Interspeech 2022
18-22 September 2022, Incheon, Korea

Densely-connected Convolutional Recurrent Network for Fundamental Frequency Estimation in Noisy Speech

Yixuan Zhang[1], *Heming Wang*[1], *DeLiang Wang*[1,2]

[1] Department of Computer Science and Engineering, The Ohio State University, USA
[2] Center for Cognitive and Brain Sciences, The Ohio State University, USA
{zhang.7388, wang.11401, wang.77}@osu.edu

Abstract

Estimating fundamental frequency ($F0$) from an audio signal is a necessary step in many tasks such as speech synthesis and speech analysis. Although high estimation accuracy has been achieved for clean speech, it is still challenging for $F0$ estimation to handle noisy speech, mainly because of the corruption of harmonic structure caused by noise. In this paper, we view $F0$ estimation as a multi-class classification problem and train a frequency-domain densely-connected convolutional neural network (DC-CRN) to estimate $F0$ from noisy speech. The proposed model significantly outperforms baseline methods in terms of detection rate. We find that using complex short-time Fourier transform (STFT) as input produces better performance compared to using magnitude STFT as input. Furthermore, we explore improving $F0$ estimation with speech enhancement. Although the $F0$ estimation model trained on clean speech performs well on enhanced speech, the distortion introduced by the speech enhancement model limits the estimation performance. We propose a cascade model which consists of two modules that optimize enhanced speech and estimated $F0$ in turn. Experimental results show that the cascade model brings further improvements to the DC-CRN model, especially in low signal-to-noise ratio (SNR) conditions.

Index Terms: Pitch tracking, densely-connected convolutional recurrent neural network, complex domain, cascade architecture

1. Introduction

Pitch tracking or pitch estimation is a crucial step in applications such as speech analysis and speech synthesis, and it refers to estimating the fundamental frequency ($F0$) of an audio signal.[1] While many pitch tracking methods can estimate the pitch of clean speech accurately, it is difficult to extract the correct pitch if the speech is severely interfered by noise since the harmonic structure and temporal continuities of the speech signal are corrupted.

Many signal processing algorithms are designed to estimate pitch from clean or noisy speech. In general, these algorithms can be categorized into time-domain, frequency-domain, and time-frequency domain methods. Time-domain methods such as YIN [1], PYIN [2] and RAPT [3] extract pitch by estimating the periodicity of the signal. For frequency domain methods such as SAFE [4] and PEFAC [5], the objective is to determine the fundamental frequency based on the harmonic patterns of the signal. Time-frequency domain methods such as Wu et al. [6] extract pitch by estimating the periodicity of subband signals in the time-frequency domain. To further improve pitch tracking results, post-processing methods, such as dynamic programming [5] and hidden Markov models [7, 6], are often applied to the above methods to leverage temporal continuity of pitch contours by producing a most probable pitch track from the frame-level pitch candidates.

In recent years, deep neural networks (DNNs) have been introduced to pitch tracking and achieved considerable improvements over signal processing algorithms. In the first such study, Han and Wang [8] investigated pitch state distribution modeling in the frequency domain with a feed-forward DNN and a recurrent neural network (RNN). The probabilistic pitch state outputs are connected to form the final pitch contours using Viterbi decoding. The models are trained with noisy speech and show robust pitch tracking performance in different noise conditions. Recently, time-domain pitch tracking methods such as CREPE [9] and FCN [10] have been proposed and produce state-of-the-art $F0$ estimation results. These methods take raw waveform as input and utilize convolutional neural network (CNN) models for pitch estimation. Both CREPE and FCN are trained with synthesized signals, which allows for complete control of ground truth $F0$.

In this paper, we propose to use a densely-connected convolutional recurrent neural network (DC-CRN) model for noisy speech pitch tracking. The DC-CRN architecture used in this study is designed based on the original DC-CRN model [11] proposed for speech enhancement. Our model extracts pitch from the frequency domain and incorporates the information of temporal dependency in pitch sequences with the help of an RNN. Experimental results show that our model significantly outperforms baseline methods. In addition, inspired by the success of speech enhancement in the complex domain [12, 13], we explore incorporating phase information in pitch tracking. We observe that, compared with using magnitude STFT as input, taking complex STFT as input brings consistent improvements.

