Improving Long-term F0 representation using post-processing techniques

Crisron Rudolf Lucas and **Diptasree Debnath** and **Andrew Hines**
School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
crisron.lucas@ucdconnect.ie
diptasree.debnath@ucdconnect.ie
andrew.hines@ucd.ie

Abstract

The long-term fundamental frequency of speech (LTF0) represents a speaker's F0 over longer stretches of speech. It can be used as an acoustic feature for speech, e.g. speaker identification and as a controllable parameter in speech synthesis. LTF0 estimation is a challenging task for automatic F0 estimators as they vary in sensitivity, accuracy, and robustness to noisy data. In this paper, we aim to improve the accuracy and robustness of long-term F0 representation of speaker voices using 3 model output post-processing techniques: a) thresholding, b) median filtering, and c) smoothing. We evaluated these for 6 popular F0 estimators: pYin, SWIPE, REAPER, FCNF0, CREPE and SPICE. We evaluated their performance with hand-labelled LTF0 for 5 male and 5 female speaker selected from LibriSpeech as well examining trends for a larger group of 40 speakers. We conclude from our analysis that post-processing significantly improved the CREPE model estimates. SPICE and SWIPE had minimal improvements. As for the other methods, we would not recommend the post-processing techniques.

1 Introduction and Motivation

The fundamental frequency (F0) of speech dictates the pitch and intonation at which the acoustic-linguistic units are spoken. F0 can be measured manually or by using automated F0 estimators such as pYIN or CREPE. However, F0 estimation errors can still occur especially at the high frequencies for unvoiced sounds (See Figure 1). These errors in detection could impact the accuracy and precision when estimating the long-term F0 of a speaker. The long-term F0 represents the fundamental frequency over longer duration of speech as compared to short-term F0 which represents smaller units such as vowels or phonemes (Loakes, 2006). This study investigates the performance of popular F0 estimators on LibriSpeech(Panayotov et al., 2015) and suggests post-processing methods to improve the long-term F0 (LTF0) speaker representation which can be used for better prosody analysis and modelling.

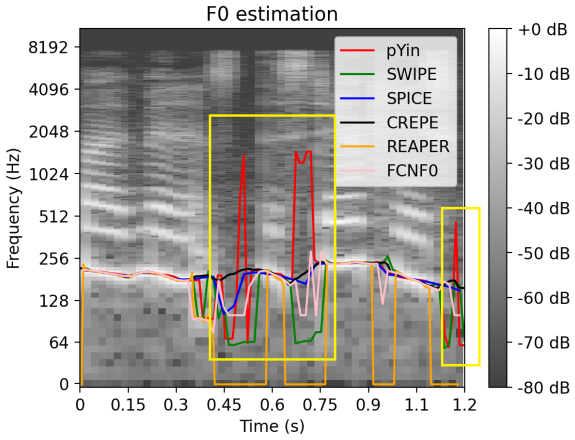

Figure 1: Spectrogram and F0 contour estimates from 6 different models on a sample speech file. Misdetected F0 (errors) are boxed in yellow

2 Related Literature

Traditional signal-based methods for pitch estimation use autocorrelation function (ACF) and spectral analysis. Recent state-of-the-art systems, on the other hand, use data-driven approach and machine learning methods such as Convolutional Neural Networks (CNN), Deep Learning (DL), and Self-Supervised Learning (SSL) techniques.

2.1 F0 estimators

Autocorrelation methods for F0 estimation. The autocorrelation of a signal is calculated by getting the product of a signal with a lagged or time-shifted version of itself. The resulting product has a high value at lags corresponding to the period of the signal. The fundamental frequency can then be calculated from the lag with a maximum autocorrelation value. The YIN algorithm implements the

ACF together with cumulative mean normalised difference function and absolute thresholding to estimate F0 (de Cheveigne and Kawahara, 2002). An improved implementation pYIN uses thresholding distribution instead of a single threshold which resulted to higher accuracy (Mauch and Dixon, 2014). Another algorithm, RAPT, uses normalised cross-correlation for F0 estimation, alongside dynamic programming to remove discontinuities in the F0 estimates. (Camacho, 2008; Talkin, 1995). REAPER[1], which is an improved implementation of RAPT, uses an *epoch tracker* to simultaneously estimate the voiced-speech *epochs* or glottal closure instants, voicing state, and F0.

