaled estimated sources.

$$L_{sep} = L_{\text{SI-SDR}} + \frac{1}{N}\left\| \sum_{k=1}^{2} \hat{\beta}_k \hat{s}_k - \sum_{i=1}^{2} s_k \right\|_1, \tag{7}$$

$$\hat{\beta}_k = \arg\min_{\beta} \|\beta \hat{s}_k - s_k\|_2^2, \tag{8}$$

where N denotes the number of samples, $\|\cdot\|_1$ computes L_1 norm, $\|\cdot\|_2^2$ computes L_2 norm, $\hat{\beta}_k$ is the scaling parameter for the estimate, and $\hat{s}_k$ denotes the re-synthesized signal for speaker k.

2.3.2. Pitch tracking

To learn the probabilistic outputs $\hat{\boldsymbol{y}}_k$, we minimize the binary cross-entropy loss L_{pitch},

$$\begin{aligned} L_{pitch}(\hat{\boldsymbol{y}}_k, \boldsymbol{y}_k) &= -y_k^1 \log \hat{y}_k^1 - (1 - y_k^1) \log (1 - \hat{y}_k^1) \\ &+ w * \left(\frac{1}{P} \sum_{i=2}^{P+1} [-y_k^i \log \hat{y}_k^i - (1 - y_k^i) \log (1 - \hat{y}_k^i)] \right), \end{aligned} \tag{9}$$

$$L_{pitch} = \sum_{k=1}^{2} L_{pitch}(\hat{\boldsymbol{y}}_k, \boldsymbol{y}_k), \tag{10}$$

where P is the number of pitch estimates, which equals 486 in our setting, w is set to 10 to align the numerical scales of the voicing and pitch tracking components (see Zhang et al. (2023)), and the matched permutation from the speaker separation module is used to calculate the loss.

3. Experimental setup

3.1. Datasets

3.1.1. Training data

For experiments in Sections 4.1 and 4.2, we create our training sets by randomly mixing utterances sampled from the LibriSpeech train-clean-360 subset (Panayotov et al., 2015). The training set contains 60k mixtures with samples mixed at −5 to 5 dB. All signals are resampled to 8 kHz. For experiments conducted under noisy and reverberant conditions, we further enlarge the training set to 100k mixtures and adapt the data to align with the specific experimental setups, which are detailed in Sections 4.4 and 4.5.

3.1.2. Test data

The PTDB test set used in Sections 4.1 and 4.2 is constructed from utterances in the PTDB-TUG dataset (Pirker et al., 2011). For either same-gender or different-gender cases, 100 speaker pairs are randomly selected from different speakers to create audio mixtures. Each pair is mixed at 0 dB, ±3 dB, ±6 dB, and ±9 dB. The provided ground-truth labels from PTDB-TUG, obtained from laryngograph data, are used for evaluation. All audio files are resampled to 8 kHz.

The ReverbTIMIT test set used for evaluating models trained in reverberant conditions is constructed from a part of the evaluation test set in Jin and Wang (2010) that focuses on multi-talker pitch tracking in reverberation, comprising a total of 175 pairs of same-gender and different-gender mixtures. Similar to the PTDB test set, each pair is mixed at 0 dB, ±3 dB, ±6 dB, and ±9 dB. The provided ground-truth labels are generated using an interactive pitch detection algorithm (PDA) described in Jin and Wang (2010). Depending on how the ground truth is defined in reverberant conditions, we create two versions of the ReverbTIMIT test set: ReverbTIMIT-a treats pitch labels from anechoic speech as the ground truth, while ReverbTIMIT-r treats pitch labels from reverberant speech as the ground truth. We also create two versions of the ReverbTIMIT-noisy test set - ReverbTIMIT-noisy-a and ReverbTIMIT-noisy-r - for noisy and reverberant conditions by adding factory noise to subsets of the test set where speakers are mixed at 0 dB. The noise is added so that the signal-to-noise ratio (SNR) between the first speaker and the noise is set to −5 dB, 0 dB, or 5 dB.

3.2. Evaluation metrics

Unless explicitly stated, we report the performance of multi-pitch tracking using the total error measure E_{total} (Wohlmayr et al., 2010), which considers pitch estimation error, voicing detection error, and speaker assignment error, but using a 5% frequency deviation threshold instead of the original 20% specified in Wohlmayr et al. (2010),

$$\begin{aligned} E_{total} = {} & E_{01} + E_{02} + E_{10} + E_{12} + E_{20} + E_{21} \\ & + E_{gross} + E_{fine} + E_{perm}, \end{aligned} \tag{11}$$

where E_{ij} denotes the percentage of time frames in which i pitch points are misclassified as j pitch points. For example, E_{01} represents the percent of time frames where unvoiced frames (0 pitch points) are incorrectly classified as having 1 pitch point. E_{gross} represents the percent of time frames with a frequency deviation exceeding 5% for one or both references. E_{fine} is the average frequency deviation in percent for time frames where the deviation is less than 5%. Lastly, E_{perm} quantifies the error due to incorrect speaker assignment in multi-talker pitch tracking. It equals the percent of time frames where the pitch frequency deviation exceeds 5% from the ground-truth pitch of the corresponding speaker but within 5% of that of the other speaker, indicating a swapped speaker assignment.

