OBJECTIVE MEASUREMENT OF PITCH EXTRACTORS’ RESPONSES TO FREQUENCY MODULATED SOUNDS AND TWO REFERENCE PITCH EXTRACTION METHODS FOR ANALYZING VOICE PITCH RESPONSES TO AUDITORY STIMULATION

Hideki Kawahara* K. Yatabe† K. Sakakibara‡ T. Kitamura§ H. Banno§ M. Morise§

*Wakayama University, Japan †Waseda University, Japan ‡Health Science University of Hokkaido, Japan
§Konan University, Japan ¶Meijo University, Japan

*kawahara@sys.wakayama-u.ac.jp, †k.yatabe@asagi.waseda.jp, ¶kis@hoku-iryo-u.ac.jp
‡t-kitamura@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. The method simultaneously measures the linear and the non-linear time-invariant responses and random and time-varying responses. It uses extended time-stretched pulses combined by binary orthogonal sequences. Our recent finding of involuntary voice pitch response to auditory stimulation while voicing motivated this proposal. The involuntary voice pitch response provides means to investigate voice chain subsystems individually and objectively. This response analysis requires reliable and precise pitch extraction. We found that existing pitch extractors failed to correctly analyze signals used for auditory stimulation by using the proposed method. Therefore, we propose two reference pitch extractors based on the instantaneous frequency analysis and multi-resolution power spectrum analysis. The proposed extractors correctly analyze the test signals. We open-sourced MATLAB codes to measure pitch extractors and codes for conducting the voice pitch response experiment on our GitHub repository.

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

1. INTRODUCTION

We reported that voice fundamental frequency \( f_0 \) involuntarily shows compensatory response to frequency modulation of auditory stimuli applied while voicing \( [2] \). The experiment requires to measure \( f_0 \) of periodic sound \( [3] \) without introducing nonlinearities and glitches for measuring voice response to auditory stimulation. This requirement led us to propose an objective measurement method of pitch extractors’ response to frequency-modulated tone \( [4] \). We assigned the modulation signal as input and the extracted pitch value as output and fed them to the proposed pitch extractor measurement method. We found that existing \( f_0 \) extractors fail to meet this requirement. This issue motivated us to introduce reference pitch extractors.

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 also introduced two reference pitch extractors.

2. BACKGROUND

The first author introduced an altered auditory feedback technique \( [4] \) and reported that our voice pitch control consists of two responses to feedback pitch modification, the voluntary and the involuntary responses a quarter-century ago \( [5,6] \). Despite decades of research, altered feedback still is a hot topic for investigating speech chain and underlying neural basis \( [7,11] \). This research mainly focused on voluntary responses represented by the pitch shift paradigm and adaptation paradigm \( [12,13] \). Introduction of CAPRICEP-based method \( [3] \) and voice pitch response to auditory stimulation, which is not an altered feedback voice, opened a new possibility \( [2] \). The combination of the method and response to non-feedback sounds solved difficulties in investigating the involuntary response \( [2,15] \).

The experimental procedure for measuring the voice pitch response to FM test sounds consists of the following steps.

Generation of test signal: Combine orthogonal sequences made from extended time-stretched pulses followed by smoothing to yield the modulation signal. We applied frequency modulation to four types of signals. They are; SINE: a sinusoid, SINES: a sum of multiple harmonic sinusoids, MFND: a sum of multiple harmonic sinusoids without the fundamental component, and MFUNDH: a sum of multiple harmonic sinusoids without the first eight harmonic components.

Voicing with auditory stimulation: The subject keeps voicing with a constant pitch while listening to the test sound. The test system records the produced voice and the test signal for auditory stimulation simultaneously.

Response analysis: Apply pulse recovering and orthogonalization procedure to recover the stimulation pulse from the test signal

\*This work was supported by JSPS (Japan Society for the Promotion of Science) Grants-in-Aid for Scientific Research Grant Numbers JP18K00147, JP19K10708, JP19K21618, and JP21H04900.

\*The term “altered auditory feedback” is here because it is common practice now. We used the term “transformed auditory feedback” to represent our paradigm a quarter-century ago.

\*For detailed historical background and discussions on altered feedback research and relation to CAPRICEP-based method, refer to [2,13].
We applied band-pass filtering using an analytic signal to extract the sine worked fine with the machine epsilon. We use Gaussian windows throughout this paper because it simplifies discussion and derivation.

Note that Harvest [28] shows the non-LTI response close to the lower bound.

Because FM is a non-linear operation, and its sideband is not bandlimited. The simulation signal is a Gaussian smoothed mixture of three CAPRICEP-based orthogonal signals. The modulation depth (standard deviation) made pitch extraction using an analytic signal impossible. These difficulties (missing fundamental and significant deviation) let us test available pitch extractors by replacing the analytic-signal-based procedure. The objective measurement method of pitch extractors is our first contribution.

3. OBJECTIVE MEASUREMENT METHOD

Pitch extraction has long research history [16–25] and still is a hot topic in speech processing [26–30]. We propose to apply the CAPRICEP-based method [3] to measure the responses of pitch extractors. We replicated the FM-based procedure introduced to measure the voice pitch response to auditory stimulation [24, 25].