Does speech enhancement help pitch tracking in noisy conditions? We observe that extracting pitch from enhanced speech using a model trained on clean speech functions reasonably well. However, the distortion introduced by the speech enhancement model places a limit on pitch tracking performance. To reduce the influence of such distortion in enhanced speech, we propose a cascade architecture that contains two modules. The first module takes noisy audio as input and focuses on speech enhancement. The second module takes the enhanced speech from the first module and noisy audio as input and generates pitch estimates. The two modules are jointly trained using

This work was supported in part by an NIDCD grant (R01DC012048) and the Ohio Supercomputer Center.

[1] The definitions of pitch and fundamental frequency are not identical. Pitch is defined as an auditory attribute of a sound, which is a perceptual measure. On the other hand, fundamental frequency is a physical property of an audio signal. However, these two terms are often used interchangeably since they are closely related. In this paper, we will use the two terms interchangeably for convenience.

10.21437/Interspeech.2022-11156

a loss function consisting of a pitch estimation objective and a speech enhancement objective. Experimental results show that the cascade architecture further improves the performance, especially in low SNR scenarios.

The rest of this paper is organized as follows. In the next section, we describe the details of our proposed method. The experiments and evaluation results are presented in Section 3. Section 4 concludes the paper.

2. Model Description

2.1. Speech Synthesis for Data Generation

For training purposes, the dataset should have enough utterances and provide reliable ground truth $F0$ labels. One way to obtain such datasets is by gathering datasets such as PTDB-TUG [14] and KEELE [15] which provide laryngograph recordings. A good estimate of ground truth $F0$ can be obtained by applying an $F0$ estimator on the laryngograph recordings. However, as mentioned in [10, 16], the pitch estimates from laryngograph data are not always reliable and may contain octave errors. Besides, only limited datasets provide laryngograph recordings, which makes it more difficult to build a large dataset for pitch tracking.

To have complete control on ground truth $F0$, methods such as [9, 10, 17] build their datasets with synthesized audios. The audios are re-synthesized from real recordings and can match the pitch contours provided for synthesis. Usually, the pitch contours estimated from the real recordings with pre-existing $F0$ estimators are used for synthesis. In this study, we employ a high-quality speech synthesizer WORLD [18] to re-synthesize audios for the whole database. WORLD estimates fundamental frequency, aperiodicity, and spectral envelope and synthesizes speech based on the estimated parameters. We substitute the pitch tracker in WORLD with torchcrepe [19], an implementation of CREPE [9] with pre-trained models and additional submodules for silence detection and filtering out unreliable pitch estimates, which we find can be tuned to detect the silence and unvoiced regions more precisely.

2.2. Problem Formulation

Similar to [8, 9, 10], we view pitch tracking as a classification problem. Following the setups in FCN [10], the outputs of our network are 486-dimensional vectors, where each dimension corresponds to a pitch class. A target frequency range of [30-1000] Hz is divided into 486 pitch classes $c_1, c_2, ..., c_{486}$ with a step size of 12.5 cents. It is assumed that all possible $F0$ of vocal sounds can be covered by the range, including corner cases such as high $F0$ from soprano singing and low $F0$ from the voice in fry mode [10]. Each training target vector y is generated based on ground truth $F0$. To reduce the penalty for near-correct estimation, the target vector is Gaussian-blurred with a 25 cents standard deviation, as shown in Eq. 1.

$$y_i = \exp[-\frac{(c_i - c_{true})^2}{2 \cdot 25^2}] \tag{1}$$

where y_i is the value of ith bin in the target vector, c_i represents the pitch class of ith bin in cents and c_{true} corresponds to ground truth $F0$ in cents.

To learn the probabilistic output, we train DNN by minimizing the binary cross-entropy loss between target vector y and estimated vector $\hat{y}$.

$$\mathcal{L}_{pitch}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} [-y_i \log \hat{y}_i - (1 - y_i) \log (1 - \hat{y}_i)] \tag{2}$$

where N is the number of pitch classes and is 486 here.

