INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

MF-PAM: Accurate Pitch Estimation through Periodicity Analysis and Multi-level Feature Fusion

Woo-Jin Chung[1], *Doyeon Kim*[1] *Soo-Whan Chung*[2] *Hong-Goo Kang*[1]

[1]Dept. of Electrical & Electronic Engineering, Yonsei University, South Korea
[2]NAVER Cloud, South Korea

woojinchung@dsp.yonsei.ac.kr ehyeon24@dsp.yonsei.ac.kr, soowhan.chung@navercorp.com, hgkang@yonsei.ac.kr

Abstract

We introduce Multi-level feature Fusion-based Periodicity Analysis Model (MF-PAM), a novel deep learning-based pitch estimation model that accurately estimates pitch trajectory in noisy and reverberant acoustic environments. Our model leverages the periodic characteristics of audio signals and involves two key steps: extracting pitch periodicity using periodic non-periodic convolution (PNP-Conv) blocks and estimating pitch by aggregating multi-level features using a modified bi-directional feature pyramid network (BiFPN). We evaluate our model on speech and music datasets and achieve superior pitch estimation performance compared to state-of-the-art baselines while using fewer model parameters. Our model achieves 99.20 % accuracy in pitch estimation on a clean musical dataset. Overall, our proposed model provides a promising solution for accurate pitch estimation in challenging acoustic environments and has potential applications in audio signal processing.
Index Terms: Neural pitch estimation, multi-level fusion

1. Introduction

Voice is an intrinsic attribute of humans and depends on the physiological articulatory anatomy of each person. When producing speech, an intricate combination of organs, from the lungs to the mouth and throughout the vocal tract, collaborates to produce an individual's unique voice. In particular, pitch or fundamental frequency is widely regarded as a prominent characteristic of a speaker's voice among other acoustic features, and it is essential for various speech-oriented tasks such as speech enhancement [1, 2], speech separation [3, 4], speech synthesis [5, 6], and speaker verification [7, 8].

Previously, stochastic approaches such as normalized autocorrelation or zero-crossing intervals were primarily used to estimate the periodicity of speech in the time domain [9–11]. Other techniques such as difference functions and spectral analysis have been used to explore the harmonicity of signals in the frequency domain [12]. pYIN [13] used the local minima of the cumulative mean normalized difference function and hidden Markov models for probabilistic modification, whereas SWIPE [14] estimated the pitch in the frequency domain using the sawtooth waveform spectrum. Hybrid approaches that analyzed both the time and frequency domains propose for more stable performance [15, 16]. However, despite their lightweight approach, these stochastic methods have shown unstable pitch estimation performances due to various limitations. Typically, an accurate estimation of pitch is challenging due to its dependence on multiple factors including intonation, emotion, and even physiological factors that may vary over time. Moreover, pitch estimation in observed speech remains a challenging task due to the potential distortions caused by environmental factors.

Deep learning techniques in speech processing have significantly improved the performance of pitch estimation. In [17], the authors have proposed effective estimation networks based on the sequential modeling of neural networks. CREPE [18] leveraged convolution neural networks (CNNs) considering the noise distortion, while DeepF0 [19] utilized a dilated causal convolution network for large receptive field observation. Both methods have proved that neural networks are effectively used to analyze the acoustic characteristics of speech signals. More recent studies have focused on the development of models that consider acoustic characteristics rather than relying solely on neural networks. SPICE [20] analyzed the pitch shift mapped by constant-Q transform (CQT), and HarmoF0 [21] captured the harmonic structure closely related to pitch from a log-spectrogram using multiple rates dilated causal convolution. These studies demonstrated that considering the acoustic characteristics of speech signals improves the performance of pitch estimation.