Figure 1 shows the scheme of the proposed method. The combination of elements test signal generator, response analyzer, and CAPRICEP-set and orthogonal matrix is the CAPRICEP-based method [3]. We used a Gaussian smoothing LPF and fed the output to VFO (Variable Frequency Oscillator) to modulate $f_o$ of the test signals. We normalize the pitch extractor’s responses by the LTI response of the lower path in Fig. 1 to calculate the pitch extractor’s responses. The pitch extractor’s response should only consist of the LTI response and not introduce the non-LTI or random and time-varying response.

Figure 2 shows response examples of representative pitch extractors. The input signal is a frequency modulated synthetic male vowel /a/ with 44100 Hz sampling and 120 Hz average $f_o$. The modulation signal is a Gaussian smoothed mixture of three CAPRICEP-based orthogonal signals. The modulation depth (standard deviation) is a 25 musical cent. In Fig. 2 the graph with the title showing “ND2” uses modified autocorrelation with LPC-based spectral whitening [19]. “CREPE120sim” uses convolutional neural network [29], and “Harvestsim” represents the dedicated pitch extractor [28] for a popular high-quality VOCODER WORLD [27].

The tests, including other pitch extractors, were disappointing. The pitch extractors for WORLD [29] and legacy-STRAIGHT [24] showed the best results (the lowest non-LTI and lowest random and time-varying response levels). However, many failed to handle the original test signal (sum of multiple harmonic components with the same amplitude). We needed to shape the spectral envelope of the test signals using a Japanese vowel /a/ extracted by WORLD for tested pitch extractors to work correctly. Even with this modification, no extractor succeeded in analyzing missing fundamentals.

These results reveal that response measurement of pitch extractors using the CAPRICEP-based method provides an objective and informative means for evaluating pitch extractors in addition to commonly used performance measures (GPE, FPE, FFE, and others). This measurement method is our first contribution and motivated us to propose the second contribution, reference pitch extractors.

4. REFERENCE PITCH EXTRACTORS

We introduce two reference pitch extractors to analyze test signals and voice pitch responses correctly. The first one uses instantaneous frequency for representing and extracting the fundamental component. The second one uses periodicity on the frequency axis for representing repetition.

4.1. Instantaneous frequency-based implementation

We already know that once we find the fundamental component, the instantaneous frequency of the component provides the $f_o$ value. We introduce an automatic mechanism to select the fundamental component without prior knowledge about the true $f_o$ value.
The proposed instantaneous frequency-based algorithm uses filters having the same shape on the logarithmic frequency axis. When the bandpass filter covers 0 Hz to 2f_c Hz (where f_c is the center frequency), a filter consisting of the fundamental component has the minimum amount of total AM and FM in its output \[22, 31\]. In addition to the direct implementation of a filter bank, we briefly introduce another implementation using a combination of short-time DFT (defined by the lowest pitch f_L) and matrix multiplication.

### 4.1. Filter design and implementation

Let x|_{n_0} represent the DFT of the windowed signal segment aligned at discrete time n_0. The following equation calculates the filter outputs (y|_{n_0}) at a discrete time t_{n_0} from x|_{n_0}.

\[
y|_{n_0} = W_{n_0} x|_{n_0},
\]

where \( k \)-th element of y|_{n_0} is the output of the \( k \)-th filter, and \( A^T \) represents transposition of matrix A. The parameters \( \sigma_k \) to define the band width of a \( k \)-th filter uniquely corresponds to the level \( \theta_k \) at frequency 0 and \( 2f_c \) by \( \sigma_k^2 = -2\pi^2 f_c^2 / \log(\theta_k) \). The parameter \( \sigma_k \) (defined using f_L) is for the short-time DFT at each analysis frame.

#### 4.1.2. Instantaneous frequency and representative values

Let matrix \( X|_{n_0} \) consists of short time DFT vector sequence \( x|_{n_0}, x|_{n_1}, x|_{n_2}, \ldots, x|_{n_N} \), starting from \( n_0 \) to \( n_0+N \) with a 1/f_s increment. The matrix \( Y|_{n_0} \) consists of the output vector sequence.

\[
Y|_{n_0} = W_{n_0} X|_{n_0}.
\]

Following equation yields matrix \( F|_{n_0} \). The element \( (F|_{n_0})_{k,m} \) is the instantaneous frequency of the \( k \)-th filter at \( t_{n_m-1} \).

\[
F|_{n_0}^{N+1} = \frac{f_L}{2\pi} \angle \left[ Y|_{n_0}^{N} \odot Y|_{n_0}^{N-1} \right],
\]

where function \( \angle[z] \) yields the angle (argument) of a complex number \( z \), and \( \odot \) represents Hadamard division. A notation \( A|_{n_0}^{b} \) represents the submatrix of \( A \) consisting of columns from \( a \) to \( b \). We use the sample average \( \bar{F}|_{n_0} \) of \( F|_{n_0} \) for representative value of the analysis frame, where \( f|_{n_0} \) is the \( m \)-th column of the matrix \( F|_{n_0} \).