As shown in Eq. 3, to compute a pitch estimate, the pitch class x with the highest value is first picked. Then, the pitch estimate in cents $\hat{c}$ is obtained by calculating a weighted average of pitch classes near the picked pitch class x.

$$x = \arg\max_i \; \hat{y}_i, \qquad \hat{c} = \frac{\sum_{i=x-4}^{x+4} \hat{y}_i c_i}{\sum_{i=x-4}^{x+4} \hat{y}_i} \tag{3}$$

The pitch estimate is then converted from cents to Hz.

2.3. Densely-Connected Convolutional Recurrent Network

Although CNN-based methods [9, 10, 20] have achieved state-of-the-art performance in the clean condition, the CNN models utilize only local information for pitch tracking, thus cannot capture long-term temporal variations. However, time continuity is an essential characteristic of pitch contours in audios. In this study, we develop a DC-CRN model for pitch tracking, based on the original DC-CRN model [11] proposed for speech enhancement. This architecture contains a CNN followed by an RNN, which helps model the temporal continuity of pitch contours.

The diagram of the network architecture is shown in Fig. 1. The input is the complex STFT of the input mixture signal that has three dimensions: frequency, time, and channel. The real and imaginary components of the complex STFT are treated as two separate channels. The network consists of 7 convolutional densely-connected (DC) blocks followed by a two-layer bidirectional long short-term memory (BLSTM) block and a linear layer with Sigmoid activation.

Fig. 2a illustrates the architecture and the dense connectivity pattern of a DC block. For the first four layers, each of them contains a 2-dimensional convolutional layer followed by batch normalization and exponential linear unit (ELU) activation function. The last layer, as shown in Fig. 2b, is a gated convolutional layer that contains the gated linear units. For each layer, the input is a concatenation of the outputs from the preceding layers, allowing each layer to utilize the outputs from preceding layers, which improves the information flow between layers. For the two-layer BLSTM block, a grouping strategy [21] is applied to reduce the number of trainable parameters while not introducing much performance degradation. As illustrated in Fig. 3, in the first recurrent layer, the features and hidden states are split into disjoint groups. Intra-group features are learned within each group. To model inter-group dependency, the representations are rearranged between two recurrent layers. Layer normalization is applied after each recurrent layer.

2.4. Network Configurations

The proposed model consists of 7 DC blocks and a two-layer BLSTM block. The first four layers in each DC block are layers with kernels of size 1 × 3 (times × frequency) and 8 output channels. Zero paddings are applied along the frequency dimension with a size of 1. The last layer in each DC block has a kernel size of 1 × 4, a stride of 2, and zero paddings of size 1 applied along the frequency dimension. The seven DC blocks have 4, 8, 16, 32, 64, 128, 256 output channels respectively. For the BLSTM block, the grouping strategy is applied to reduce the model size. The group number is set to 4.

2.5. Cascade Architecture

In this study, we investigate improving pitch tracking in noisy environments by incorporating speech enhancement. Inspired

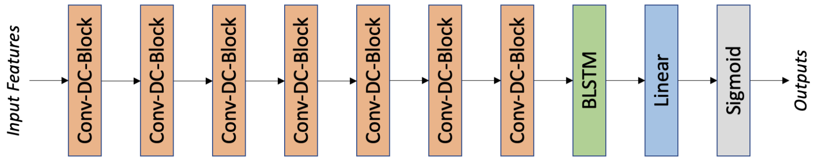
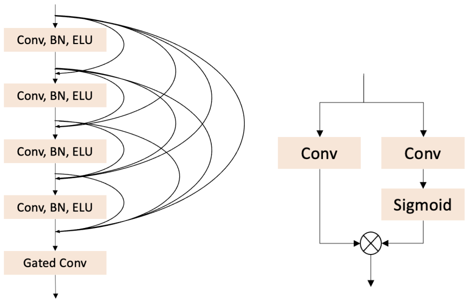


Figure 1: *Network architecture of DC-CRN pitch tracking model*



(a) (b)

Figure 2: *Diagrams of (a) densely-connected block and (b) gated convolution.* $\otimes$ *represents element-wise multiplication*