In this paper, we propose a novel lightweight pitch estimation model, Multi-level feature Fusion-based Periodicity Analysis Model (MF-PAM), which operates on the raw audio waveform. The proposed model is composed of two stages: analysis and estimation. During the analysis stage, MF-PAM extracts the periodicity of the feature maps from the input speech utilizing two submodules. The low-level submodule distinguishes the periodic and non-periodic characteristics by using periodic and non-periodic convolution (PNP-Conv) blocks. The PNP-Conv blocks analyze the input with a dual-path convolution layer using a snake function [22], which is sensitive to periodic representations. The high-level submodule employs periodic convolution (P-Conv) blocks to further extract the periodic components, while its following long short-term memory (LSTM) layer enables sequential modeling of the extracted periodic features. In the estimation stage, we utilize a modified bi-directional feature pyramid network (BiFPN) to aggregate the multi-level features extracted in the analysis stage. The multi-level feature fusion provides an accurate and effective pitch estimation by referencing various latent representations from each layer. The designed neural network is optimized to pitch tracking task with only 0.362M parameters, indicating the effectiveness of the proposed composite modules and their ability to perform well without relying on high computational power or a large number of model parameters. Our experiments on various datasets and ablation studies demonstrate the effectiveness of MF-PAM compared to that of the baselines and highlight the importance of the submodules in extracting pitch components in various environments. In addition, our lightweight model, MF-PAM-S, exhibits notable pitch estimation accuracy despite having only 0.213M parameters which is equivalent to 59% of the smallest baseline model.

10.21437/Interspeech.2023-2487

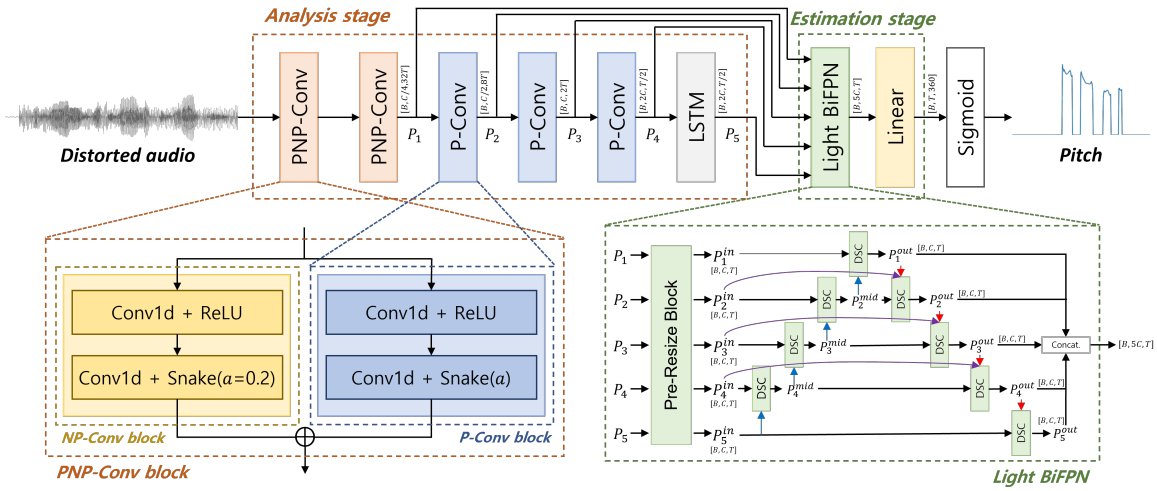

Figure 1: *Illustration of the proposed model, MF-PAM.*

2. Related works

2.1. Snake function

In [22], the authors investigated the extrapolation properties of activation functions and proposed an effective activation function sensitive to periodicity, the Snake function. Neural networks that use the Snake function demonstrate impressive results in solving periodic problems such as atmospheric prediction, body temperature prediction, and financial data prediction. In addition, the Snake activation function was found to be advantageous in the optimization of model training compared to other periodic baselines. The Snake function is formulated as follows:

$$Snake_a(x) = x + \frac{1}{a} sin^2(ax), \tag{1}$$

where a denotes a pre-fixed constant that affects the frequency range the function focuses on. The authors reported that larger values of a were more effective for input with periodic characteristics, while smaller values of a were appropriate for non-periodic or standard tasks such as image classification.

2.2. BiFPN

BiFPN [23] is one of the multi-scale fusion models [24, 25] devised for object recognition. BiFPN is a feature pyramid network that improves the target task with top-down and bottom-up pathways and cross-connections. It matches the channel difference using convolution layers and the resolution difference by re-sampling. Since different resolution features contribute unequally to the output, BiFPN provides a fast normalize fusion method with learnable weights. With the lower network complexity and time cost, BiFPN achieve better object detection performance than that of the previous methods.

3. Proposed model

The speech signal exhibits a quasi-stationary characteristic, with its periodicity primarily arises from the pitch. Therefore, our proposed model, MF-PAM, is designed to be sensitive to the periodicity of speech signals, which is beneficial for accurate pitch estimation. MF-PAM emphasizes periodic characteristics in the latent representation and estimates pitch trajectory based on the representations. The overall structure, as shown Figure 1, comprises two stages: analysis and estimation. In the analysis stage, MF-PAM extracts periodic information by eliminating non-periodic information in low-level representations. The estimation stage tracks the pitch trajectory by leveraging the representation obtained from the analysis stage using a BiFPN module optimized for pitch estimation.

