An objective test tool for pitch extractors’ response attributes

Hideki Kawahara1, Kohei Yatabe2, Ken-Ichi Sakakibara3, Tatsuya Kitamura4, Hideki Banno5, Masanori Morise6

1Wakayama University, Wakayama, 640-8510 Japan
2Tokyo University of Agriculture and Technology, Tokyo, 184-8588 Japan
3Health Science University of Hokkaido, Hokkaido, 061-0293 Japan
4Konan University, Kobe 658-8510 Japan
5Meijo University, Nagoya 468-8502 Japan
6Meiji University, Tokyo, 164-8525 Japan

kawahara@wakayama-u.ac.jp, yatabe@go.tuat.ac.jp, kia@hoku-iryo-u.ac.jp, t-kitamu@konan-u.ac.jp, banno@meijo-u.ac.jp, mmorise@meiji.ac.jp

Abstract

We propose an objective measurement method for pitch extractors’ responses to frequency-modulated signals. It enables us to evaluate different pitch extractors with unified criteria. The method uses extended time-stretched pulses combined by binary orthogonal sequences. It provides simultaneous measurement results consisting of the linear and the non-linear time-invariant responses and random and time-varying responses. We tested representative pitch extractors using fundamental frequencies spanning 80 Hz to 800 Hz with 1/48 octave steps and produced more than 2000 modulation frequency response plots. We found that making scientific visualization by animating these plots enables us to understand different pitch extractors’ behavior at once. Such efficient and effortless inspection is impossible by inspecting all individual plots. The proposed measurement method with visualization leads to further improvement of the performance of one of the extractors mentioned above. In other words, our procedure turns the specific pitch extractor into the best reliable measuring equipment that is crucial for scientific research. We open-sourced MATLAB codes of the proposed objective measurement method and visualization procedure.

Index Terms: Voice chain, fundamental frequency, frequency modulation, time-stretched pulse, linear time invariant system

1. Introduction

Scientific investigations need reliable and accurate measuring equipment. A pitch extractor serves as critical equipment for investigating voice fundamental frequency. We introduce a method to objectively measure pitch extractors’ performance regarding the frequency transfer function, total distortions, and signal-to-noise ratio. We implemented the method using MATLAB and open-sourced. We measured representative pitch extractors and reported raw data, scientific visualization movies and characterized pitch extractors on several performance maps. These measurement and visualization led us to make one of the pitch extractors a promising tool for voice scientific research.

The contributions of this article are as follows. We introduced a new objective measurement method of pitch extractors which provides useful supplemental information to existing evaluation measures. We applied this method on existing pitch extractors and tuned NINJAL extractor [1, 2] a reliable and accurate measuring equipment for speech science research.

2. Background

Revisiting a research topic a quarter-century ago motivated this research. The first author introduced an altered auditory feedback technique and reported that our voice fundamental frequency (f0) responds to feedback pitch modification [4]. Despite decades of research, altered feedback still is a hot topic for investigating pitch perception and underlying neural basis [5–9]. Those research focused on voluntary responses represented by the pitch shift paradigm and adaptation paradigm [10, 11].

Introduction of CAPRICEP-based method [14] and voice pitch response to auditory stimulation, which is not an altered feedback voice, opened a new possibility for investigating less studied involuntary response [12, 13]. The experiment requires the measurement of f0 of periodic sounds without introducing nonlinearities and glitches for measuring voice response to auditory stimulation. This requirement led us to propose an objective response measurement method of pitch extractors’ response to frequency-modulated tones [5]. We assigned the modulation signal as input and the extracted pitch of the produced voice as output and fed them to the proposed response analysis method. We found that this assignment enables us to acquire pitch extractors’ behavior objectively and in detail.

3. Principles of operation

Pitch extraction has long research history [16–22] and still is a hot topic in speech processing [23–25]. We propose to apply the CAPRICEP-based method [14] that is adopted to investigate voice pitch response to auditory stimulation [12, 13] to measure the responses of pitch extractors.

Figure 1 shows the scheme of the proposed method. The combination of elements Base signal generator, Response analyser, and CAPRICEP-set and orthogonal matrix is the CAPRICEP-based method [14]. We used a Gaussian smoother for LPF and fed the output to VFO (Vari...
Figure 1: Measurement Scheme of the proposed method. (VFO: Variable Frequency Oscillator with harmonic components. The fundamental carrier frequency is the average \( f_o \))

Figure 2: Schematic diagram of the response analyzer of Fig. 1. We normalize the pitch extractor’s responses (Target system measurement) by the test system response (Test system calibration) to get the Target response. The pitch extractor’s response should only consist of the LTI response and not introduce the non-LTI or random and time-varying responses.