Spectrum-based methods for F0 estimation. SWIPE is an example of pitch estimator using spectral analysis. It determines F0 from the frequency that maximises the Average Peak to Valley Distance (APVD) in the frequency domain. SWIPE was shown to outperform pYIN, RAPT for musical instruments and canonical speech (Camacho, 2008). Other spectral-based methods estimates F0 by calculating the power-spectrum. TANDEM-STRAIGHT defines a fluctuation spectrum for the periodic components and a separate model for aperiodicity (Kawahara et al., 2008). A more recent method based on pseudo Wigner-Ville distribution using spectral density achieves faster F0 estimation (Liu et al., 2023).

Machine Learning F0 Estimators With the development of speech and audio corpora such as VCTK Corpus (Veaux et al., 2017), PTDB (Pernkopf), MDB (Salamon et al.), and MIR (Lerch) database comes the development of data-driven machine learning models for F0 estimation (Chung et al., 2023; Kim et al., 2018). CREPE (Kim et al., 2018) is a deep convolutional neural network architecture trained using MDB dataset. It has been shown to outperform pYIN and SWIPE in terms of raw pitch accuracy (RPA) across RWC-synth and MDB-stem-synth datasets. MF-PAM (Chung et al., 2023) also uses CNN architecture with periodic and non periodic CNN blocks alongside bi-directional feature pyramid network (Bi-FPN) . The system was shown to outperform pYIN, SWIPE, CREPE, DeepF0, HarmoF0 across the databases mentioned. RMVPE is another pitch estimator that uses log mel spectrogram features, residual CNN with BiGRU and fully-connected layers with sigmoid activation function (Wei et al., 2023). TAPE uses a modified version of CREPE which is connected to a Transformer layer (Tamer et al., 2023). FCNF0 is another modified version of CREPE with fully-convolutional inference, zero-padding omitted from the convolutional layers, and with different number of convolutional channels (Morrison et al., 2023). Lastly, SPICE is a Self Supervised Learning (SSL) approach with Constant-Q Transform (CQT) features and attention layers. SPICE was shown to outperform CREPE, and SWIPE across MIR (1k), MDB-stem-synth, and Singing Voices datasets (Gfeller et al., 2020).

Recent machine learning methods use more complex computations and require training from large datasets compared to the traditional signal processing methods using the autocorrelation function and spectral analysis. However, state-of-the-art models such as CREPE and SPICE are able to achieve high accuracy, outperforming autocorrelation and spectral-based F0 estimators on large music and singing datasets (Kim et al., 2018; Gfeller et al., 2020).

2.2 Calculating the long-term F0 for speech

Mean F0 and F0 histograms extracted using autocorrelation (via PRAAT software (Boersma and Weenink)) were used as complementary features to Mel Frequency Cepstral Coefficient (MFCC) and have been shown to improve text-independent speaker recognition (Kinnunen and Hautamaki, 2005). Another study analysed the effects of using long-term mean F0, standard deviation, kurtosis, skew, modal F0, and the modal density for forensic speaker classification on Japanese speakers (Kinoshita et al., 2008). A more recent study (Arantes et al., 2017) compared mean, median and base values extracted using autocorrelation (via PRAAT software) for long-term F0 estimation and found out that the base value which is defined as 1.43 standard deviations below the mean stabilises faster than the mean and the median. In this study, we will focus on improving long-term F0 using three post-processing techniques: a) thresholding, b) median filtering, and c) smoothing.

2.3 Post-Processing Techniques

1. Thresholding - Single absolute thresholding of confidence score was used in YIN in selecting the smallest period corresponding to the F0 among candidates (Mauch and Dixon, 2014). For pYIN, probabilistic thresholding with beta distributions

[1] Google, 'REAPER: Robust Epoch And Pitch EstimatoR', 2019, https://github.com/google/REAPER

was used to improve the F0 candidate selection of YIN. Peak thresholding in the residuals calculated via autocorrelation was done in REAPER [1] in selecting the glottal closure instants candidates (GCI).