We adopt the 5% frequency deviation threshold instead of the original 20% specified in Wohlmayr et al. (2010) based on the following considerations. The 20% standard (approx. 318 cents) was originally designed for evaluating traditional signal-processing methods with limited accuracy. Contemporary deep-learning models have reached a level of accuracy where the 20% threshold is too coarse for meaningful performance differentiation. For instance, state-of-the-art single-pitch tracking models (Kim et al., 2018; Ardaillon and Roebel, 2019) commonly utilize a 50-cent threshold (≈2.93%) derived from musical pitch perception. Consequently, our selection of the 5% threshold (≈86.3 cents) provides a more stringent and appropriate metric for evaluating modern multi-pitch tracking systems. It also provides a conservative and reliable bound that safely covers the just-noticeable difference (JND) in pitch perception while isolating structurally meaningful acoustic shifts.

3.3. Training protocols

We train our models using the Adam optimizer (Kingma and Ba, 2014). We use the PyTorch OneCycleLR scheduler with a maximum learning rate of 0.001, where the number of steps per epoch is defined by the size of the training dataset. Gradient clipping is applied with a maximum value of 1 to avoid gradient explosion. We set the maximum training epoch number to 200 for the speaker separation models and 150 for the pitch tracking models; all models converge within this limit. For models trained with cascaded training, we set α in Eq. (1) to 10 to balance speaker separation loss and pitch tracking loss during training. Similar to the selection of w, this hyperparameter choice is guided by the empirical findings in Zhang et al. (2023). As demonstrated in that study, directly combining loss terms with disparate numerical scales can lead to suboptimal performance. In this work, setting $\alpha = 10$ effectively scales the pitch tracking loss L_{pitch} to a range comparable to that of the speaker separation loss L_{sep}. This choice prevents the separation task from dominating the learning process.

4. Evaluation results and comparisons

4.1. Training data and label comparisons

Training label accuracy is a widely discussed issue in pitch tracking, particularly in single-talker scenarios. Conventional approaches (Liu and Wang, 2018; Li et al., 2020) rely on signal processing algorithms, such as RAPT (Talkin and Kleijn, 1995), to extract pitch labels from clean speech, which are then treated as ground truth. However, this introduces limitations since the accuracy of a pitch label extraction algorithm constrains the performance of the trained model. A common solution to this problem is the use of synthetic data (Kim et al., 2018; Ardaillon and Roebel, 2019; Zhang et al., 2023), where speech is synthesized based on known pitch tracks, ensuring perfectly accurate labels. While this method has been explored for single-talker scenarios, its application to multi-talker environments remains unexplored. Following Zhang et al. (2023), we employ the WORLD vocoder (Morise et al., 2016) to synthesize audio, which is then used to create training data by mixing synthesized speech samples. Specifically, we use the train-clean-360 subset of LibriSpeech to synthesize speech. A total of 60k training mixtures are created by randomly mixing synthesized utterances with SNRs between −5 and 5 dB. For multi-pitch tracking, we adopt a structure similar to that described in Section 2, but replace the speaker separation module with Sudo rm-rf (Tzinis et al., 2020), a network architecture designed for speaker separation. We evaluate the model on the PTDB test set described in Section 3 and find that its performance is significantly worse compared to the models trained on real mixtures with RAPT-extracted pitch labels. As shown in Table 1, the total error rates exceed 60%, which is considerably higher than the total error rates of the models trained using real recordings with RAPT-extracted labels, which range from 23% to 24%. This problem occurs because the separation module, trained on synthetic speech mixtures, does not generalize well to real recordings, and vice versa. To validate this observation, we randomly sample 10,000 speech mixtures from both the synthetic training set and the training set containing real mixtures. An experiment is conducted using Sudo-rm-rf to evaluate speaker separation performance across these different types of data. Training on mixtures of real-recorded speech achieves an SI-SDR of 16.04 dB on the real-recorded test set, whereas training exclusively on synthetic data results in a significantly lower SI-SDR of 0.64 dB on the same test set. Conversely, training on real mixtures yields an SI-SDR of 8.98 dB on the synthetic test set, while training on synthetic data performs well on the synthetic test set, achieving 16.29 dB. The observed results highlight the challenges of generalizing across real and synthetic speech domains.

This substantial performance gap noted above indicates a feature distribution mismatch between synthetic and real-recorded data. While a single-talker pitch tracker can effectively learn periodicity from synthetic waveforms, the speaker separation module relies on capturing subtle spectrotemporal features for separation. Synthetic vocoders often produce idealized harmonic structures and phase relation, which appear incapable of modeling complex talker interactions in recorded multi-talker mixtures. Consequently, the speaker separation network, when trained exclusively on synthetic data, fails to generalize to real speech features, leading to poor pitch tracking performance.

