JUND-F0: A NOVEL DEEP LEARNING FRAMEWORK FOR JOINT UNVOICED/VOICED DETECTION AND F0 ESTIMATION

Yuang Chen[†,1,2], *Rui Feng*[†,1], *Yin-Long Liu*[1], *Yu Hu*[1], *Jiahong Yuan*[*,1,2]

[1] National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, P. R. China
[2] Interdisciplinary Research Center for Linguistic Sciences, University of Science and Technology of China, Hefei, P. R. China

ABSTRACT

This paper addresses the limitations of the confidence-driven unvoiced estimation (CUE) strategy commonly adopted in deep learning-based F0 extraction methods, particularly with respect to adaptability and explicit voiced/unvoiced (V/UV) modeling. To overcome these challenges, we propose a novel deep learning framework, JUND-F0 (Joint V/UV Detection and F0 extraction). JUND-F0 integrates two innovative strategies: the Joint F0 and V/UV learning method (JFUV) and the Unified F0 and V/UV detection method (UFUV). JFUV incorporates V/UV information directly into the network, enabling simultaneous learning of voice activity and pitch distributions. UFUV, in contrast, treats unvoiced frames as an explicit class within the network output, supporting end-to-end V/UV judgments alongside precise pitch predictions. Experimental results demonstrate that JUND-F0 achieves comprehensive improvements over state-of-the-art models across multiple benchmark datasets, yielding superior performance in terms of mean absolute error (MAE), gross pitch error (GPE), raw pitch accuracy (RPA), and voicing decision error (VDE). For reproducibility, our code is publicly accessible at **`https://github.com/RuiFeng-USTC/jund_cya_and_fr`**.

***Index Terms*—** Fundamental frequency (F0), phoneme prediction, multi-task learning, deep learning.

1. INTRODUCTION

Fundamental frequency (F0) extraction, also known as pitch tracking/estimation, is a fundamental task in the field of speech processing [1, 2, 3, 4]. F0 extraction plays a crucial role in speech signal processing, with wide-ranging applications in areas such as speech synthesis [5], speech emotion analysis [6, 7], and speaker recognition [8]. Currently, although existing research has achieved effective and high-quality extraction of F0 from speech signals under clean conditions[3], key factors such as noise, unvoiced segments, and varying speech characteristics significantly degrade the performance of F0 extraction [9, 3].

Moreover, many F0 extraction methods typically struggle to accurately distinguish between voiced and unvoiced frames, leading to substantial errors in F0 estimation [10, 4]. Recently, the vast majority of deep learning-based approaches, taking the Crepe model as an example [11], have demonstrated robustness to noise by proposing a data-driven F0 extraction method based on deep convolutional neural networks that operate on time-domain signals. However, these models exhibit significant limitations when dealing with unvoiced frames since their original labels correspond to classification probabilities for voiced frames within the range of 30-2000 Hz, which results in their outputs not directly estimating whether a given frame is in an unvoiced state (i.e., at 0 Hz) [3, 12, 13].

1.1. Challenges and Motivations

Many deep learning-based F0 extraction methods usually output predicted pitch values accompanied by their corresponding confidence levels [11, 3, 2, 13]. The determination of voicing requires post-processing the output based on these confidence levels. We refer to this fixed-threshold-based post-processing approach as the confidence-driven unvoiced estimation (CUE) strategy, which leads to several challenges:

- **Over-reliance on confidence thresholds**: CUE relies on a manually set confidence threshold for unvoiced or voiced (V/UV) classification, which is difficult to adjust across environments. A high threshold may misclassify normal pitch as unvoiced, while a low threshold may incorrectly detect unvoiced frames as abnormal pitch values.
- **No explicit V/UV modeling**: CUE relies on confidence-based post-processing instead of explicitly modeling "unvoiced", leading to potential misjudgments in ambiguous spectral states.
- **Weak robustness to noise and non-speech**: In noisy or unclear speech, CUE struggles to distinguish between noise and actual speech, especially in complex cases like vibrato or breathiness.

[†] These authors are co-first author. [*] Corresponding author.
{yuangchen21, fengruimse}@mail.ustc.edu.cn, jiahongyuan@ustc.edu.cn.