In a similar way, we hypothesise that thresholding can be used in extracting the LTF0 from the F0 contours of a given speaker. We propose to threshold based on three parameters which are the primary basis for extracting LTF0: harmonic, periodic and voiced sounds. We propose to threshold these parameters: a) voiced probability for pYIN, b) confidence scores for CREPE, and SPICE, c) strength (pitch) for SWIPE, d) correlation for REAPER, and e) periodicity for FCNF0.

2. Median filtering - For effective LTF0 representation, the appropriate measure of central tendency must be properly selected. In F0 estimation wherein outliers among the pitch estimates are naturally occurring, we suggest that the median should be a better measure for F0 representation. In a related study on analysis of LibriSpeech data, characteristic median of F0 estimates from pYIN and CREPE were used to characterise intra- and inter- speaker range distributions from which they observed a bimodal distribution across genders(Debnath et al., 2023). A related study on duration modeling demonstrated median as a better estimator than mean for human speech (Ronanki et al., 2016).

3. Smoothing - Temporal smoothing has already been implemented in pYIN and has been shown to improve the precision and F-score in F0 estimation of synthetic singing voice data (Mauch and Dixon, 2014). CREPE (Kim et al., 2018) also included an option for Viterbi smoothing in their repository. There are still other methods for time-series smoothing such as convolutional smoothing, polynomial smoothing, gaussian smoothing, etc.[2]. For post-processing, we hypothesise that applying a smoothing function on any of the F0 estimators will still improve the robustness of the pitch estimators. We select Kalman filter as a robust temporal smoothing algorithm as it considers prior estimates and could perform well on non-stationary time series data (Lotysh and Larysa Gumeniuk, 2023).

Using the combination of these three post-processing techniques, we aim to determine whether these could improve LTF0.

[2]Marco Cerliani, 'A python library for time-series smoothing and outlier detection in a vectorised way', 2023, https://github.com/cerlymarco/tsmoothie

3 Methodology

The flowchart of the methodology is shown in Figure 2:

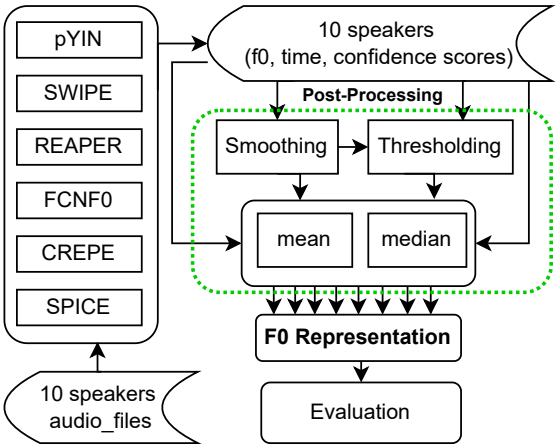

Figure 2: Diagram showing the process of applying post-processing techniques from the F0 estimates of the speech dataset to improve F0 representation

3.1 Data Preparation

Five (5) male and five (5) female speaker data were randomly selected from the *dev-clean* set of LibriSpeech. The speaker folders contain at least one chapter of read audiobooks sampled at 16k Hz with utterance lengths varying from 3-20 s with around 10 minutes per speaker.

3.2 F0 Estimators

Available repositories of the estimators were consolidated and used for benchmarking.

An example showing the F0 estimate of all the 6 models is shown in Figure 1. Minimum F0 was set to 55 Hz and Maximum F0 to 1760 Hz which covers the human voice range of 40 Hz - 450 Hz (Bäckström et al., 2022).

3.3 Post processing techniques

Optimal threshold values were set upon observing the histogram distributions of the confidence scores (see Appendix). For temporal smoothing a Kalman filtering function from *tsmoothie* [2] was applied for all the models. After smoothing and thresholding, the central measures (mean and median) were then calculated.

3.4 Evaluation

Mean Absolute Deviation (MAD), a measure of variability (Amir, 2016), was used to determine the

robustness of the F0 representation with and without the post-processing functions. Mean Absolute Error (MAE) was also computed for the systems with respect to manual labels. For Table 1, variability was compared when using mean LTF0 estimates versus when using post-processed median LTF0 estimates. In Figure 3, improvement in accuracy was determined by getting the difference of mean LTF0 estimates with the ground truth as well as the difference of the post-processed median LTF0 estimates from the ground truth.

and

4 Results and Discussion

4.1 Speaker F0 representation

Accuracy

Manually labelled median F0 estimates (see Appendix tables 2 and 3) were obtained through spectrogram inspection with PRAAT for 10 utterances from each of the 10 speakers. Figure 3 shows the MAE improvement with post-processing versus a simple mean calculation for the different models with respect to the ground truth labels. It is observed that the error for REAPER increased after post-processing while the other systems improved. CREPE exhibited the largest reduction in error (127 Hz), with a lower MAE than SPICE.