3.1. Analysis stage

In the analysis stage, MF-PAM analyzes speech signals by leveraging periodicity-sensitive modules, which include periodic convolution (P-Conv), non-periodic convolution (NP-Conv), and periodic and non-periodic convolution (PNP-Conv) blocks. Our analysis structure is designed to first eliminate non-periodic information from the input and then enforce periodic characteristics. In particular, the structure comprises two PNP-Conv blocks followed by three P-Conv blocks and an LSTM layer. The PNP-Conv block consists of a dual-path convolution block, where one path is a P-Conv block, and the other is a NP-Conv block. Both modules have two convolution layers in a stack, activated by the ReLU function and rectified by the Snake function. The main difference between the P-Conv and NP-Conv block is the parameter a in Eq. (1), which controls the frequency range of the periodicity. Referring the findings in [22], the Snake function effectively processes periodic information with large a values (5-50), while small a (0.2-0.5) is suitable for processing non-periodic characteristics. Therefore, we set a as 0.2 for all NP-Conv blocks to eliminate non-periodic components, while a of P-Conv blocks are set differently with larger values. As the receptive field size of each layer increases, it becomes capable of capturing a larger range of temporal information. Based on our preliminary experiments, we gradually reduced the values of a in higher-level P-Conv blocks that have larger receptive fields, to (17, 13, 11, 7, 5) in order to increase sensitivity to the low-frequency range. Subsequently, an LSTM layer enables powerful sequential modeling for pitch tracking.

3.2. Estimation stage

In this work, we transform the pitch estimation problem into a classification task similar to [21], which estimates the level at which the pitch exists among 360 quantized levels of a limited frequency range. Therefore, in the estimation stage, MF-PAM estimates the discrete pitch frequency quantized in logarithmic scale by aggregating the multi-level features from the analysis stage using a modified BiFPN module. In Figure 1, there is an overall structure of the BiFPN optimized for MF-

PAM, which has half number of channels compared to the vanilla BiFPN. The pre-sizing block aims to adjust the temporal resolution of the multi-level features to half that of the third-level feature (P_3), while maintaining the channel size. This is achieved through up- and down-sampling and using a depthwise separable convolution layer. The resized five multi-level features are fused as below:

$$P_i^{mid} = DSC\left(\frac{w_1 \cdot P_i^{in} + w_2 \cdot P_{i+1}^{in}}{w_1 + w_2 + \epsilon}\right), \quad (2)$$

$$P_i^{out} = DSC\left(\frac{w_1' \cdot P_i^{in} + w_2' \cdot P_i^{mid} + w_3' \cdot P_{i-1}^{out}}{w_1' + w_2' + w_3' + \epsilon}\right), \quad (3)$$

where P_i and w_i denote the i-th level feature and its learnable weight factor, respectively. DSC indicates the depthwise separable convolution layer activated by a Swish function [26]. The ϵ is set as 1e-4.

The BiFPN output is projected onto 360-dimensional quantized frequency bins using a projection layer, which has a fully-connected layer followed by a sigmoid function. There is a 25-cent interval between consecutive quantization levels, and the frequency range is from 32.7Hz to 5834.5Hz. The level can be converted to the frequency in Hertz using $f(i) = 32.7 \times 2^{25i/1200}\ [Hz]$, where i denotes the index of the level.

3.3. Training criteria

The entire model is trained in an end-to-end manner, and we follow a similar training criterion as in [21], which involves minimizing the binary cross-entropy loss between the target one-hot vector y and the predicted output vector $\hat{y}$ as follows:

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^{360} \left(-y_i \log \hat{y_i} - (1 - y_i) \log\left(1 - \hat{y_i}\right)\right) \quad (4)$$

4. Experiments

4.1. Experimental details

Datasets. We trained and evaluated pitch estimation models in four different datasets that were resampled to 16kHz as listed.

