# EVALUATION OF MACHINE-LEARNING PITCH ESTIMATION ALGORITHMS

**Takeshi Ikuma, PhD[1,2], Andrew J. McWhorter, MD[1,2], Melda Kunduk, PhD[1,2,3]**

[1]Dept. of Otolaryngology-Head and Neck Surgery, LSU Health Sciences Center, New Orleans, Louisiana, U.S.A.
[2]Voice Center, The Our Lady of The Lake Regional Medical Center, Baton Rouge, Louisiana, U.S.A.
[3]Dept. of Communication Sciences & Disorders, Louisiana State University, Baton Rouge, Louisiana, U.S.A.

*Keywords*: Disordered Voice; Acoustic analysis; Pitch; Subharmonics

## Abstract

**Objectives / Introduction:**

This presentation aims to demonstrate the performance of existing machine-learning (ML) based pitch detectors, especially for handling so-called subharmonic errors. Accurate pitch estimation is vital for objective voice analysis. Most acoustic parameters rely on either periods or fundamental frequencies $f_o$ of vocal cycles with exception of only a handful parameters (e.g., cepstrum peak prominence and loudness). In a recent study[1], ML algorithms—CREPE[2] and FCN-F0[3]—were shown to outperform the Praat pitch analysis algorithm[4], especially in estimating $f_o$ of the voices of head and neck cancer patients. However, the study did not reveal what type of errors these ML algorithms mitigate.

The present study was focused on analyzing a particular type of pitch estimation error: incorrect selection of an integer-divisor of the true $f_o$ as the estimated. We call this type of error a subharmonic error because it is registered mostly when a voice signal is nonmodal and contains subharmonics. The challenge here is that the true fundamental frequency of a subharmonic signal is indeed an integer-divisor of the speaking $f_o$, i.e., the pitch intended by the speaker. As such, a pitch detector cannot solely depend on the periodicity of the signal and requires additional information to make a correct choice.

**Methods:**

***Data*:** All 709 sustained /a/ recordings of KayPENTAX Disordered Voice Database [5] were included. While each recording was processed at once, pitch estimates in a 50-millisecond (ms) interval were pooled for the analysis, yielding 16174 total sample points (frames).

$f_o$ **Annotation.** The truth values for the fundamental frequency $f_o$ in all signal intervals were evaluated in three steps. First, the initial estimates were gathered from the Praat. Then, these estimated $f_o$'s were reviewed, and those which Praat incorrectly estimated were adjusted manually with a custom computer program. Finally, the estimates were refined using the time-varying harmonic model with a gradient-based optimization[6].

**Algorithms.** *Praat* pitch detector is based on the autocorrelation function and uses the hidden Markov model (HMM, an unsupervised machine-learning technique) as the postprocessor to correct the errors. The default configuration was used except for the minimum pitch was adjusted to 60 Hz to handle the lowest $f_o$ present in the dataset. This results in Praat to set its analysis frame size to be 50 ms. The recordings were resampled to 8 kHz first. The *CREPE* algorithm[2] uses a six-layer convolutional neural network (CNN) model. The model signals to be resampled to 16 kHz with input signal frame size of 1024 samples (64 ms; 28% frame overlap). The original model coefficients were used. The *FCN-F0* algorithm FCN-F0 algorithm[3] is an extension of the CREPE with a 7-layer model, taking input signals sampled at 8-kHz. Specifically, the FCN-933 model with input frame size of 993 (124 ms; 148% frame overlap) and the original model coefficients were used. Finally, the best Praat pitch candidates before postprocessing (i.e., the estimates of an autocorrelation-function based algorithm, *ACF*) was evaluated as a reference without employing any machine learning techniques.

**Performance Metric.** To detect subharmonic errors accurately, the output of each algorithm was refined using the time-varying harmonic model as the annotated truth. The refined estimate was recorded as $\hat{f}_o$. In other words, the algorithm output was considered correct if it yields the same harmonic model as the annotated. The ratio of the truth and estimated, $f_o/\hat{f}_o$, was used as the performance metric.

If $f_o/\hat{f}_o \approx 1$, the estimated $f_o$ matches the truth (labeled "Correct"). On the other hand, if $f_o/\hat{f}_o$ is approximately an integer greater than 1, the algorithm committed a subharmonic error (labeled "Subh"). Note that this error does not guarantee that the signal contains subharmonics as the error can also be caused by chance. Finally, non-integral $f_o/\hat{f}_o$ values indicate either that the estimator picked an $\hat{f}_o$ with no apparent relationship to $f_o$ or that the truth $f_o$ was too aggressively annotated (labeled "Other"). To account for a numerical error, $f_o/\hat{f}_o$ value within ±0.01 of an integer was treated as an integral outcome, i.e., either "Correct" or "Subh".