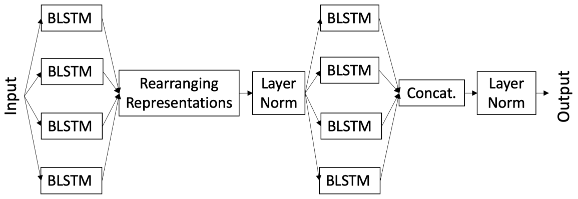


Figure 3: *Group strategy for two-layer BLSTM*

by a recent study [22], a cascade architecture is developed for this purpose. As shown in Fig. 4, the model consists of a speech enhancement module and a pitch tracking module. The complex STFT of the noisy input is fed into the speech enhancement module first. The speech enhancement module produces a complex STFT estimate of the clean speech, which is concatenated with the complex STFT of the noisy input and then fed into the pitch tracking module. The pitch tracking module produces pitch estimates. The two modules are trained jointly by minimizing the loss function that is defined as,

$$\mathcal{L} = \alpha\mathcal{L}_{enh} + \mathcal{L}_{pitch} \tag{4}$$

$$\mathcal{L}_{enh} = \frac{1}{2TF}\sum_{t,f}[(\hat{S}_r(t,f) - S_r(t,f))^2 + (\hat{S}_i(t,f) - S_i(t,f))^2] \tag{5}$$

where the loss for pitch tracking L_{pitch} is the binary entropy loss defined in Eq. 2. The enhancement module directly learns the real and imaginary spectra of clean speech ($S_r(t,f)$, $S_i(t,f)$) by minimizing L_{enh} (Eq. 5). T, F denotes the number of time frames and frequency bins. Coefficient α is set to 0.01 to balance the value ranges of L_{pitch} and L_{enh}.

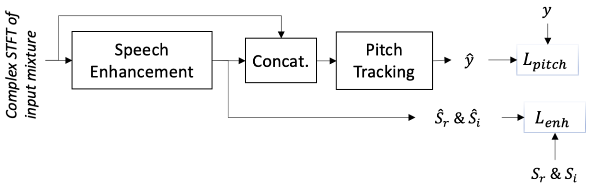


Figure 4: *Cascade architecture*

3. Evaluation Results and Comparisons

3.1. Experimental Setup

Data preparation

The synthetic dataset for training is created from the original audios selected from the LibriSpeech [23] 360 hours training corpus, which has clean speech audios from 921 speakers, where 439 speakers are female and 482 speakers are male. To build the training set, 4152 utterances are randomly picked from the LibriSpeech 360 hours training corpus. All utterances with audio lengths greater than 6s are cut to 6s long. Then, as described in Section 2.1, we use torchcrepe [19] to estimate the pitch contours for all utterances. We set less reliable pitch estimates and $F0$ in the silence region to 0 with the help of sub-modules in torchcrepe. With the estimated pitch contours, each utterance is re-synthesized using the WORLD [18] speech synthesizer. The dataset is further augmented by re-synthesizing each utterance with pitch contours that are an octave lower and one octave higher compared with the original pitch contours. Synthesized audios that contain $F0$ out of the target range are removed from the dataset. All synthesized audios are re-sampled from 16 kHz to 8 kHz.

To generate noisy mixtures for the training set, we use the 10,000 noises with a total duration of 126 hours, from a sound effect library[2]. Each clean utterance from the synthetic training set is mixed with a random segment from the 10,000 noises with a signal-to-noise ratio (SNR) randomly chosen from {-5, -4, -3, -2, -1, 0} dB.

To avoid the possible bias brought by the synthesizer, we build the validation and test set on real recordings. Since octave errors are observed in the ground truth $F0$ obtained from laryngograph data, we adopt consensus ground truth $F0$ from [16], which derives ground truth $F0$ by looking for the consensus from state-of-the-art fundamental frequency estimation algorithms. It is observed that the provided consensus ground truth is broadly compatible with laryngograph-based ground truth and more representative in edge cases. For datasets, we choose Mocha-TIMIT [24] for validation and the FDA database [25] for testing. The utterances from the validation set are mixed