- **VCTK-corpus (VCTK)** [27] contains 44 hours of clean speech obtained from 109 speakers. We used 100 speakers (40,212 utterances) for the training set and unseen 9 speakers (4,030 utterances) for the test set. The ratio of male to female numbers is close to one.
- **PTDB-TUG (PTDB)** [28], typically used to evaluate the pitch tracking performance, consists of 576 minutes (4,720 utterances) of speech recorded by 20 English speakers. It includes laryngograph and reference pitch trajectories.
- **MDB-stem-synth (MDB)** [29] contains 418 minutes of 230 solo tracks, re-synthesized from the MedleyDB dataset [30]. It consists of various instrumental sounds and singing voices with the F0 annotations.
- **MIR-1k (MIR)** [31] contains 133 minutes of singing voices (11 males, 8 females) recorded with the musical accompaniment, and pitch annotations.

We splitted the datasets into training, validation, and test sets in ratio of 3:1:1, except for the VCTK dataset. To evaluate the robustness of models to environmental distortion, we created a dataset called VCTK-Distortion (VCTK-DT). This dataset was generated by convolving speech signals from the VCTK dataset with room impulse responses obtained from the MIT Impulse Response Survey [32], and adding noise from the NOISEX-92 dataset [33]. The signal-to-noise ratio (SNR) was randomly selected from the ranges (-7, -2, 3, 8, 13) dB for the training set and uniformly selected from the ranges (-5, 0, 5, 10, 15) dB for the test set. To acquire ground-truth pitch trajectory, we used DIO [10, 11] algorithm.

Table 1: *Performance results on four clean datasets. Average raw pitch accuracy (RPA) and raw chroma accuracy (RCA). Both high RPA and RCA scores indicate better performances.*

Model	Params. (M)	Metrics (%)	VCTK	PTDB	MDB	MIR
pYIN [13]	-	RPA ↑	54.20	50.51	90.12	90.47
		RCA↑	55.00	51.30	90.71	91.06
SWIPE [14]	-	RPA ↑	77.74	67.45	92.50	96.36
		RCA↑	73.44	69.50	93.34	96.73
CREPE [18]	22.240	RPA ↑	89.92	81.44	96.34	96.41
		RCA↑	91.23	84.26	96.74	96.72
DeepF0 [19]	4.961	RPA ↑	90.82	93.14	98.38	97.82
		RCA↑	91.33	93.47	98.44	98.28
HarmoF0 [21]	0.377	RPA ↑	95.00	93.56	98.40	98.34
		RCA↑	95.01	93.59	98.46	98.46
MF-PAM	0.362	RPA ↑	**96.62**	**97.12**	**99.20**	**98.97**
		RCA↑	**96.62**	**97.13**	**99.20**	**98.99**
MF-PAM-S	**0.213**	RPA ↑	96.33	96.62	99.05	98.93
		RCA↑	96.33	96.62	99.05	98.96

Network configurations. The input (C_{in}) and output (C_{out}) channel sizes of the PNP-Conv and P-Conv blocks are (1, 6, 12, 24, 48) and (6, 12, 24, 48, 96), respectively, with a stride of 4 and dilation of 1. The kernel sizes (K) are sequentially increased by 4, 4, 8, 8, and 12 in each block to utilize a larger receptive field. For the light BiFPN, depthwise separable convolution layers have a kernel size of 5, strided and dilated by 1. In addition, we up-sampled the input by a factor of 4 using the sinc interpolation filter [34] before the analysis stage to provide richer context information.

Evaluation protocols. We evaluated the pitch estimation performance using raw pitch accuracy (RPA) and raw chroma accuracy (RCA) [35], and the threshold was set to 50 cents. RPA and RCA measure the percentage of the number of frames, in which the pitch errors are smaller than the threshold value. The difference between RCA and RPA is that the RCA ignores the error by a single octave since the chroma represents 12 different pitch classes without the concept of an octave in musical datasets. We measured the mean absolute error (MAE) on the VCTK dataset to evaluate the pitch error in Hz.

4.2. Results

To evaluate the accuracy of the estimated pitch, we compared the pitch estimation performance of the proposed model with the two signal processing based methods (pYIN [13], SWIPE [14]) and three deep learning-based models (CREPE [18], DeepF0 [19], HarmoF0 [21]).