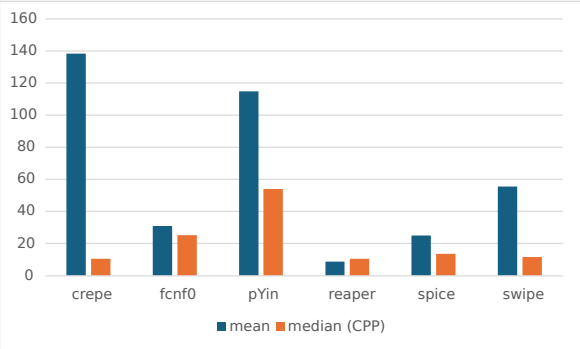

Figure 3: MAE before and after post-processing

Precision

Table 1: Average MAD difference values across the 10 speakers

Algorithms	crepe	fcnf0	pYin	reaper	spice	swipe
Average MAD	25.61	-3.80	28.24	0.03	1.72	4.49

Table 1 shows the average of MAD score differences before and after post-processing for all the algorithms across the 10 speakers. Positive values which indicate reduced variability can be observed mostly for CREPE, pYin, and SWIPE. REAPER and SPICE models have less reduction in variability with 0.03 Hz and 1.72 Hz improvement respectively as compared to 25.61 Hz for CREPE and 28.24 Hz for pYin. Variability in F0 estimates from FCNF0 increased as indicated by the negative values.

4.2 Intergender F0 representation

Figure 4 shows the interquartile range (IQR) of the CREPE F0 estimates across the 10 speakers. It can be observed that median ranking becomes more definitive across gender after applying the the combination of the post-processing techniques as shown by the clearer separation between male and female voices. The rankings were also investigated across all 40 speakers in the *dev-clean* and results are consistent with only one male speaker clustered among female speakers. See Appendix for details.

5 Conclusion and Recommendations

Based on our analysis, we conclude that the post-processing was yielded a significant benefit for CREPE. Post processing also helped pYin but the thresholding is not robust to varied data (details in the appendix). SPICE and SWIPE exhibited minimal improvement with post-processing. REAPER worsened in accuracy and had minimal improvement in variability while FCNF0 had minimal accuracy improvement and increased variability. When used on CREPE, the suggested approach can yield better LTF0 representation which can be used to improve the quality of speech models.

Ethics Statement

We declare that the results presented above are our honest work and there are no ethics issues.

Acknowledgments

This research was conducted with the financial support of Research Ireland under Grant Agreement No. 13/RC/2106_P2 and 12/RC/2289_P2.