#### 4.1.3. AM and FM deviation measure and calibration

For estimating total amount of AM and FM deviation \( \sigma_E \) we use sample standard deviation of the instantaneous frequency difference \( s_f|_{n_0} \) that of the relative absolute value change difference \( s_a|_{n_0} \).

\[
\sigma_E^2 = s_f^2|_{n_0} + s_a^2|_{n_0},
\]

where \( s_f|_{n_0} = \left( \frac{2\pi f_L}{N} \sum_{m=n_0}^{n_0+N} \left| f|_{n_0} - \bar{f}|_{n_0} \right|^2 \right)^{\frac{1}{2}}, \]

\[
s_a|_{n_0} = \left( \frac{f_L}{N} \sum_{m=n_0}^{n_0+N} \left| a|m|_{n_0} - \bar{a|m|}_{n_0} \right|^2 \right)^{\frac{1}{2}},
\]

where vector \( a|m|_{n_0} \) is \( m \)-th column of matrix \( A|_{n_0}^{N+1} \).

[footnote:Note that the segment center \( x|_{n_0} \) is at the beginning of the DFT buffer. The discrete time and the frequency of DFT buffer are circular. The first part of DFT buffer of \( f_k \) is \( f_0, f_1, \ldots, f_{L_{DFT}/2} \). The last part of \( f_k \) is \( f_{L_{DFT}/2+1}, f_{L_{DFT}/2+2}, \ldots, f_{-1} \). The discrete time \( t_0 \) has the same ordering. The windowed input signal element \( x|_{n_0} \) aligns at \( t_0 \).]

#### 4.2. Repetitive structure-based implementation

The instantaneous frequency-based method cannot find the fundamental component for missing fundamentals. The second algorithm, the repetitive structure-based method, can. Power spectra of periodic sounds have periodic variation. Removing envelope from a power spectrum of a voiced sound and calculating inverse Fourier transform yields a modified autocorrelation with local peaks at integer multiples of the fundamental period. This is one of common strategies of many pitch extractors \[17, 21\]. We revisit this idea (selecting \[20\] and \[21\] for starting point) using Gaussian window and make it \( f_0 \) independent self tuning method.

Let \( x \) represent a signal segment. For making it self tuning, we start from a window \( w_t \) having a narrower frequency resolution than half of the lowest (possible) \( f_0 \). We also use a window \( w_r \) on a lag axis. The following equations calculate the power spectrum \( p \) and smoothed power spectrum \( \tilde{p} \) via smoothing by weighting the autocorrelation \( v \) using the lag window \( w_r \).

\[
p = |\mathcal{F}[w_t \odot x]|^2, \quad v = \mathcal{F}^{-1}[p], \quad \tilde{p} = \mathcal{F}[w_r \odot v],
\]

where \( \odot \) represents Hadamard product.

#### 4.2.1. Recursive design and analysis

The following equations recursively define a set (Gaussian) time windows and lag windows. Their durations (for example, defined

\[11\]Note that different \( \theta_{c} \) setting introduces bias. Lower \( \theta_{c} \) value makes linear relation saturate in the high SNR end.

\[12\]x, w_r, and w_r are column vectors on the discrete time and lag axes.
where $f_0$ represents the fundamental period. Combining this information and instantaneous frequencies of harmonic components gives the correct $f_0$ value for missing fundamental test signals. Pitch extraction of missing fundamentals is the task that no existing pitch extractors succeeded. These two extractors are the second contribution.

5. DISCUSSION

We found that some pitch extractors showed surprisingly poor results using our proposed test method. We placed test signals and test codes in our GitHub repository [33] and made anyone be able to verify. These poor results are possible because they consist of pre-and post-processes specialized for their target applications and these processes are usually non-linear. Our proposed reference extractors do not have such processes and leave that specialized processing for users’ customization. To illustrate example cases of applying the measurement method and the reference pitch extractors, we prototyped an interactive and real-time tool for visualizing vocal tract shapes using the simplified versions of the reference pitch extractors. Our proposed test method was beneficial for implementing these simplified extractors.

Our approach using Gaussian windows without sidelobes is inefficient because of its excessive window length. Using windows without sidelobes (effectively) and have more compact support [44][36] for the short time DFT at the first stage provides the answer. However, we leave substantiation of this solution for further research.

Figure 5 shows a snapshot of an EGG (Electroglottograph) database inspector under development. We made heat maps of the AM and FM total deviation measure and the periodicity measure. With heat maps and interactive visualization, they were informative for detailed inspection. For example, around the green cursor line, it shows that the pitch period alternates short and long in each cycle.

6. 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 LTI, the non-LTI, and random and time-varying responses. Unexpectedly poor existing pitch extractors’ performance motivated us to introduce two reference pitch extractors, one uses instantaneous frequency, and the other uses frequency domain periodicity. We placed the code and the data on our GitHub repository [33] and made it open-sourced.

7. REFERENCES

[1] Ingo R. Titze, R. J. Baken, and K. W. Bozeman et al., “Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization,” J. Acoust. Soc. Am., vol. 137, no. 5, pp. 3005–3007, 2015.