Comparison with baseline models. Table 1 shows the pitch estimation performance of the proposed model and baselines on the four clean datasets. Our proposed model outperformed baselines in every metric and dataset. In general, the number of periodic components in VCTK and PTDB datasets is less than that of musical datasets. Thus the pitch estimation performance of the baseline models showed a more severe degradation than that of MF-PAM. For the MDB dataset, our model showed an estimation accuracy of >99% in terms of RCA and RPA, even with the smallest number of network parameters (0.362 M). We achieved higher pitch estimation performance with over

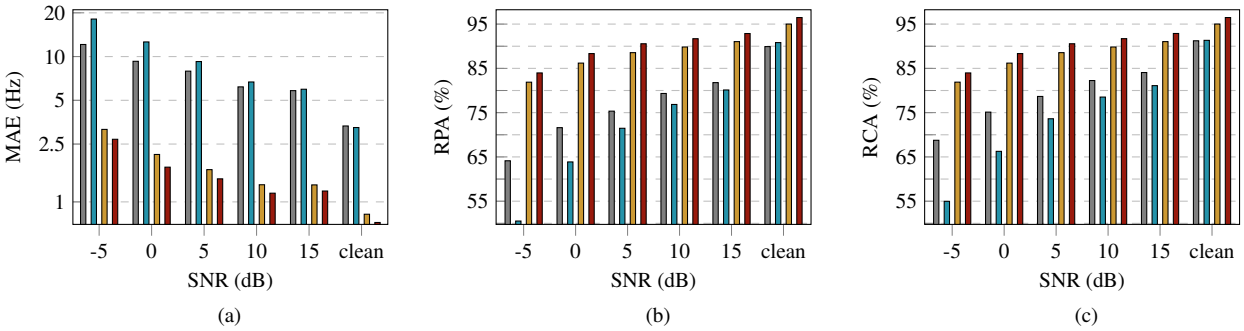

Figure 2: *Pitch estimation performances in various SNRs on the VCTK-DT test set. (a) MAE (Hz) in log-scale; (b) RPA (%); (c) RCA (%). The gray, blue, yellow, and red bars indicate the results for CREPE, DeepF0, HarmoF0, and the proposed model MF-PAM. The low MAE, high RPA, and RCA indicate better performance.*

Table 2: *Performance results on VCTK-DT. Average mean absolute error (MAE), raw pitch accuracy (RPA), and raw chroma accuracy (RCA). Up arrow indicates that higher score is better while down arrow indicates lower score is better.*

Model	FLOPs (G)	Metrics	All	Noise	Reverb	Clean
CREPE [18]	480.875	MAE (Hz)↓	8.24	6.59	5.12	3.33
		RPA (%)↑	74.51	80.43	83.31	89.92
		RCA (%)↑	77.83	83.15	85.36	91.23
DeepF0 [19]	1378.921	MAE (Hz)↓	10.49	8.24	5.54	3.25
		RPA (%)↑	68.66	77.82	80.56	90.82
		RCA (%)↑	70.97	79.28	81.68	91.33
HarmoF0 [21]	43.705	MAE (Hz)↓	1.91	1.66	1.22	0.82
		RPA (%)↑	87.52	90.52	91.18	95.00
		RCA (%)↑	87.53	90.53	91.19	95.01
MF-PAM	**0.101**	MAE (Hz)↓	**1.64**	**1.35**	**1.10**	**0.69**
		RPA (%)↑	**90.05**	**92.20**	**93.29**	**96.62**
		RCA (%)↑	**90.05**	**92.20**	**93.29**	**96.62**
MF-PAM-S	**0.101**	MAE (Hz)↓	2.08	1.56	1.30	0.88
		RPA (%)↑	87.89	91.82	91.45	95.32
		RCA (%)↑	87.89	91.82	91.45	95.32

40 % fewer parameters by eliminating the LSTM layer in MF-PAM (MF-PAM-S), compared to HarmoF0. These results indicate the effectiveness of the proposed modules.

Figure 2 depicts the estimation performance of the baseline models and MF-PAM in various SNRs based on the VCTK-DT. Evidently, MF-PAM significantly outperforms the baselines across all the metrics (MAE, RPA, and RCA), especially in -5 and 0 dB SNRs. Table 2 presents the model performance in various environments; Clean, Noise, Reverberation (Reverb), and all distortions (All) based on the VCTK-DT. The table demonstrates that MF-PAM accurately estimated the pitch in all environments and exhibited a minor performance degradation in harsh conditions.