**Results:**

Out of 16174 frames, 15956 (98.7%) were found to contain periodic behavior with annotated $f_o$'s. Based on the $f_o/\hat{f}_o$ metric, Table 1 shows the overall performance of the pitch detectors. It is apparent that the ACF is prone to making subharmonic errors (34.5%) and Praat's HMM successfully eliminated 75.2% of these errors. For these two algorithms, the subharmonic errors are dominant over the other types of errors. Meanwhile, both CNN solutions clearly outperform the former two with above 95% correctness. More importantly, they demonstrated their resiliency to the subharmonic errors: reduced the amount of subharmonic error by 94.3% for CREPE and 96.1% for FCN-F0

relative to ACF. The CREPE and FCN-F0 committed fewer subharmonic errors than the other types. A minor downside of both CREPE and FCN-F0 was that they introduced a few frequency-doubling errors ($f_o/\hat{f}_o = 0.5$; 17 for CREPE and 11 for FCN-F0) which Praat had none of and ACF had in 3 frames.

Table 2 shows how Praat, CREPE, and FCN-F0 improved over ACF. The top row of the table reveals that the CNN algorithms selfdom made errors on the frames which the ACF was correct (<1.0%), especially the subharmonic errors (<0.2%). In comparison, the HMM postprocessor of Praat incorrectly flipped 1.5% of the correct ACF estimates to subharmonic errors.

Table 1: Pitch Detection Outcomes (occurrences)

| | ACF | Praat | CREPE | FCN-F0 |
|---|---|---|---|---|
| Correct | 9830 (61.6%) | 13931 (87.3%) | 15200 (95.3%) | 15335 (96.1%) |
| Subh | 5506 (34.5%) | 1368 (8.6%) | 316 (2.0%) | 217 (1.4%) |
| Other | 620 (3.9%) | 657 (4.1%) | 440 (2.8%) | 404 (2.5%) |
| Subh, % rel ACF | 100.0% | 24.8% | 5.7% | 3.9% |

Table 2: Contingency matrix between ACF and other algorithms (occurrences)

| ACF | Praat | | | CREPE | | | FCN-F0 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct | Subh | Other | Correct | Subh | Other | Correct | Subh | Other |
| Correct | 9646 | 152 | 32 | 9740 | 15 | 75 | 9762 | 3 | 65 |
| Subh | 4269 | 1209 | 28 | 5239 | 247 | 20 | 5315 | 179 | 12 |
| Other | 16 | 7 | 597 | 221 | 54 | 345 | 258 | 35 | 327 |

**Conclusions:**

The CNN-based pitch detectors—FCN-F0 and CREPE—have demonstrated dominant performance over the pitch detector of the widely used Praat software by correctly detecting the pitches on over 95% of the sustained /a/ audio frames. Remarkably, these detectors are driven by model coefficients which were trained with synthesized data (CREPE) or with speech database without any voice disorder samples (FCN-F0). The CNN detectors especially excelled in avoiding committing subharmonic errors likely by learning the details of the cyclic behavior of the signals beyond their periodicity. These CNN (and other deep learning) models could improve their performance further by training them with clinically relevant data and by using a postprocessing technique similar to Praat.

## References

1. Vaysse R, Astésano C, Farinas J. Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *J Acoust Soc Am*. 2022;152(5):3091-3101. doi:10.1121/10.0015143
2. Kim JW, Salamon J, Li P, Bello JP. Crepe: A Convolutional Representation for Pitch Estimation. In: *2018 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*.; 2018:161-165. doi:10.1109/ICASSP.2018.8461329
3. Ardaillon L, Roebel A. Fully-Convolutional Network for Pitch Estimation of Speech Signals. In: *Interspeech 2019*.; 2019:2005-2009. doi:10.21437/Interspeech.2019-2815
4. Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc Inst Phonet Sci*. 1993;17:97-110.
5. KayPENTAX, Massachusetts Eye and Ear Infirmary. Disordered Voice Database and Program [Model 4337]. Published online 2006.
6. Ikuma T, Story B, McWhorter AJ, Adkins L, Kunduk M. Harmonics-to-noise ratio estimation with deterministically time-varying harmonic model for pathological voice signals. *J Acoust Soc Am*. 2022;152(3):1783-1794. doi:10.1121/10.0014177