3.1. Implementation

Figure 2 shows a schematic diagram of the response analyzer of Fig. 1. The test signal is a mixture of three sequences. Convolution with three unit-CAPRICEP signals with time reversal (“A,” “B,” and “C”) in Fig. 2 separates the test signal into three sequences. The following orthogonalization procedure removes cross-correlation between sequences. The LTI-response is the average of all element responses because the impulse response is identical irrespective of the test signals. The difference between impulse responses obtained in each sequence provides sample variance representing random responses. The difference between impulse responses obtained in different test sequences provides sample variance representing non-linear time-invariant response. Please refer to details for [13].

We refined the implementation of algorithms described in [13] by taking advantage of the signal periodicity using FFT to calculate discrete Fourier transform. Figure 2 shows that the test signal is a periodic signal with the length of the period 4\( N_o \), where \( N_o \) is the allocation interval of each unit-CAPRICEP. We made periodic signals \( x_k[n] \) and \( y_k[n] \) (where, \( k \in \{1, 2, 3\} \), and \( n = 0,\ldots, 4N_o - 1 \) represents the discrete time) from the separated test signals \( x_k[n + n_0] \) and the measured signals \( y_k[n + n_0] \) using the following equation.

\[
\begin{align*}
\hat{x_k}[n] &= w[n]x_k[n + n_0] + u[4N_o - 1 - n]x_k[n + 4N_o + n_0] \\
\hat{y_k}[n] &= w[n]y_k[n + n_0] + u[4N_o - 1 - n]y_k[n + 4N_o + n_0],
\end{align*}
\]

where \( w[n] \) represents a weighting function that satisfies \( w[n]+u[4N_o - 1 - n] = 1 \), and \( n_0 \) represents the initial position of \( \hat{x}_k[n] \).

\( \hat{y}_k[n] \) is identical irrespective of the test signals. The difference between impulse responses obtained in different test sequences provides sample variance representing non-linear time-invariant response. Please refer to details for [13].

3.2. Interface function for pitch extractors

We prepared an interface function for each pitch extractor and called it from the response analyzer. This interface makes it easy to test different pitch extractors.

4. Measurement of pitch extractors

We used MATLAB to implement the response analyzer and interface functions. We tested representative pitch extractors to illustrate the proposed objective measurement method.

4.1. Tested extractors

The tested pitch extractors are listed below with abbreviations for representing them in plots. Some extractors’ test results are in supplement materials due to the limited available space.

MATLAB function (LHS,CEPSRH,NCF,PEF,CREPE)

Scientific computing environment MATLAB has built-in functions for pitch extraction. They consist of classical methods (“CEP” cepstrum-based method [17], “NCF” LPC-based method [18], and “LHS” harmonic summation-based method [26]), and recent methods (“PEF” [27] and “SRH” summation of residual harmonics [22]). In addition to these “Deep Learning Toolbox” has a deep learning-based method “CREPE” [25].

YIN

YIN is an absolute difference-based method [20].

SWIPEP

SWIPEP uses a harmonic model-based approach. It estimates the fundamental frequency of the best matching sawtooth signal [28].

RAPT

RAPT (VOICEBOX) RAPT uses a two-stage autocorrelation-based method followed by post processing [29] for robust processing. We used implementation in voicebox [30].

REAPER REAPER simultaneously estimate GCI (Glottal Closure Instant), V/UV (voiced or unvoiced), and pitch. We used the open-source implementation [31].

Praat Praat is a popular tool for doing phonetics using computers [32]. We used recommended procedure “Sound: To Pitch...” which uses the autocorrelation-based method [33] with the default setting.

openSMILE

openSMILE is widely applied in automatic emotion recognition for affective computing [34]. It has two configuration files for pitch analysis, “prosodyShs.conf” which uses the Sub-Harmonic-Sampling (SHS) method, and “prosodyAcf.conf” which uses an autocorrelation and cepstrum-based method (ACF).

STRAIGHT

STRAIGHT (NDF, XSS) Two Vocoder packages, legacy-STRAIGHT [19] and TANDEM-STRAIGHT [35] use \( f_o \) adaptive spectral envelope recovery. Their \( f_o \) adaptive procedure led to development and refinement of dedicated pitch extractors, NDF [21] for legacy-STRAIGHT, and XSS [36] for TANDEM-STRAIGHT.

WORLD

WORLD (Harvest) A high-quality Vocoder WORLD [37] also use \( f_o \) adaptive spectral envelope recovery. The latest pitch extractor for WORLD is Harvest [24].

We used the integrated and normalized rectangular windowing function to remove level discontinuity of the truncated observation noise. This weighting needs 1.76 dB compensation of the calculated sample variance. Integrated and normalized half-wave rectified cosine windowing function assures continuity up to the first-order derivative with 1.25 dB compensation.