Motivated by these challenges, this paper introduces two pioneering solutions that enhance the robustness and accuracy of F0 detection across multiple application scenarios:

Joint F0 and V/UV learning (JFUV), which adds a V/UV output head to jointly learn voicing and pitch distributions. It replaces confidence thresholds with an unvoiced classification head for direct V/UV discrimination, while joint modeling enhances performance through positive transfer effects.

Unified F0 and V/UV detection (UFUV), which treats "unvoice" as an explicit output class with 361 dimensions (1 unvoiced and 360 pitch bins). Through a single `softmax` inference, it selects the class with the highest probability, eliminating threshold settings. It enables end-to-end V/UV judgment and pitch prediction, with smoother transitions for weakly voiced frames due to competition within the same probability distribution.

1.2. Main Contributions

This paper develops an innovative deep learning framework named JUND-F0 that implements a multi-level convolutional block-based feature extractor for frame-level joint V/UV detection and F0 extraction from raw audio signals. JUND-F0 can significantly reduce dependence on post-processing by modeling V/UV states within the model itself. The primary contributions are summarized as follows:

- Compared to CUE, JFUV enhances pitch prediction by incorporating a V/UV output module that predicts silence probabilities for each frame. Moreover, it can co-train F0 and V/UV, forming a multi-task learning framework.
- Through joint training, JFUV jointly learns pitch distributions and silent states during training, thus reducing reliance on manual thresholds and providing more direct support for unvoiced detection.
- For UFUV, the unvoiced class is directly integrated into pitch prediction, resulting in an output dimension of 361 that encompasses both 360 pitch classes and one silent class.
- Moreover, UFUV treats unvoiced modeling as an intrinsic component of pitch prediction, allowing the model to learn when to output unvoiced frames through a unified cross-entropy loss during training. It enables end-to-end joint prediction of pitch and unvoiced frames, streamlining post-processing procedures.
- Compared with state-of-the-art models including dio_stone [14], pYIN [15], Crepe [11], Wav2f0 [1], and Yaapt [16], JUND-F0 performs well on popular datasets CSTRFDA [17] and KEELE [18]. The performance in terms of mean absolute error (MAE), gross pitch error (GPE), raw pitch accuracy (RPA), and voicing decision error (VDE) shows significant superior improvements.

2. RELATED WORKS

F0 extraction has been extensively studied in the literature, spanning both traditional signal processing methods and machine learning-based approaches. Traditional methods typically identify periodicity from frequency and time domain features. Time-domain approaches include autocorrelation [19, 20], average magnitude difference function (AMDF) [21], and cross-correlation used in Rapt [22]. YIN improves autocorrelation for noisy conditions [23], while PYIN outputs probabilistic pitch candidates with Viterbi decoding [15]. Frequency domain methods involve cepstrum-based algorithms [24] and SWIPE, which reduces harmonic estimation errors [25]. Hybrid methods, such as Yaapt [26] and NCCF-based approaches [16], integrate temporal and spectral features.

Machine learning has advanced F0 extraction, with methods categorized into supervised and unsupervised learning. Supervised methods train models on labeled F0 data, while unsupervised methods estimate pitch using statistical techniques. Representative methods include Crepe, a convolutional neural network trained on waveforms [11], and a neural cascade architecture for noisy speech [3]. For unsupervised learning, a self-supervised model using constant-Q transform achieves accuracy comparable to supervised methods [2]. MAJL [13] addresses data scarcity with a two-stage training framework and dynamic weighting.

3. JUND-F0 FRAMEWORK

As illustrated in Fig. 1, the pipeline of the developed JUND-F0 framework is introduced, which develops a multi-level convolutional block-based feature extractor to extract high-level speech features and integrates F0 prediction and V/UV detection, achieving joint prediction of F0 and voiced segments.

3.1. Data Preprocessing

3.1.1. Raw F0 Extraction and Cent Representation

For each audio frame, we extract its F0. When direct F0 annotations are unavailable, we estimate F0 from the Electroglottography (EGG) signal by applying the peak detection algorithm [27] to the differentiated EGG (dEGG). This algorithm identifies critical landmarks in dEGG signals, determines vocal fold vibration periods, and derives frame-level F0 from the periodicity. For audio frames with $F_0 > 0$, the value is further converted into cents on a logarithmic scale, as follows:

$$\text{cent} = 1200 \times \log_2\left(\frac{F_0}{10}\right). \tag{1}$$

The cent value is subsequently encoded into a 360-dimensional one-hot vector, where each bin represents a 20-cent interval six octaves from 32.70 Hz (C1) to 1975.5 Hz (B7). Following the label processing strategy in Crepe [11], we further apply Gaussian smoothing with a standard deviation of 25 cents, as follows:

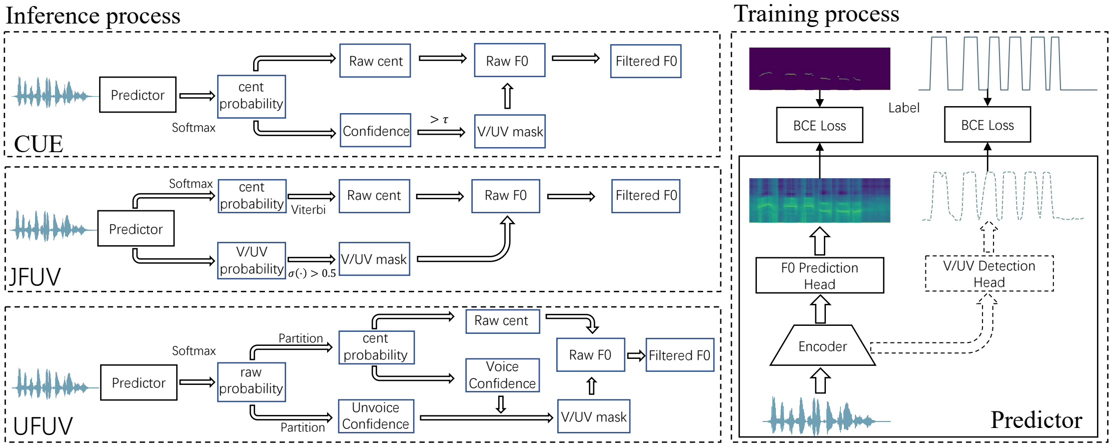


Fig. 1. The pipeline of the developed JUND-F0 framework.

$$y_i = \exp\left(-\frac{(\text{cent}_i - \text{cent})^2}{2 \times 25^2}\right) \tag{2}$$

where y_i denotes the value of the i-th element, and cent_i represents the center value of the corresponding 20-cent interval. If $F_0 = 0$, all 360 elements are set to zero. Depending on the modeling approach, we generate the following frame-level label matrices for an audio segment of N frames:

- CUE: An $N \times 360$ matrix, where each row is the smoothed one-hot vector representing the cent value.
- JFUV: An $N \times 360$ matrix as above, with an additional $N \times 1$ binary vector indicating voicing, where 1 denotes $F_0 > 0$ and 0 otherwise.
- UFUV: An $N \times 361$ matrix, where the first dimension represents the *unvoiced* category, where 1 if $F_0 = 0$, otherwise 0), and dimensions 2 to 361 encode the cent value if $F_0 > 0$, otherwise, they are set to zero.

3.2. Neural Network Architecture

As shown in Fig. 1, our neural network leverages the proposed multi-level convolutional blocks-based feature extractor to derive frame-level features from the second hidden layer, denoted as hidden_states [28]. Compared to the final layer, this intermediate layer preserves more localized acoustic details and contextual information, improving sensitivity to frame-level F0 variations. The Cent Head generates a 360- or 361-dimensional vector, while the V/UV Head outputs a binary voicing value. During inference, the model produces frame-level F0 distributions and V/UV probabilities, with the final F0 sequence obtained via Viterbi decoding. Fine-tuning uses a learning rate of 2×10^{-5}, a batch size of 32, and 48,000 frames per batch.

3.3. Pitch extraction post-processing methods

JUND-F0 employs three post-processing strategies to convert frame-wise model outputs into final pitch estimates. Firstly, the CUE strategy applies `softmax` to frame-wise cent outputs ($\textbf{cent output} \in \mathbb{R}^{n \times 360}$) to obtain pitch probabilities, using the Viterbi algorithm to identify the most likely pitch sequence and classifying frames below a confidence threshold τ as unvoiced. Secondly, the JFUV strategy incorporates an additional V/UV output ($\textbf{V/UV output} \in \mathbb{R}^{n}$) for voicing probability, applying `softmax` to cent outputs and using Viterbi to select the optimal sequence, with frames having sigmoid outputs below 0.5 marked as unvoiced. Thirdly, the UFUV strategy splits frame-wise outputs into unvoiced probability ($\textbf{unvoiced output} \in \mathbb{R}^{n}$) and cent probabilities ($\textbf{cent output} \in \mathbb{R}^{n \times 360}$), applying `softmax` to both parts. A frame is deemed unvoiced if the unvoiced probability exceeds the highest cent probability; otherwise, it is assigned the cent with the maximum probability. **CUE** uses a fixed threshold, **JFUV** leverages V/UV output, and **UFUV** integrates pitch and voicing predictions into a unified output.