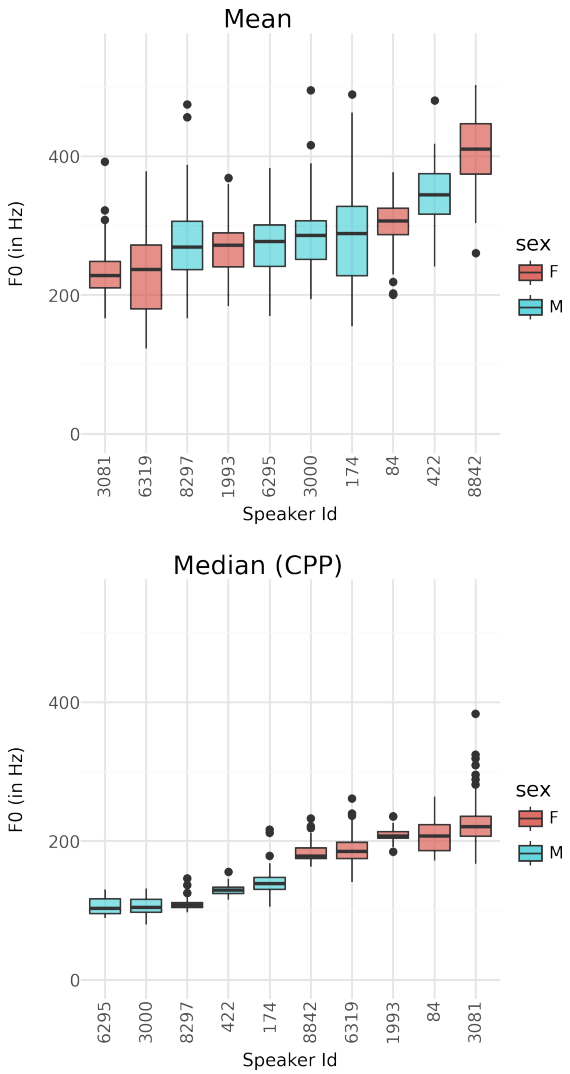

Figure 4: Box plot for the F0 estimation using CREPE with (left) and without (right) post-processing; The center line represents the median F0 and the color indicates male (Blue) and female (Red).

References

Elsayed Ali Habib El Amir. 2016. On uses of mean absolute deviation: Decomposition, skewness and correlation coefficients. In *METRON – International Journal of Statistics LXX (23): 145 – 164*.

Pablo Arantes, Anders Eriksson, and Suska Gutzeit. 2017. Effect of language, speaking style and speaker on long-term f0 estimation. In *Interspeech 2017*. ISCA.

Paul Boersma and David Weenink. Praat: doing phonetics by computer. https://www.fon.hum.uva.nl/praat/. Accessed: 2024-09-25.

Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Marieum Bouafif Mansali, Daniel Ramos, Sudarsana Kadiri, Paavo Alku, and Mohammad Hassan Vali. 2022. *Introduction to Speech Processing*, 2 edition.

Arturo Camacho. 2008. Swipe: A sawtooth waveform inspired pitch estimator for speech and music. In *Journal of Acoustical Society America (JASA)*. Acoustical Society of America.

Woo-Jin Chung, Doyeon Kim, Soo-Whan Chung, and Hong-Goo Kang. 2023. Mf-pam: Accurate pitch estimation through periodicity analysis and multi-level feature fusion. In *Interspeech 2023*. ISCA.

Alain de Cheveigne and Hideki Kawahara. 2002. Yin, a fundamental frequency estimator for speech and music. In *Journal of Acoustical Society America (JASA)*. Acoustical Society of America.

Diptasree Debnath, Helard Becerra Martinez, and Andrew Hines. 2023. Well said: An analysis of the speech characteristics in the librispeech corpus. In *Irish Signals and Systems Conference (ISSC) 2023*. IEEE.

Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirovi. 2020. Spice: Self-supervised pitch estimation. In *IEEE Transaction on Audio, Speech, and Language Processing*. IEEE.

Hideki Kawahara, Masanori Morise, Toru Takahashi, Toshio Irino Ryuichi Nisimura, and Hideki Banno. 2008. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE.

Tomi Kinnunen and Rosa Gonzalez Hautamaki. 2005. Long-term f0 modeling for text-independent speaker recognition. In *Tenth International Conference on Speech and Computers*.

Yuko Kinoshita, Shunichi Ishihara, and Phil Rose. 2008. Beyond the long-term mean: Exploring the potential of f0 distribution parameters in traditional forensic speaker recognition. In *Odyssey 2008: The Speaker and Language Recognition Workshop*. ISCA.

Alexander Lerch. list of mir datasets. https://gist.github.com/alexanderlerch/e3516bffc08ea77b429c419051ab793a. Accessed: 2024-09-25.

Yisi Liu, Peter Wu, Alan Black, and Gopala Anumanchipalli. 2023. A fast and accurate pitch estimation algorithm based on the pseudo wigner-ville distribution. In *2023 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE.

Deborah Loakes. 2006. Variation in long-term fundamental frequency: Measurements from vocalic segments in twins' speech. In *Proceedings of the 11th Australian International Conference on Speech Science and Technology*. Australian Speech Science and Technology Association Inc.

Volodymyr Lotysh and and Pavlo Humeniuk Larysa Gumeniuk. 2023. Comparison of the effectiveness of time series analysis methods: Sma, wma, ema, ewma, and kalman filter for data analysis. In *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*.

Matthias Mauch and Simon Dixon. 2014. Pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE.