Given the limited generalization between recorded and synthetic speech, and the importance of speaker separation, we opt for recorded speech utterances. To utilize the label accuracy of synthetic data, one approach is to train a single-talker pitch tracker on clean synthetic speech and use it to generate ground-truth labels. Although limitations in label accuracy remain, this method benefits from the inherent accuracy of synthetic data. For label creation, we train the DC-CRN pitch tracking model following the setup in Zhang et al. (2023) but on clean speech and use it to generate pitch labels for voiced segments. As shown in Table 1, this approach consistently improves performance compared to the models trained on RAPT-labeled data, with an average improvement of 2.48%. However, in terms of voicing detection accuracy, the model trained on synthetic audio does not outperform RAPT. To address this, we adopt the method from Zhang et al. (2024), dubbed Robust Voicing Detection (RVD), where the voicing detection model is trained on laryngograph-based data to enhance accuracy. Using RVD further boosts performance by 2.69% on average. The models are trained in a cascaded manner, using Eq. (1) to jointly optimize the speaker separation and pitch tracking modules.

4.2. Training strategy comparisons

Additionally, we explore finetuning the model with laryngograph-based cross-corpus data, using 10k speech mixtures created from samples in CMU-Arctic (Kominek and Black, 2004), Mocha-TIMIT,[1] Keele (Plante et al., 1995), and FDA (Bagshaw et al., 1993) datasets mixed at −5 to 5 dB. We find that finetuning with laryngograph data further improves the performance, especially in different-gender scenarios, benefiting from the high accuracy of laryngograph-based ground-truth labels (see Table 2).

[1] https://data.cstr.ed.ac.uk/mocha.

Table 1
Total error rate (%) of the models trained with various training label generation methods. 'Voiced $F0$' refers to $F0$ estimation in voiced frames. 'V/UV' stands for voiced/unvoiced classification.

Training label		Same gender				Different gender			
Voiced $F0$	V/UV	0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
WORLD (Morise et al., 2016)	WORLD (Morise et al., 2016)	61.09	60.82	60.96	61.09	61.47	61.11	61.69	62.30
RAPT (Talkin and Kleijn, 1995)	RAPT (Talkin and Kleijn, 1995)	23.63	23.71	23.94	24.28	23.43	23.43	23.46	23.92
DC-CRN (Zhang et al., 2023)	RAPT (Talkin and Kleijn, 1995)	21.89	21.90	21.94	21.96	20.20	20.36	20.61	21.13
DC-CRN (Zhang et al., 2023)	RVD (Zhang et al., 2024)	**19.70**	**19.62**	**19.20**	**19.02**	**17.43**	**17.60**	**17.88**	**17.99**

Table 2
Total error rate (%) of models trained using different training strategies.

Training strategy	Same gender				Different gender			
	0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
Multi-stage independent	21.60	23.56	21.78	21.60	18.26	18.21	18.38	18.33
Sequential training	**19.24**	**19.29**	**19.02**	**18.89**	17.70	17.70	**17.77**	**17.89**
Cascade training	19.70	19.62	19.20	19.02	**17.43**	**17.60**	17.88	17.99
Sequential & Cascade	20.33	19.87	18.97	19.47	18.07	17.99	18.20	18.08
Cascade & Laryn. finetune	18.51	18.71	18.69	18.28	14.26	14.35	14.85	15.84

Table 3
Total error rate (%) of multi-pitch tracking models.

Methods	Same gender				Different gender			
	0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
uPIT-Pitch (Liu and Wang, 2018)	25.1	24.9	24.4	25.2	14.8	14.8	15.4	16.7
uPIT-SS-Perm-Pitch (Liu and Wang, 2018)	23.8	23.4	23.5	25.3	14.3	14.5	15.1	16.5
EMSSL-Pitch-v2 (Li et al., 2022)	28.3	27.8	28.2	28.9	17.6	18.0	18.4	19.2
Proposed method	**16.2**	**16.4**	**16.7**	**17.0**	**12.8**	**12.9**	**13.0**	**13.2**

4.3. Comparisons with baseline models

We compare our proposed model with several baseline models, following the training and test setups outlined in Liu and Wang (2018). The results are presented in Table 3, which demonstrates significant improvements obtained by our model in both same-gender and different-gender scenarios, particularly in the same-gender case. Note that, to align with the evaluation results of the baseline methods, we use the original 20% frequency deviation for calculating the total error rate. For instance, at 0 dB in the same-gender condition, our model reduces the total error rate from 23.8%, obtained by the best baseline of uPIT-SS-Perm-Pitch, to 16.2%, which corresponds to a relative improvement of over 31.93%. Similar improvements are observed across other scenarios, indicating the effectiveness of the proposed model in multi-pitch tracking, especially in challenging same-gender scenarios.

4.4. Evaluation in reverberant and noisy conditions

To evaluate model performance under reverberant (reverb for short) and noisy conditions, we extend the training set described in Section 3 to two training sets: a reverb-only training set and a noisy-reverb training set. For the reverb-only training set, we generate 15,000 room impulse responses (RIRs) using the image method (Allen and Berkley, 1979), with random room characteristics and reverberation times (T60) ranging from 0.2 s to 0.6 s. For each RIR generation, we define the room dimensions by randomly selecting the length and width from $[5, 10]$ m and the height from $[3, 4]$ meters. The microphone is placed within 0.5 m of the center of the room, at a height between 1 and 2 m. The speakers are positioned at the same height as the microphone, maintaining a distance of 0.75 to 2.5 m from it, while ensuring a minimum of 0.5 m from any wall. For each training mixture, an RIR pair for two speakers is randomly selected to simulate room reverberation. For the noisy-reverb training set, we further add noise to the reverberant mixtures by mixing each training speech mixture with a noise segment cut from the 10,000 noises extracted from a sound effect library (Chen et al., 2016). We apply a gain to the noise such that the SNR between the first speaker and the noise is set to a random value uniformly sampled from −5 to 5 dB. Under this evaluation setup, the speaker separation module is trained to directly estimate anechoic speech (see Kalkhorani and Wang (2024)), effectively performing joint speaker separation, and speech dereverberation and denoising, hence facilitating robust multi-pitch tracking.