4. EXPERIMENTS

4.1. Datasets

As shown in Table 1, we evaluate JUND-F0 against existing algorithms using seven datasets. The PTDB_TUG consists of 2,342 phonetically rich sentences from the TIMIT corpus [29] [1]. The CLD, originating from the 'Computational Linguistic Documentation for 2025' project, provides Vietnamese language data with F0 extracted from EGG signals during speech generation [2] [3]. The CMU_ARCTIC [4], MOCHA_TIMIT, CSTRFDA [17] and KEELE [18] are English-language datasets, detailed in the respective references [30], [31], [17], and [18]. PTDB_TUG, CLD, CMU_ARCTIC, and MOCHA_TIMIT serve as training datasets for F0 extraction, while CSTRFDA and KEELE are used for validation. The WORLD vocoder is optionally used to re-synthesize the reference F0 to make the signal more matched with the label.

[1] `http://www2.spsc.tugraz.at/databases/PTDB-TUG`
[2] `https://pangloss.cnrs.fr/corpus/`
[3] `https://github.com/MinhChauNGUYEN/CLD2025_EGG`
[4] `http://festvox.org/cmu_arctic/`

Table 1. Datasets used for evaluating the proposed F0 extraction framework

Dataset	Content	Num. of Speakers	Audio Scale	EGG	F0
PTDB_TUG [29]	English	10F, 10M	576 min	+	+
CLD	Vietnamese	10F, 10M	364 min	+	-
CMU_ARCTIC [30]	English	1F, 2M	171 min	+	-
MOCHA_TIMIT [31]	English	5F, 4M	278 min	+	-
CSTRFDA [17]	English	1F, 1M	5.53 min	+	+
KEELE [18]	English	5F, 5M	5.62 min	+	+

4.2. Experimental results

We compare JUND-F0 with state-of-the-art methods, including dio_stone [14], pYIN [15], Crepe [11], Wav2f0 [1], and Yaapt [16], with all experiments conducted on an NVIDIA RTX 4090 GPU. F0 extraction performance is evaluated using MAE, GPE, RPA, and VDE. Lower MAE and GPE values indicate higher accuracy, while higher RPA values are preferable. VDE reflects the percentage of frames misclassified in voicing, with lower values reflecting better performance:

$$\mathrm{VDE} = \frac{N_{p\to n} + N_{n\to p}}{N}, \tag{3}$$

where N denotes the number of total frames, $N_{n\to p}$ and $N_{p\to n}$ indicate the number of non-voiced frames that are misclassified as voiced and the number of the voiced frames that are misclassified as non-voiced, respectively.

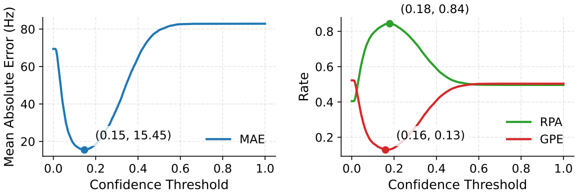


Fig. 2. F0 extraction of CUE heavily relies on confidence levels.

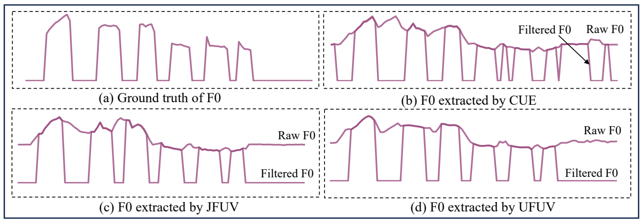
As shown in Fig. 2, we investigated the CUE strategy's F0 extraction performance using MAE, RPA, and GPE across various confidence levels. The curves highlight optimal performance points, i.e., MAE at (0.15, 15.45), RPA at (0.18, 84%), and GPE at (0.16, 13%). To the left of these markers, lower confidence thresholds lead the algorithm to favor the pitch_output, outputting results even at low F0 confidence when the threshold is 0. To the right, higher thresholds lead to rejecting pitch_output, setting F0 to 0, with thresholds above 0.3 resulting in nearly all F0 values being set to 0. These findings indicate that the CUE strategy's F0 extraction performance heavily depends on the chosen confidence threshold, limiting adaptability and introducing robustness challenges, particularly in unvoiced segments.