Max Morrison, Caedon Hsieh, Nathan Pruyne, and Bryan Pardo. 2023. Cross-domain neural pitch and periodicity estimation. In *IEEE Transaction on Audio, Speech, and Language Processing*. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Franz Pernkopf. Ptdb-tug: Pitch tracking database from graz university of technology. https://www.spsc.tugraz.at/databases-and-tools/. Accessed: 2024-09-25.

Srikanth Ronanki, Oliver Watts, Simon King, and Gustav Eje Henter. 2016. Median-based generation of synthetic speech durations using a non-parametric approach. In *IEEE Workshop on Spoken Language Technology (SLT 2016)*. IEEE.

Justin Salamon, Rachel Bittner, Jordi Bonada, Juan Jose Bosch, Emilia Gómez, and Juan Pablo Bello. Synth datasets.

David Talkin. 1995. A robust algorithm for pitch tracking (rapt). In *Speech Coding and Synthesis*. Elsevier.

Nazif Can Tamer, Yigitcan Ozer, Meinard Muller, and Xavier Serra. 2023. Tape: An end-to-end timbre-aware pitch estimator. In *2023 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. 2023. Rmvpe:arobustmodel for vocal pitch estimation in polyphonic music. In *Interspeech 2023*. IEEE.

A Appendix

A.1 Manual pitch annotation comparison

Two annotators manually estimated the pitch of 10 random samples from each of 10 speakers. The results, presented in Table 2 and Table 3, show a pitch difference of 5 – 20 Hz due to varying annotation methods and sample selection. Annotator 1 chose all samples from a single chapter, while Annotator 2 selected samples from multiple chapters. This variation in pitch is likely influenced by changes in recording settings, chapter content, and background noise. Despite these differences, the high Pearson correlation values indicate strong similarity between the two annotation sets. Moreover, the speaker rankings based on pitch remained consistent for most speakers, with only one exception.

Table 2: A1: Manual annotation for the 10 speakers using 10 utterances each

spk_id	Mean F0	Median F0	Std_dev
84	185.21	184.80	3.08
174	149.59	149.00	2.64
422	115.60	115.20	1.58
1993	204.87	204.70	2.70
3000	85.57	85.37	3.31
3081	237.26	222.50	23.59
6295	95.52	93.32	4.65
6319	194.90	194.70	10.89
8297	105.63	107.20	2.59
8842	174.22	174.80	2.08

Table 3: A2: Manual annotation for the 10 speakers using 10 utterances each

spk_id	Mean F0	Median F0	Std_dev
84	191.08	184.80	17.21
174	131.19	131.05	15.66
422	134.12	135.86	7.24
1993	217.61	219.30	10.11
3000	110.98	105.20	12.65
3081	230.15	230.55	18.01
6295	113.52	112.20	10.17
6319	189.77	189.90	14.76
8297	114.16	114.35	6.44
8842	192.36	187.30	20.15

The Pearson correlation coefficients for the Mean and Median F0 annotations are R=0.9672, and R=0.9693 respectively. The Spearman rank correlation for the Mean and Median F0 annotations are R=0.93939, and R=0.97576 respectively. These results indicate high correlation and high reliability between the two manual annotations.

A.2 Mean Absolute Deviation difference values

Table 4: MAD difference values for the F0 estimators

speaker_id	crepe	fcnf0	pYin	reaper	spice	swipe
84	8.15	-3.07	44.49	0.22	-1.98	-5.25
174	40.30	-1.08	47.23	-2.23	0.46	16.15
422	26.31	-0.64	23.60	-0.89	-0.19	5.65
1993	25.29	-3.77	29.05	0.00	2.78	9.91
3000	29.57	4.49	35.49	-0.66	0.09	10.62
3081	4.76	-20.32	-23.72	1.96	-0.43	-27.60
6295	27.06	2.31	37.46	-0.19	1.36	15.55
6319	27.06	-15.48	9.28	-0.68	0.17	-9.45
8297	34.52	1.94	32.74	1.35	6.40	11.42
8842	33.04	-2.38	46.74	1.46	8.57	17.87
Average	**25.61**	-3.80	**28.24**	**0.03**	**1.72**	**4.49**

Table 4 shows the computed MAD difference (with vs without post-processing) across the 10 speakers. Positive values indicate improvement (reduced variability) while negative values indicate the opposite.