[2] https://www.soundideas.com

Table 1: $F0$ *Detection Rates (in %) of proposed and baseline methods*

Method	Babble				Factory				Cafeteria			
	-10 dB	-5 dB	0 dB	5 dB	-10 dB	-5 dB	0 dB	5 dB	-10 dB	-5 dB	0 dB	5 dB
PEFAC	34.44%	55.00%	71.60%	81.14%	59.32%	71.05%	80.21%	85.16%	32.62%	54.85%	70.27%	79.27%
Han and Wang RNN	25.54%	57.09%	82.17%	92.27%	67.08%	84.81%	92.66%	95.41%	20.97%	59.06%	81.95%	93.17%
FCN-noisy	64.64%	83.32%	92.04%	96.65%	80.05%	91.03%	96.11%	97.78%	57.23%	79.34%	90.69%	96.70%
DC-CRN-mag	71.51%	88.05%	94.86%	97.07%	88.49%	95.25%	97.54%	97.92%	71.64%	88.00%	95.38%	97.76%
DC-CRN-complex	75.11%	89.64%	95.38%	97.92%	89.09%	95.95%	97.67%	98.40%	72.40%	88.84%	96.38%	97.89%

Table 2: $F0$ *Detection Rates (in %) of separate training and cascade architecture*

Method	Babble				Factory				Cafeteria			
	-10 dB	-5 dB	0 dB	5 dB	-10 dB	-5 dB	0 dB	5 dB	-10 dB	-5 dB	0 dB	5 dB
Separate Training	67.52%	83.35%	91.66%	95.22%	80.72%	91.20%	95.54%	96.87%	57.83%	79.89%	91.34%	95.52%
Cascade Architecture	77.83%	90.66%	96.26%	98.01%	90.43%	96.41%	98.09%	98.27%	76.41%	90.15%	96.96%	98.26%

with cafeteria noise from an Auditec CD[3] at -5 dB SNR. For the test set, babble noise, factory noise from NOISEX92 [26] and the cafeteria noise are used for creating noisy mixtures. Four SNRs {-10, -5, 0, and 5} dB are considered for testing. We use a Hamming window of 128 ms duration with a 10 ms hop size for STFT computation.

Comparison baselines

We compare our methods with three pitch tracking methods: PEFAC [5], Han and Wang's RNN model [8] and FCN model [10]. PEFAC [5] is a signal processing algorithm that performs relatively well in low SNR conditions. The method attenuates narrow-band noise and smoothly varying noise components by combining non-linear amplitude compression and applying comb-filter while estimating $F0$. Han and Wang's RNN model [8] is a speaker-independent model which learns probabilistic pitch states from noisy speech data and produces the pitch contour by applying Viterbi decoding. FCN [10] is an end-to-end system that takes raw waveform as input, which achieves state-of-the-art pitch tracking performance on clean speech. We retrain Han and Wang's RNN model (Han and Wang RNN) and the FCN model (FCN-noisy) with our datasets for a fair comparison.

Evaluation metric

Models are evaluated in terms of detection rate (DR) as defined in [27]. Detection rate is calculated on voiced frames. An estimated F0 is considered as correct if the estimated $F0$ differs from ground truth $F0$ by less than 5%.

$$DR = \frac{N_{0.05}}{N_p} \tag{6}$$

$N_{0.05}$ is the number of frames whose estimated $F0$ has a deviation of less than 5% of ground truth $F0$. N_p is the number of voiced frames.

3.2. Results and Comparisons

We compare the proposed DC-CRN model with baseline methods on the FDA dataset in Table 1. We observe that FCN-noisy and DC-CRN models significantly outperform PEFAC and Han and Wang RNN, and the proposed DC-CRN-complex model yields the best results in all SNR conditions. FCN-noisy has strong performance in less noisy conditions but DC-CRN models substantially outperform FCN-noisy in the low SNR scenarios, with much fewer parameters (4.1 Million compared to 12.3 Million). For example, under -10 dB SNR, the detection rate is improved by 11.56% on average. In addition, we examine different input types on the DC-CRN model. We find that the model using complex STFT input (DC-CRN-complex) consistently outperforms the model trained with magnitude STFT input (DC-CRN-mag).