Table 2 reports the evaluation results on the CSTR and KEELE datasets across MAE, RPA, GPE, and VDE metrics. On the CSTR dataset, the proposed UFUV method achieves the best overall performance, yielding the lowest MAE (8.6 Hz), GPE (4.7%), and VDE (4.4%), while also obtaining the highest RPA (91.7%), clearly outperforming traditional approaches such as dio_stone, pYIN, and Crepe. Although JFUV does not match UFUV in overall accuracy, it consistently improves upon pYIN across all metrics and remains competitive in terms of voiced/unvoiced detection. On the KEELE dataset, similar trends are observed. UFUV again exhibits superior performance, achieving the lowest MAE (11.1 Hz), GPE (7.9%), and VDE (7.0%), together with a high RPA (87.4%), surpassing all benchmark models. JFUV also provides competitive results, with VDE (7.4%) significantly lower than most baselines, demonstrating its effectiveness in unvoiced/voiced detection. These results confirm that the proposed framework, particularly UFUV, provides accurate F0 estimation and voiced/unvoiced detection across different datasets.

Table 2. Evaluation results for different models on CSTR and KEELE datasets.

Model	CSTR MAE ↓ (Hz)	RPA ↑ (%)	GPE ↓ (%)	VDE ↓ (%)	KEELE MAE ↓ (Hz)	RPA ↑ (%)	GPE ↓ (%)	VDE ↓ (%)
dio_stone [14]	14.1	87.4	8.3	9.0	13.0	82.8	9.3	9.1
pYIN [15]	51.6	54.1	33.8	35.4	48.4	57.5	27.2	28.1
crepe [11]	15.9	87.8	7.9	9.5	14.0	87.1	9.6	9.6
wav2f0 [1]	15.2	81.8	9.7	8.2	13.1	78.0	10.3	6.4
yaapt [16]	12.3	84.9	8.3	6.7	16.4	79.8	12.5	9.4
CUE	32.8	75.2	22.7	11.0	15.6	83.5	13.3	12.1
JFUV	24.4	72.7	16.9	16.2	12.8	79.1	8.3	**6.4**
UFUV	**8.6**	**91.7**	**4.7**	**4.4**	**11.1**	**87.4**	**7.9**	7.0

Fig. 3 illustrates the visualization of F0 extraction for CUE, JFUV, and UFUV. CUE often misclassifies unvoiced segments as voiced, leading to poor F0 extraction performance. In contrast, JFUV and UFUV effectively overcome these limitations, delivering F0 extraction closer to the ground truth and significantly improving accuracy without requiring an explicit V/UV decision threshold. These results demonstrate that JFUV enhances performance through joint pitch and V/UV modeling, while UFUV enables end-to-end V/UV decisions and precise pitch prediction, achieving smoother transitions in weak or transitional speech frames via competition between silent and tonal classes in a unified probability distribution.



(a) Ground truth of F0

(b) F0 extracted by CUE

(c) F0 extracted by JFUV

(d) F0 extracted by UFUV

Fig. 3. Visualization of F0 Extraction for CUE, JFUV, and UFUV.

5. CONCLUSIONS

For the first time, this paper identifies the limitations of the commonly used CUE strategy in existing deep learning-based F0 extraction methods. To this end, we propose a novel deep-learning framework, named JUND-F0, that incorporates JFUV and UFUV to enhance the robustness and accuracy of F0 extraction across diverse scenarios. Extensive experimental results prove that JUND-F0 achieves significant performance improvements on several popular datasets compared with state-of-the-art benchmarks.

6. REFERENCES

[1] Rui Feng, Yin-Long Liu, Zhen-Hua Ling, and Jia-Hong Yuan, “Wav2F0: Exploring the potential of wav2vec 2.0 for speech fundamental frequency extraction,” in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, pp. 169–173.

[2] Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirović, “SPICE: Self-supervised pitch estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.