A.3 Mean Absolute Error difference values

Table 5: Accuracy improvement for the F0 estimators

	mean	median (CPP)
crepe	138.26	10.52
fcnf0	30.92	25.29
pYin	114.82	53.97
reaper	8.75	10.52
spice	25.11	13.59
swipe	55.47	11.58

Table 5 shows the computed average MAE across the 10 speakers using the simple mean F0 estimation (column 1) vs post-processing/median (CPP) (column 2). These are the actual values from Figure 3.

A.4 Inter-gender rankings for 40 speakers

Expanding on Section 4.2, we estimated the pitch for all 40 speakers in the *dev-clean* subset of LibriSpeech using the CREPE algorithm, both with and without post-processing. As illustrated in Figures 5 and 6, similar to Figure 4, there was a significant improvement in speaker ranking based on mean pitch. Consistent with our findings, all male speakers exhibited lower pitch than female speakers. However, speaker 7976 displayed an unusually high pitch compared to other male speakers, potentially due to gender preference or mislabeling.

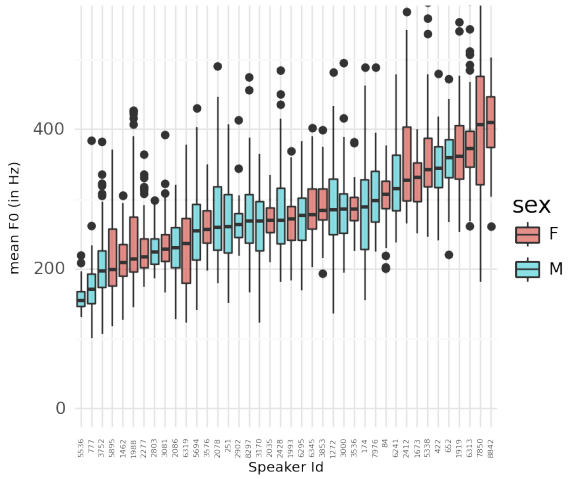

Figure 5: Inter-gender rankings for 40 speakers in LibriSpeech *dev-clean* without post-processing (using mean)

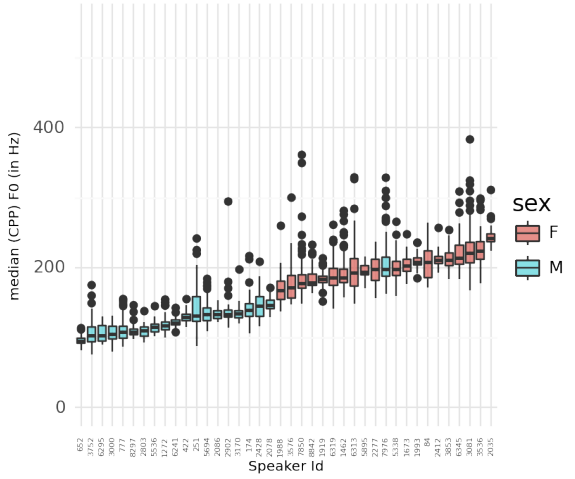

Figure 6: Inter-gender rankings for 40 speakers in LibriSpeech *dev-clean* with post-processing (*median (CPP)*

A.5 Histogram of Threshold values

Figure 7 shows the histogram plots of the confidence scores. The plots show bimodal distribution for CREPE, FCNF0, REAPER, and SPICE. Having thresholds between the peaks would be intuitive for optimal performance of each algorithm. However, for pYin and SWIPE, a continuous trend distribution does not support a justifiable robust threshold selection.

A.6 MAE vs Threshold values

Parameter sweep was done to check how the MAE varies across different thresholds. The succeeding figures show the experimental results from sweeping across a) the 10 speakers, b) only the 5 male speakers and c) only the 5 female speakers.