In Table 2, we explore different ways of integrating speech enhancement into pitch tracking. We first train a speech enhancement model and a clean pitch tracker separately and evaluate the pitch tracker on enhanced speech. A DC-CRN [11] model for speech enhancement is trained with our training set. We also train a DC-CRN-complex pitch tracking model for clean speech, with the clean synthetic training set before mixing. For testing, the noisy input is enhanced by the speech enhancement model. The enhanced speech is then used as input to the pitch tracker. From Table 2, it is observed that the pitch tracker trained on clean speech performs reasonably well on enhanced speech. But the distortion introduced by speech enhancement seems to limit the pitch tracking performance. On the other hand, in the proposed cascade architecture in Section 2.5, the speech enhancement module and the pitch tracking module are jointly optimized by a composite loss function. Compared with separate training, the pitch tracker in the cascade architecture can adapt to such distortions. As shown in Table 2, the cascade architecture substantially outperforms separate training. In addition, compared with the DC-CRN-complex model in Table 1, the cascade architecture brings further improvements, especially in low SNR conditions. For example, in -10 dB SNR scenarios, the detection rate is improved by 2.69% on average.

4. Conclusion

In this study, we investigate pitch tracking for noisy speech with a focus on speaker-independent and noise-independent scenarios. We perform pitch tracking in the frequency domain and treat pitch tracking as a multi-class classification problem. The proposed DC-CRN model significantly outperforms baseline methods. It is found that, as the input, complex STFT is preferable to magnitude STFT. In addition, we notice that the distortion in enhanced speech makes it suboptimal for a pitch tracker trained on clean speech to estimate F0. A cascade architecture is then proposed which integrates speech enhancement into pitch tracking. We demonstrate that the cascade architecture reduces the effects of distortion introduced by speech enhancement and brings further improvements to pitch estimation results. Future work will explore voicing detection and pitch tracking in multi-talker speech mixtures.

[3] https://auditec.com/

5. References

[1] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

[2] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. ICASSP*, 2014, pp. 659–663.

[3] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, vol. 495, p. 518, 1995.

[4] W. Chu and A. Alwan, "SAFE: A statistical approach to F0 estimation under clean and noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 933–944, 2011.

[5] S. Gonzalez and M. Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 518–530, 2014.

[6] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 229–241, 2003.

[7] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1091–1102, 2010.

[8] K. Han and D. L. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, pp. 2158–2168, 2014.

[9] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. ICASSP*, 2018, pp. 161–165.

[10] L. Ardaillon and A. Roebel, "Fully-convolutional network for pitch estimation of speech signals," in *Proc. Interspeech*, 2019.

[11] K. Tan, X. Zhang, and D. L. Wang, "Deep learning based real-time speech enhancement for dual-microphone mobile phones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1853–1863, 2021.

[12] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language processing*, vol. 24, pp. 483–492, 2015.

[13] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[14] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[15] F. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," *Proc. Eurospeech*, vol. 8, pp. 30–50, 1995.

[16] B. Bechtold, "Pitch of voiced speech in the short-time fourier transform: Algorithms, ground truths, and evaluation methods," Ph.D. dissertation, Carl von Ossietzky Universität Oldenburg, 2021.

[17] J. Salamon, R. M. Bittner, J. Bonada, J. J. Bosch, E. Gómez Gutiérrez, and J. P. Bello, "An analysis/synthesis framework for automatic F0 annotation of multitrack datasets," in *Proc. ISMIR*, 2017.

[18] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, pp. 1877–1884, 2016.

[19] M. Morrison, "torchcrepe," 2020. [Online]. Available: https://github.com/maxrmorrison/torchcrepe

[20] S. Singh, R. Wang, and Y. Qiu, "DEEPF0: end-to-end fundamental frequency estimation for music and speech signals," in *Proc. ICASSP*, 2021, pp. 61–65.

[21] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 799–808.

[22] H. Wang and D. L. Wang, "Neural cascade architecture with triple-domain loss for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 734–743, 2021.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[24] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

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

[26] A. Varga and H. J. Steeneken, "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 Communication*, vol. 12, pp. 247–251, 1993.

[27] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 2067–2079, 2010.