The original YIN is compatible with the latest MATLAB version (R2022a) on Windows11. However, only one specific calling sequence of YIN \( r = yin(x, p) \) worked adequately on macOS.
Figure 3: Upper plot: An example absolute value of unit-
CAPRICEP. The vertical lines represent the allocation interval. Lower plot: frequency modulation signal (blue line) and the extracted pitch by NCF (red line).

NINJAL. This pitch extractor [1, 2] is a refined version of its predecessor [38]. They use a log-linear filter-bank and their instantaneous frequency and residual levels of outputs. We set the smoothing length parameter to 10 ms (named NINJALX2) and 40 ms (named NINJAL: default). The setting of NINJALX2 is the result of fine-tuning enabled by the proposed objective measurement.

4.2. Test signal generation and pitch extraction

We generated test signals using a 44100 Hz sampling rate. For pitch extractors implemented using MATLAB, we fed the test signal directly. We used a WAVE format file with 24 bit quantization for feeding the test signal to other pitch extractors. The length of the signals is 20 s. We set the fundamental frequencies of the test signals from 80 Hz to 800 Hz with 1/48 octave step. We generated 1000 unit-CAPRICEP signals and selected a set consisting of 10 signals having minimum cross-correlation.

The upper plot of Fig. 3 shows an example of the absolute amplitude and allocation interval. We selected three elements from the set of unit-CAPRICEP and allocated them 36 times with an overlap-and-add procedure changing polarity as shown in Fig. 2.

We convolved the signal with a Gaussian shape and used it for modulating $f_o$ of multiple harmonic signals. Then we applied a spectral shaper simulating a Japanese vowel /a/. The lower plot of Fig. 3 shows an example of the generated modulation signal and example output of a pitch extractor NCF. Note that the estimated pitch contour shows a quantization effect due to the sampling interval of discrete-time signals.

4.3. Analysis example

We selected consecutive segments consisting of practically periodic signals. First, we recovered the pulse by convolving time-reversed versions of each unit-CAPRICEP. Then, we calculated the standard deviation of the difference of consecutive segments. The length of each segment is $4N_u$, the fundamental period of the modulation signal.

Figure 4 shows the standard deviations of differences between consecutive segments. We set a -150 dB threshold level for selecting practically periodic pairs and selected six pairs. Then, we made six periodic segments from the reference and the measured segments using Eq. 1 and Eq. 2.

We applied the procedure in Fig. 2 and got 72 ($3 \times 4 \times 6$) impulse responses (length: $N_u$). We used them to get six extended pulse responses (length: $4N_u$), average extended impulse response, and random and nonlinear responses. The upper plot of Fig. 5 shows examples of individual extended impulse response and the corresponding impulsive modulation. The lower plot of Fig. 5 shows the frequency characteristics of the responses.

4.4. Response of representative extractors

This section shows the frequency response of representative pitch extractors. We placed the other extractors’ responses in supplemental materials.

Figure 6 shows responses of popular pitch extractors. SHS configuration of openSMILE provided better performance than ACF configuration. Its default frame interval (10 ms) is the main cause of the narrow bandwidth. Praat has wider bandwidth and relatively good total distortion. REAPER has a significantly wide bandwidth. However, the total distortion is poor. This distortion is mainly because of the frequency quantization due to the sampling interval of the signal.

Pitch extractors in Fig. 7 show better performance both in the bandwidth and the total distortion than other extractors tested. The significantly lower random response in NINJALX2 is due to its output rate (44100/6 Hz this case). The relatively narrow bandwidth of Harvest is the result of tuning to avoid perceptually annoying pitch errors.

4.5. Visualization and performance map

The tests produced more than 2000 plots. It was challenging to understand their behavior by inspecting all plots. Instead, we compiled a movie of about 5 seconds, placing 16 plots having the same fundamental frequency in each movie frame. Figure 8 shows a snapshot. We found that movie visualization makes it easy to characterize the behavior of each extractor as a whole.

11We applied modulation in the logarithmic frequency domain (musical cent) regarding $f_o$ dynamics [39].

12It can operate at 44100 Hz. The default setting introduces automatic downsampling for reducing the processing time.
These visualizations suggested an improvement of our recent extractor, NINJAL. We found the smoothing time of the original NINJAL is excessive by inspecting the movie and plots. Figure 9 shows scatter plots of the extractors on the bandwidth-SNR plane for 240 Hz in the left plot. We noticed that gains of some pitch extractors are not smooth and change depending on both the modulation frequency and the fundamental frequency. These changes are not desirable for measuring equipment. The right plot shows the scatter plot of these changes. The extractors in the bottom left are the better ones. The movie and these plots indicate that the revised version of NINJAL (NINJALX2) best fits scientific research for voice fundamental frequencies.