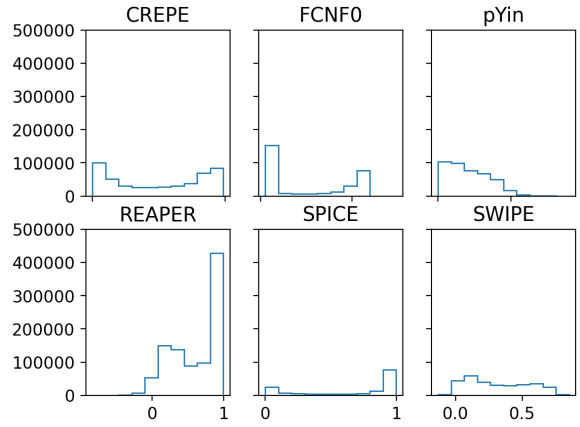

Figure 7: Histogram plot of values for confidence scores (pYIN, CREPE, SPICE) strength (SWIPE), correlation (REAPER), and periodicity (FCNF0)

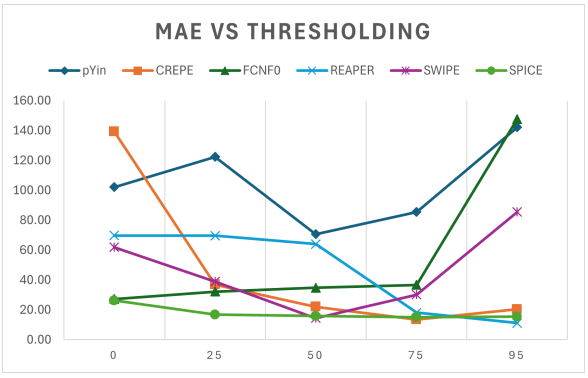

Figure 8: Mean Absolute Error vs Threshold values for all speakers

Based on Figure 8, optimal threshold values are: 0.5 for pYin and SWIPE, 0.75 for CREPE and SPICE, 0 for FCNF0 and 0.95 for REAPER.

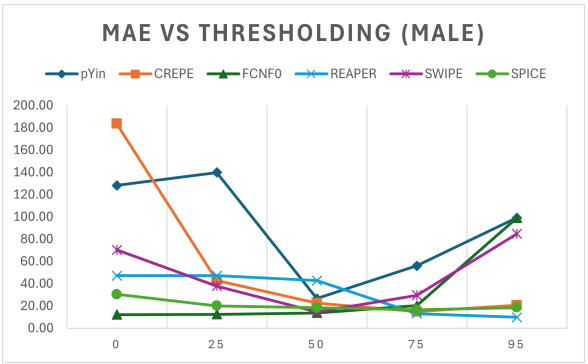

Figure 9: Mean Absolute Error vs Threshold values for male speakers

Same trend can be seen in Figure 9 with just only the male speakers.

Using only female speakers in Figure 10, optimal threshold value for pYin is now at 0 while 0.95 for

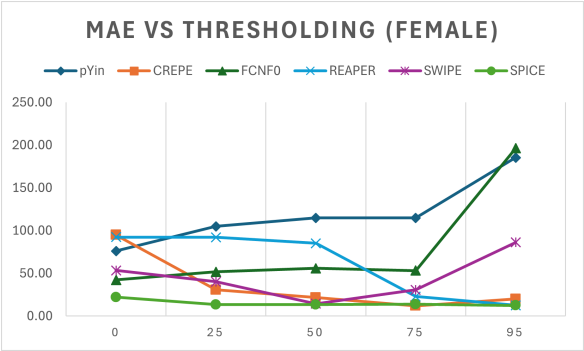

Figure 10: Mean Absolute Error vs Threshold values for female speakers

SPICE.

From the experiments on the threshold values, we observe that applying threshold for pYin will not be robust when used with a different set of data. On the other hand, we can see robust thresholding performance on CREPE, and REAPER.

A.7 Post-processing effects on Spearman Rank correlation

Table 6: Spearman Rank coefficient vs central measures

	mean	**base value**	**median**	**median (CPP)**
crepe-reaper	-0.28	-0.03	0.87	0.99
crepe-spice	0.05	0.33	0.93	0.99
crepe-pyin	0.56	0.77	0.96	-0.36
crepe-swipe	0.65	0.83	0.84	1.00
crepe-fcnf0	0.24	0.71	0.87	0.81

Another experiment was done to see how using the post-processed median, the mean, the median and the base value (Arantes et al., 2017) affects the rankings across the different algorithms. Looking at Table 6, we can see that rankings between CREPE, REAPER, SPICE, and SWIPE become highly correlated after post-processing.