4.4.1. Evaluation in reverberant conditions

To evaluate the performance of the proposed model in reverberant conditions and choose a speaker separation module, we train and evaluate two separation models under this setup, the proposed model described in Section 2 and a variant in which the separation module is replaced with TF-GridNet (Wang et al., 2023), a recent speaker separation method showing high performance. The two models are trained using the reverb-only training set, with anechoic speech serving as the training target for the speaker separation

Table 4
Total error rate (%) of multi-pitch tracking models in reverberant conditions, with pitch contours from anechoic speech as ground truth.

Test set	Sep.	Same gender				Different gender			
	module	0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
Libri-reverb	TF-GridNet	13.47	13.67	14.43	15.43	13.62	13.66	14.48	15.18
Libri-reverb	TF-CrossNet	**12.55**	**12.61**	**13.33**	**13.88**	**12.71**	**12.46**	**13.16**	**13.61**
ReverbTIMIT-a	TF-GridNet	23.62	**23.25**	23.68	24.82	21.56	21.68	21.89	22.48
ReverbTIMIT-a	TF-CrossNet	**23.36**	23.54	**23.48**	**24.69**	**20.44**	**20.61**	**21.01**	**21.88**

Table 5
Total error rate (%) of the proposed model trained under noisy-reverberant conditions, evaluated on different test sets. {}/{} dB indicates the SNR between the first speaker and noise, and the speaker-to-interference ratio.

Test set	Same gender			Different gender		
	5/−1.19	0/−3.01	−5/−6.19	5/−1.19	0/−3.01	−5/−6.19
Libri-reverb-noisy	18.01	24.21	38.45	18.08	23.99	37.65
ReverbTIMIT-noisy-a	28.36	36.60	52.43	28.29	35.34	51.92

module. We evaluate the trained models on two test sets, namely ReverbTIMIT-a as described in Section 3 and a Libri-reverb test set created from the Librispeech test-clean subset. Similar to the PTDB-TUG test set, we create same-gender and different-gender subsets, each with 100 speaker pairs. Each audio pair is mixed at 0 dB, ±3 dB, ±6 dB, and ±9 dB. To add room reverberation, we generate 200 RIR pairs, which are randomly selected to simulate reverberant speech pairs, following the same procedure used for creating the training set. The ground-truth labels are generated in accordance with the training label generation method described in Section 4.1. Table 4 presents the evaluation results. The proposed model consistently outperforms the TF-GridNet variant across both test sets, achieving lower total error rates. The improvement is more significant on the Libri-reverb test set. On average, the proposed model reduces the total error rate of the TF-GridNet variant by 1.2% on the Libri-reverb test set, and by 0.49% on the ReverbTIMIT-a test set. In addition, TF-CrossNet is significantly easier to train than TF-GridNet.

4.4.2. Evaluation in noisy and reverberant conditions

To further evaluate the model in noisy and reverberant scenarios, we train the proposed model using the noisy-reverb training set and treating reverberant speech as the training target for the speaker separation module. We evaluate the models on two test sets, namely ReverbTIMIT-noisy-a described in Section 3 and Libri-reverb-noisy created from Libri-reverb. For the latter, we consider only same-gender and different-gender mixtures with a mixing ratio of 0 dB and add factory noise from the NOISEX92 dataset (Varga and Steeneken, 1993). The noise is added such that the SNR of the first speaker and the noise is set to −5 dB, 0 dB, or 5 dB. Table 5 presents the evaluation results in noisy-reverberant scenarios. For the speaker-to-interference ratio, interference refers to the sum of the interfering talker and noise. This ratio offers a more accurate characterization of the acoustic environment. Compared to the reverb-only scenario, which corresponds to the 0 dB columns in Table 4, we observe similar levels of increase in total error rates for same-gender mixtures on both test sets. For example, the error rate increases from 12.55% to 18.01% for Libri-reverb and from 23.36% to 28.36% for ReverbTIMIT-a when the SNR is 5 dB. Furthermore, the total error rate increases significantly as the SNR decreases, showing the sensitivity of the pitch tracking module to the performance of the speaker separation module. For example, in the different-gender scenario of the ReverbTIMIT-noisy-a test set, the total error rate increases from 28.29% to 51.92% as the SNR decreases from 5 dB to −5 dB.

4.5. Analysis of using reverberant speech as reference

Should the pitch associated with vocal fold vibration always be regarded as the ground truth for pitch tracking in reverberant speech? While it is straightforward to define the ground-truth pitch in clean or anechoic speech, it is not the same for reverberant speech. This question has been discussed in Jin and Wang (2010). In reverberant environments, room reflections alter the relationship between the excitation signal and the received speech signal. As explained in the image source model (Allen and Berkley, 1979), room reverberation corresponds to an infinite number of image sources created by the reflections of the original source by room walls and surfaces. The resulting signal is an aggregate of these delayed and attenuated sources, which may not accord with the vocal fold characteristics of the original speech. It is observed that pitch in reverberant speech is elongated and modified compared to that in clean or anechoic speech.