Ablation study. We further investigated the individual contributions of each proposed module, the PNP-Conv block and the light BiFPN as shown in Figure 3. We replaced the PNP-Conv block with the P-Conv block with a larger hidden channel size to match the model size with the 'w/o PNP-Conv block' setup. For the 'w/o light BiFPN' setup, we removed the light BiFPN layer, and for the 'w/o multi-level' setup, we only used the last feature of the analysis stage as the input for the light BiFPN. As shown in Figure 3, while the pitch estimation accuracy of 'w/o PNP-Conv' was similar to that of MF-PAM in the clean environment (RPA: 96.62% vs. 96.07%), MF-PAM demonstrated more robust pitch estimation performance in low SNRs compared to 'w/o PNP-Conv' (RPA: 84.52% vs. 81.28% in -5 dB SNR). The results demonstrate that the PNP-Conv block encouraged the proposed model to extract only pitch-related information even in extremely noisy conditions. Although 'w/o multi-level' has a bigger model size than 'w/o light BiFPN', both showed similar pitch estimation performance (RPA: 95.42% vs. 95.32% in clean signal). These results indicate that the light BiFPN architecture as well as the multi-level features are crucial for improving the pitch estimation performance.

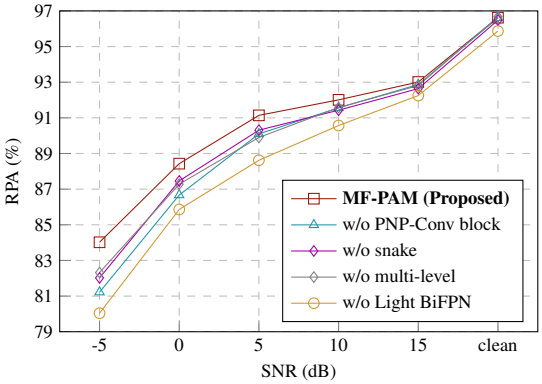

Figure 3: *Ablation study for each module in various SNR levels on the VCTK-DT test set. Average raw pitch accuracy (RPA). High RPA scores indicate better performance.*

5. Conclusions

In this paper, we proposed a novel pitch estimation model, MF-PAM, which extract periodic-related information effectively from the raw audio input using periodicity-sensitive blocks. The pitch-related representation was processed by leveraging a multi-level feature fusion model, BiFPN, and projected onto quantized frequency levels for the pitch estimation. Our experimental results demonstrated that MF-PAM outperformed state-of-the-art baseline models in various datasets and conditions, thanks to its structural configurations that consider the periodicity of speech signals. We further conducted ablation studies to investigate the contributions of the submodules of MF-PAMand confirmed their effectiveness. Moreover, the lightweight version of our proposed model, MF-PAM-S, achieved competitive performance in terms of RPA and RCA with significantly fewer parameters, over 40% less than the smallest baseline model.

6. References

[1] R. Makhijani, U. Shrawankar, and V. M. Thakare, "Speech enhancement using pitch detection approach for noisy environment," *International Journal of Engineering Science and Technology*, vol. 3, no. 2, pp. 1764–1769, 2013.

[2] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "Hgcn: Harmonic gated compensation network for speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.

[3] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[4] K. Wang, F. Soong, and L. Xie, "A pitch-aware approach to single-channel speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[5] E. Song, Y.-S. Joo, and H.-G. Kang, "Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[6] M. K. Reddy and K. S. Rao, "Excitation modelling using epoch features for statistical parametric speech synthesis," *Computer Speech & Language*, vol. 60, p. 101029, 2020.

[7] Y. Cheng and H. C. Leung, "Speaker verification using fundamental frequency." in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1998.

[8] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[9] 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, no. 4, pp. 1917–1930, 2002.

[10] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, 2009.

[11] M. Morise, H. Kawahara, and T. Nishiura, "Rapid f0 estimation for high-snr speech based on fundamental component extraction," *IEICE Transactions on Information and Systems (Japanese Edition)*, vol. 93, pp. 109–117, 2010.

[12] P. Boersma *et al.*, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, 1993.

[13] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

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

[15] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[16] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 1, 2002, pp. I–361.

[17] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2158–2168, 2014.

[18] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[19] S. Singh, R. Wang, and Y. Qiu, "Deepf0: End-to-end fundamental frequency estimation for music and speech signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[20] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "Spice: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.

[21] W. Wei, P. Li, Y. Yu, and W. Li, "Harmof0: Logarithmic scale dilated convolution for pitch estimation," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2022.

[22] L. Ziyin, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[23] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[27] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[28] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *INTERSPEECH*, 2011.

[29] 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 *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[30] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[31] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2009.

[32] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[33] 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, no. 3, pp. 247–251, 1993.

[34] J. Smith and P. Gossett, "A flexible sampling-rate conversion method," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1984.

[35] J. Salamon, E. Gómez, D. P. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.