[3] Yixuan Zhang, Heming Wang, and DeLiang Wang, “F0 estimation and voicing detection with cascade architecture in noisy speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3760–3770, 2023.

[4] Rui Feng, Yuang Chen, Yu Hu, Jun Du, and Jiahong Yuan, “EGGCodec: A robust neural encodec framework for EGG reconstruction and F0 extraction,” *arXiv preprint arXiv:2508.08924*, 2025.

[5] Andreas Triantafyllopoulos, Björn W. Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, Ruibo Fu, and Jianhua Tao, “An overview of affective speech synthesis and conversion in the deep learning era,” *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1355–1381, 2023.

[6] Changchong Sheng, Gangyao Kuang, Liang Bai, Chenping Hou, Yulan Guo, Xin Xu, Matti Pietikäinen, and Li Liu, “Deep learning for visual speech analysis: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6001–6022, 2024.

[7] Ringki Das and Thoudam Doren Singh, “Multimodal sentiment analysis: a survey of methods, trends, and challenges,” *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–38, 2023.

[8] Rashid Jahangir, Ying Wah Teh, Henry Friday Nweke, Ghulam Mujtaba, Mohammed Ali Al-Garadi, and Ihsan Ali, “Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges,” *Expert Systems with Applications*, vol. 171, pp. 114591, 2021.

[9] Sam Karimian-Azari, Jesper Rindom Jensen, and Mads Græsbøll Christensen, “Computationally efficient and noise robust DOA and pitch estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1613–1625, 2016.

[10] Anderson Queiroz and Rosângela Coelho, “Noisy speech based temporal decomposition to improve fundamental frequency estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2504–2513, 2022.

[11] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, “Crepe: A convolutional representation for pitch estimation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[12] Takanori Ashihara, Shigeto Furukawa, and Makio Kashino, “Estimating pitch information from simulated cochlear implant signals with deep neural networks,” *Trends in Hearing*, vol. 28, pp. 23312165241298606, 2024.

[13] Haojie Wei, Jun Yuan, Rui Zhang, Quanyu Dai, and Yueguo Chen, “MAJL: A model-agnostic joint learning framework for music source separation and pitch estimation,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8623–8632.

[14] Masanori Morise, Hideki Kawahara, and Haruhiro 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*. Audio Engineering Society, 2009.

[15] Matthias Mauch and Simon Dixon, “pYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 659–663.

[16] Kavita Kasi, *Yet another algorithm for pitch tracking:(Yaapt)*, Ph.D. thesis, Citeseer, 2002.

[17] Paul C Bagshaw, Steven M Hiller, and Mervyn A Jack, “Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching.,” 1993.

[18] F Plante, G Meyer, and W Ainsworth, “A pitch extraction reference database,” *Children*, vol. 8, no. 12, pp. 30–50, 1995.

[19] Lawrence Rabiner, “On the use of autocorrelation analysis for pitch detection,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no. 1, pp. 24–33, 1977.

[20] Lawrence Rabiner, Md Cheng, A Rosenberg, and C McGonegal, “A comparative performance study of several pitch detection algorithms,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[21] Myron Ross, Harry Shaffer, Andrew Cohen, Richard Freudberg, and Harold Manley, “Average magnitude difference function pitch extractor,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[22] David Talkin and W Bastiaan Kleijn, “A robust algorithm for pitch tracking (RAPT),” *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[23] Alain De Cheveigné and Hideki 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.

[24] A Michael Noll, “Cepstrum pitch determination,” *The journal of the acoustical society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[25] Arturo Camacho and John 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.

[26] Lyudmila Sukhostat and Yadigar Imamverdiyev, “A comparative analysis of pitch detection methods under the influence of different noise conditions,” *Journal of voice*, vol. 29, no. 4, pp. 410–417, 2015.

[27] Nathalie Henrich, Christophe d’Alessandro, Boris Doval, and Michèle Castellengo, “On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation,” *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1321–1332, 2004.

[28] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “Wav2Vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[29] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf, “A pitch tracking corpus with evaluation on multipitch tracking scenario.,” in *Interspeech*, 2011, pp. 1509–1512.

[30] John Kominek, “CMU ARCTIC databases for speech synthesis,” *CMU-LTI*, 2003.

[31] Alan Wrench, “‘mocha multichannel articulatory database: English,” 1999.