Many studies (Unoki et al., 2004; Prasanna and Yegnanarayana, 2004; Flego and Omologo, 2006) aim to remove the effects of reverberation, recover the original speech signal, and extract glottal information. Such a definition of ground truth prioritizes the original content of the speech, often regarded as the default, and is employed for the evaluation in Section 4.4. In contrast, other studies such as Jin and Wang (2010) consider the pitch of the reverberant signal itself as the ground truth, emphasizing the actual periodicity of the received signal, aligning more closely with human perception. Given that both definitions have their use cases, we also address the second definition in this section, based on the periodicity of reverberant speech.

Table 6
Evaluation of pitch tracking algorithms on the reverberant utterances provided in Jin and Wang (2010). VDE (%) represents voicing decision error, as defined in Zhang et al. (2024), while DR (%) denotes detection rate, as defined in Zhang et al. (2022).

Algorithm	VDE ↓	DR ↑
RAPT (Talkin and Kleijn, 1995)	15.25	78.50
SWIPE (Camacho and Harris, 2008)	22.28	69.90
PRAAT (Boersma, 2001)	10.09	86.70
YIN (De Cheveigné and Kawahara, 2002)	19.37	33.44
DIO (Morise et al., 2009)	17.64	75.71
PENN (Morrison et al., 2023)	19.44	67.63
DC-CRN (Zhang et al., 2023) & RVD (Zhang et al., 2024)[a]	**8.54**	85.38
Proposed	**8.54**	**89.66**

[a] RVD requires laryngograph data during training, which may not always be available for standard speech data.

Table 7
Evaluation of pitch tracking algorithms on the Libri-manual test set.

Algorithm	VDE ↓	DR ↑
RAPT	12.50	78.02
SWIPE	18.11	73.95
PRAAT	12.65	82.97
YIN	41.33	36.42
DIO	23.73	63.09
PENN	18.22	66.60
DC-CRN & RVD	**8.05**	81.88
Proposed	**8.05**	**86.68**

4.5.1. Training label extraction

When reverberant speech pitch is treated as the reference, our proposed training label extraction method for clean or anechoic speech becomes less effective due to a mismatch when applied to reverberant conditions. To produce pitch labels for reverberant speech, we explore alternative methods that combine the strengths of conventional and data-driven pitch tracking methods. Specifically, we examine five conventional and two data-driven pitch tracking algorithms, as summarized in Table 6. The evaluation is conducted on the simulated reverberant speech samples used to create the evaluation corpus in Jin and Wang (2010), consisting of 105 samples in total. These samples are derived from 15 utterances from the TIMIT dataset (Garofolo et al., 1993), with each utterance simulated under seven different conditions, including one anechoic setup and six reverberation setups. The reverberation setups involve two acoustic rooms with T60 of 0.3 and 0.6 s, each with three configurations detailed in Jin and Wang (2010). Reference pitch contours are obtained through an interactive PDA (McGonegal et al., 1975; Jin and Wang, 2010), which combines automatic pitch determination and human correction. All simulated utterances are downsampled to 8 kHz to be compatible with the model input.

From Table 6, we observe that PRAAT achieves the highest detection rate, while the RVD model trained on laryngograph data achieves the lowest voicing decision error. To further enhance the utility of these algorithms, we propose a label generation method that combines the strengths of the best-performing algorithms. More specifically, we use RVD for voicing detection. In estimated voiced frames, we adopt PRAAT for pitch estimation if the frames are also considered voiced by PRAAT, or DC-CRN if PRAAT fails to detect voicing. The proposed method achieves the highest detection rate which is 2.96% higher than that of PRAAT.

4.5.2. Libri-manual test set

To further expand the test corpus, in addition to the manually labeled evaluation set provided in Jin and Wang (2010), we create another test set Libri-manual, which is derived from the LibriSpeech test-clean subset, with pitch manually labeled using the interactive PDA. The created Libri-manual dataset comprises 140 utterances, from 10 male and 10 female utterances randomly selected from the LibriSpeech test-clean subset. Each selected utterance is simulated under seven different configurations, following the setup outlined in Table A.II of Jin and Wang (2010). Table 7 presents the evaluation results for all pitch tracking algorithms listed in Table 6, tested on the Libri-manual test set. The performance trend closely aligns with the results in Table 6, with the proposed label generation method consistently showing the best performance. This consistency also shows the reliability of created pitch labels. As a result, we employ the proposed label generation method to produce ground-truth pitch labels for the training data and incorporate Libri-manual for model evaluation.

4.5.3. Evaluation in reverberant conditions

For training, we use the reverb-only training set described in Section 4.4. Since we treat reverberant speech pitch as the reference, we employ the proposed label generation method to create training pitch labels. Additionally, the training target for the separation module is set to reverberant speech rather than anechoic speech. To evaluate the effectiveness of the proposed training label

Table 8
Total error rate (%) of multi-pitch tracking models, with pitch labels from reverberant speech as ground truth.