5. Conclusion

We proposed an objective and informative measurement method of pitch extractors’ response to frequency-modulated tones. It uses a measurement procedure based on a new member of time-stretched pulses and simultaneously yields the modulation frequency response and the total distortion. We applied various pitch extractors and placed them on the bandwidth-SNR plane. We also applied the proposed method to tune one of the extractors to make it best fit scientific research of voice fundamental frequencies. We placed the code and the data on our GitHub repository [40] and made it open-sourced. It consists of example MATLAB functions for applying this method to openSMILE, Praat, and other pitch extractors. The motivation of this research was particular. However, the outcome is a general-purpose tool applicable to the more comprehensive range of tuning tasks of pitch extractors, for example, noise tolerance, reverberation tolerance, and tolerance to recording conditions.

6. Acknowledgements

We appreciate Heiga Zen of Google Brain and Kikuo Maekawa of NINJAL (National Institute for Japanese Language and Linguistics) for inviting the first author for their projects. Participating in their projects enabled us to develop the reference pitch extractor NINJAL. The naming is an acknowledgment for the Center for Corpus Development, NINJAL. This research was also supported by Grants in Aid for Scientific Research (Kakenhi) by JSPS numbers, JP18K00147, JP18K10708, JP19K21618, JP21H03468, JP21H04900, and JP21H00497.

7. References

[1] H. Kawahara, “Application of time-frequency representations of aperiodicity and instantaneous frequency for detailed analysis of filled pauses,” Journal of the Phonetic Society of Japan, vol. 21, no. 3, pp. 63–73, 2017.
[2] H. Kawahara, K.-I. Sakakibara, M. Morise, H. Banno, and T. Toda, “Accurate estimation of \( f_0 \) and aperiodicity based on pe-
H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Iribo, “Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT,” in Proc. Interspeech, 2005, pp. 537–540.

T. Drugman and A. Alwan, “Joint robust voicing detection and pitch estimation based on residual harmonics,” in Proc. Interspeech, 2011, pp. 1973–1976.

G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP—A collaborative voice analysis repository for speech technologies,” in Proc. ICASSP, 2014, pp. 960–964.

M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in Proc. Interspeech, 2017, pp. 2321–2325.

J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “CREPE: A convolutional representation for pitch estimation,” in Proc. ICASSP, 2018, pp. 161–165.

D. J. Hermes, “Measurement of pitch by subharmonic summation,” J. Acoust. Soc. Am., vol. 83, no. 1, pp. 257–264, 1988.

S. Gonzalez and M. Brookes, “PEFAC - a pitch estimation algorithm robust to high levels of noise,” IEEE/ACM Trans. ASLP, vol. 22, no. 2, pp. 518–530, 2014.

A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” J. Acoust. Soc. Am., vol. 124, no. 3, pp. 1638–1652, 2008.

D. Talkin, “A robust algorithm for pitch tracking (RAPT),” in Speech coding and synthesis, D. Talkin and W. B. Kleijn, Eds, Elsevier Science, 1995, pp. 497–512.

S. G. E. Brukal, A. D. M. Africa, and E. P. Dadios, “Female voice recognition using artificial neural networks and MATLAB voice-box toolbox,” J. Telecommunication, Electronic and Computer Engineering, vol. 10, no. 1-4, pp. 133–138, 2018.

D. Talkin, REAPER: Robust epoch and pitch estimator, (retrieved 21 March 2022). [Online]. Available: https://github.com/google/REAPER

P. Boersma and D. Weenink. (1992–2022) Praat: doing phonetics by computer [computer program] version 6.2.10. (retrieved 17 March 2022). [Online]. Available: http://www.praat.org/

P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in Proc. ICPhS, vol. 17, no. 1193, 1993, pp. 97–110.

F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE – the Munich versatile and fast open-source audio feature extractor,” in Proc. 18th ACM Multimedia, 2010, pp. 1459–1462.

H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Iribo, and H. Banno, “Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” in Proc. ICASSP, 2008, pp. 3933–3936.

H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, H. Banno, and T. Iribo, “A unified approach for F0 extraction and aperiodicity estimation based on a temporally stable power spectral representation,” in Proc. ISCA ITW on Speech Analysis and Processing for Knowledge Discovery, 2008, p. paper 043.

M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” JEICE Trans. Information and Systems, vol. 99, no. 7, pp. 1877–1884, 2016.

H. Kawahara, Y. Agiomyrgiannakis, and H. Zen, “Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis,” in 9th ISCA Speech Synthesis Workshop, 2016, pp. 221–228. [Online]. Available: http://dx.doi.org/10.21437/SSW.2016-36

H. Fujisaki, “Analysis of voice fundamental frequency contours for decorative sentence of Japanese,” J. Acoust. Soc. Jpn. (E), vol. 5, no. 4, pp. 133–142, 1984.

H. Kawahara. GitHub repository for speech and hearing research/education tools. (retrieved 21 March 2022). [Online]. Available: https://github.com/HidekiKawahara