Test set	Training label	Same gender				Different gender			
		0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
Libri-manual	PRAAT	34.59	34.75	35.45	36.55	32.90	33.19	33.75	34.97
Libri-manual	Proposed	**29.43**	**29.54**	**30.21**	**31.60**	**28.15**	**28.38**	**28.99**	**30.26**
ReverbTIMIT-r	PRAAT	30.61	30.40	30.79	32.04	27.85	28.05	28.67	29.96
ReverbTIMIT-r	Proposed	**27.43**	**27.32**	**27.72**	**29.28**	**23.88**	**24.12**	**24.99**	**26.26**

Table 9
Total error rate (%) of the proposed model trained under noisy-reverberant conditions considering reverberant speech as the reference, evaluated on different test sets. {}/{} dB indicates the SNR between the first speaker and noise, and the speaker-to-interference ratio.

Test set	Same gender			Different gender		
	5/−1.19	0/−3.01	−5/−6.19	5/−1.19	0/−3.01	−5/−6.19
Libri-manual-noisy	36.53	41.21	54.11	34.69	38.86	51.88
ReverbTIMIT-noisy-r	32.52	41.38	61.96	31.80	40.90	59.11

Table 10
Total error rate (%) defined in Wu et al. (2003) of multi-pitch tracking models evaluated on ReverbTIMIT-noisy-r test set under 5/−1.19 dB SNR/SIR.

Test set	Same gender	Different gender
Jin and Wang (2010)	76.37	62.37
Proposed	**26.50**	**26.92**

generation method, we also train the proposed model using the same training data but with PRAAT for label generation. Libri-manual and ReverbTIMIT-r test sets are used for evaluation.

The evaluation results are shown in Table 8. We observe that, on both test sets, employing the proposed label generation method for ground-truth label extraction yields improvements compared to using PRAAT. For example, when evaluated on the Libri-manual dataset, the proposed model demonstrates a decrease in the average total error rate from 34.52% to 29.57%. Furthermore, when evaluated on the ReverbTIMIT-r dataset, the total error rate improves from 29.80% to 26.38%, validating the effectiveness of the proposed label extraction method. We observe that the total error rates on Libri-manual are higher than those on ReverbTIMIT-r, which may be attributed to variations in manual labeling. As shown in Tables 6 and 7, both PRAAT and the proposed label extraction method align more closely with the labels in ReverbTIMIT-r.

Comparing Tables 8 and 4, we observe that the total error rates of the proposed model are significantly higher when using reverberant speech pitch as the reference compared to anechoic speech pitch. Several possible explanations may account for this observation. First, manual labeling errors are more likely to occur in the reverberant case due to the complex harmonic patterns in reverberant speech. Second, using reverberant speech as the training target may increase the difficulty of speaker separation. Finally, although the proposed training label extraction method improves the accuracy of training labels, it remains lower than that of the labels for anechoic speech.

4.5.4. Evaluation in noisy and reverberant conditions

We further train the proposed model using the reverb-noisy training set described in Section 4.4 for evaluation in noisy-reverberant environments. For testing, we use ReverbTIMIT-noisy-r and Libri-manual-noisy test sets. To construct Libri-manual-noisy, we follow a similar procedure to ReverbTIMIT-noisy-r, considering only same gender and different gender mixtures at a mixing ratio of 0 dB, and add the factory noise to speech mixtures such that the signal-to-noise ratio between the first speaker and the noise is at −5 dB, 0 dB, or 5 dB. The evaluation results are presented in Table 9. We observe that the performance drops significantly as noise levels increase. For instance, when tested on Libri-manual-noisy in the same-gender scenario, the total error rate increases from 36.53% to 54.11% as the SNR decreases from 5 dB to −5 dB. Additionally, we find that the total error rate on Libri-manual-noisy becomes lower than on ReverbTIMIT-noisy-r at lower SNR, possibly due to the shared corpus setup between Libri-manual-noisy and the training set.

Table 10 compares the proposed model with that of Jin and Wang (Jin and Wang, 2010), the only study addressing this task to our knowledge. This baseline also treats reverberant speech as the reference but performs multi-pitch tracking without speaker assignment. We thus use the evaluation metric defined in Wu et al. (2003) to assess the models without considering speaker assignment errors. The total error rate reported in Table 10 is the summation of E_{ij}, E_{gross}, and E_{fine} defined in Wu et al. (2003), which differs from the total error rate defined in Eq. (11) that includes speaker assignment errors. We evaluate on ReverbTIMIT-noisy-r at a 5 dB SNR between the first speaker and the noise. The results show that the proposed model significantly outperforms Jin and Wang in both same-gender and different-gender scenarios by large margins; the total error rate is reduced by 49.87% in the same-gender scenario, and by 35.45% in the different-gender scenario.

Similar to our observation in reverberant-only conditions, the total error rates in Table 9 are significantly higher in all scenarios when compared to Table 5, where anechoic speech is used as the reference. Given these observations, which is the better training target for multi-talker pitch tracking in reverberant conditions, anechoic or reverberant speech? Using anechoic speech pitch as the reference aligns more closely with speech-production perspectives, making it better suited for scenarios that prioritize the content of speech. In contrast, reverberant speech pitch as the reference reflects a speech-perception perspective, such as listening to a performance in a concert hall, but is relevant to far fewer current speech processing applications. Considering pitch tracking performance, the range of applications, and the effort of obtaining ground-truth labels, we advocate anechoic speech pitch as the preferred choice, different from reverberant speech pitch suggested by Jin and Wang (2010).

5. Concluding remarks

This study investigates talker-independent multi-pitch tracking in noisy and reverberant environments. We find that while synthetic data is advantageous for single-talker pitch tracking, it is not directly applicable to multi-talker scenarios due to the poor generalization of speaker separation models trained on synthetic mixtures. Instead, we utilize a single-talker pitch tracking model trained on synthetic speech and a voicing detection model trained on laryngograph data to enhance label accuracy. In terms of training strategy, we find that sequential training provides the best results for multi-pitch tracking. Moreover, fine-tuning with laryngograph-based data leads to noticeable improvements in both voicing detection and pitch estimation accuracy, demonstrating the critical importance of high-quality ground-truth labels. Considering reverberant speech pitch as the reference, we propose a training label generation method by leveraging multiple pitch estimators, and we create a test set with manually-corrected pitch contours for evaluation purposes. Our proposed multi-pitch tracking model achieves substantial improvements over baseline models. Evaluations and comparisons suggest that anechoic speech pitch should be the preferred training target for pitch tracking in reverberant conditions, which also aligns with broader applications. It is worth noting that reverberant speech pitch remains a useful reference for auditory perception research and applications where the focus is on the periodicity of the signal as perceived by the human ear. On the other hand, optimizing for anechoic pitch targets may be preferred in clinical settings, such as hearing-aid design, where detecting the fundamental frequency of anechoic speech is helpful for enhancing speech intelligibility.

Future work will focus on overcoming the significant generalization gap between synthetic and real-recorded speech data. Specifically, we intend to investigate the underlying reasons why synthetic data serves as an effective training source for single-talker pitch estimation yet fails to generalize to speaker separation in multi-talker scenarios. Developing domain-invariant training strategies may be crucial for improving model robustness in complex, real-world acoustic environments. Another area of future research is the development of causal, computationally efficient multi-pitch tracking algorithms for real-time applications (see Laube and Panos (2025)).

CRediT authorship contribution statement

Yixuan Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Heming Wang:** Software, Conceptualization. **DeLiang Wang:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yixuan Zhang reports financial support was provided by National Institute on Deafness and Other Communication Disorders. Yixuan Zhang reports financial support was provided by Ohio Supercomputer Center. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by an NIDCD grant (R01DC012048) and the Ohio Supercomputer Center. We thank Jing Chen, Li Nan, Yuzhou Liu for assisting us in replicating their models.

Data availability

The evaluation code and data are available at https://github.com/YIXUANZ/multipitch.

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65, 943–950.

Ardaillon, L., Roebel, A., 2019. Fully-convolutional network for pitch estimation of speech signals. In: Proc. Interspeech.

Bach, F.R., Jordan, M.I., 2005. Discriminative training of hidden Markov models for multiple pitch tracking. In: Proc. ICASSP. Vol. 5, pp. 489–492.

Bagshaw, P.C., Hiller, S.M., Jack, M.A., 1993. Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. In: Proc. Eurospeech. pp. 1003–1006.

Boersma, P., 2001. PRAAT, a system for doing phonetics by computer. Glot Int. 5, 341–345.

Busso, C., Lee, S., Narayanan, S., 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Lang. Process. 17, 582–596.

Camacho, A., Harris, J.G., 2008. SWIPE: A sawtooth waveform inspired pitch estimator for speech and music. J. Acoust. Soc. Am. 124, 1638–1652.

Chen, J., Wang, Y., Yoho, S.E., Wang, D.L., Healy, E.W., 2016. Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. J. Acoust. Soc. Am. 139, 2604–2612.

Christensen, M., Jakobsson, A., 2009. Multi-Pitch Estimation, B. H. Juang Morgan & Claypool, San Rafael, CA, USA.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks. In: Proc. ICML. pp. 933–941.

De Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111, 1917–1930.

Duan, Z., Han, J., Pardo, B., 2013. Multi-pitch streaming of harmonic sound mixtures. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 138–150.

Flego, F., Omologo, M., 2006. Robust F0 estimation based on a multichannel periodicity function for distant-talking speech. In: Proc. EU-USIP.

Gao, F., Wu, L., Zhao, L., Qin, T., Cheng, X., Liu, T.-Y., 2018. Efficient sequence learning with group recurrent networks. In: Proc. NAACL-HLT. pp. 799–808.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1993. DARPA TIMIT acoustic phonetic continuous speech corpus. CDROM.

Han, K., Wang, D.L., 2014. Neural network based pitch tracking in very noisy speech. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 2158–2168.

Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S., 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In: Proc. ICASSP. pp. 31–35.

Jin, Z., Wang, D.L., 2010. HMM-based multipitch tracking for noisy and reverberant speech. IEEE Trans. Audio Speech Lang. Process. 19, 1091–1102.

Kalkhorani, V.A., Wang, D.L., 2024. TF-CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single- and multi-channel speaker separation. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 4999–5009.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In: Proc. ICASSP. pp. 3933–3936.

Kim, J.W., Salamon, J., Li, P., Bello, J.P., 2018. CREPE: A convolutional representation for pitch estimation. In: Proc. ICASSP. pp. 161–165.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. In: Proc. ICLR.

Kolbæk, M., Yu, D., Tan, Z.-H., Jensen, J., 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 25, 1901–1913.

Kominek, J., Black, A.W., 2004. The CMU arctic speech databases. In: Proc. ISCA Workshop on Speech Synthesis. pp. 223–224.

Laube, S., Panos, B., 2025. COREPIT: A convolutional real-time pitch tracker. In: Proc. ISMIR.

Li, X., Liu, R., Song, T., Wu, X., Chen, J., 2020. Single-channel speech separation integrating pitch information based on a multi task learning framework. In: Proc. ICASSP. pp. 7279–7283.

Li, X., Sun, Y., Wu, X., Chen, J., 2022. Multi-speaker pitch tracking via embodied self-supervised learning. In: Proc. ICASSP. pp. 8257–8261.

Liu, Y., Wang, D.L., 2018. Permutation invariant training for speaker-independent multi-pitch tracking. In: Proc. ICASSP. pp. 5594–5598.

Liu, Y., Wang, D.L., 2019. Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation. IEEE/ACM Trans. Audio Speech Lang. Process. 27, 2092–2102.

Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In: Proc. ICASSP. pp. 46–50.

Luo, Y., Mesgarani, N., 2019. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 27, 1256–1266.

McGonegal, C., Rabiner, L., Rosenberg, A., 1975. A semiautomatic pitch detector (SAPD). IEEE Trans. Acoust. Speech Signal Process. 23, 570–574.

Morise, M., Kawahara, H., Katayose, H., 2009. Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. In: Proc. AES 35th International Conference.

Morise, M., Yokomori, F., Ozawa, K., 2016. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst. 99, 1877–1884.

Morrison, M., Hsieh, C., Pruyne, N., Pardo, B., 2023. Cross-domain neural pitch and periodicity estimation. arXiv preprint arXiv:2301.12258.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. LibriSpeech: An ASR corpus based on public domain audio books. In: Proc. ICASSP. pp. 5206–5210.

Pirker, G., Wohlmayr, M., Petrik, S., Pernkopf, F., 2011. A pitch tracking corpus with evaluation on multipitch tracking scenario. In: Proc. Interspeech.

Plante, F., Meyer, G., Ainsworth, W., 1995. A pitch extraction reference database. In: Proc. Eurospeech. Vol. 8, pp. 30–50.

Prasanna, S.R.M., Yegnanarayana, B., 2004. Extraction of pitch in adverse conditions. In: Proc. ICASSP. Vol. 1, pp. 109–112.

Stasiak, B., Rychlicki-Kicior, K., 2012. Fundamental frequency extraction in speech emotion recognition. In: Proc. MCSS. pp. 292–303.

Taherian, H., Kalkhorani, V.A., Pandey, A., Wong, D., Xu, B., Wang, D.L., 2024. Towards explainable monaural speaker separation with auditory-based training. In: Proc. Interspeech. pp. 572–576.

Takagi, M., Nishihara, M., Hono, Y., Hashimoto, K., Nankaku, Y., Tokuda, K., 2025. PeriodCodec: A pitch-controllable neural audio codec using periodic signals for singing voice synthesis. In: Proc. Interspeech. pp. 4913–4917.

Talkin, D., Kleijn, W.B., 1995. A robust algorithm for pitch tracking (RAPT). Speech Coding Synth. 495, 518.

Tzinis, E., Wang, Z., Smaragdis, P., 2020. Sudo rm-rf: Efficient networks for universal audio source separation. In: Proc. MLSP. pp. 1–6.

Unoki, M., Furukawa, M., Sakata, K., Akagi, M., 2004. An improved method based on the MTF concept for restoring the power envelope from a reverberant signal. Acoust. Sci. Technol. 25, 232–242.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12, 247–251.

Wang, Z.-Q., Cornell, S., Choi, S., Lee, Y., Kim, B.-Y., Watanabe, S., 2023. TF-GridNet: Integrating full-and sub-band modeling for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 3221–3236.

Wohlmayr, M., Stark, M., Pernkopf, F., 2010. A probabilistic interaction model for multipitch tracking with factorial hidden Markov models. IEEE/ACM Trans. Audio Speech Lang. Process. 19, 799–810.

Wu, M., Wang, D.L., Brown, G.J., 2003. A multipitch tracking algorithm for noisy speech. IEEE Trans. Speech Audio Process. 11, 229–241.

Zhang, Y., Wang, H., Wang, D.L., 2022. Densely-connected convolutional recurrent network for fundamental frequency estimation in noisy speech. In: Proc. Interspeech. pp. 401–405.

Zhang, Y., Wang, H., Wang, D.L., 2023. F0 estimation and voicing detection with cascade architecture in noisy speech. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 3760–3770.

Zhang, Y., Wang, H., Wang, D.L., 2024. Leveraging laryngograph data for robust voicing detection in speech. J. Acoust. Soc. Am. 156, 3502–3513.