Accurate estimation of $f_o$ and aperiodicity based on periodicity detector residuals and deviations of phase derivatives

Hideki Kawahara*, Ken-Ichi Sakakibara†, Masanori Morise‡, Hideki Banno§ and Tomoki Toda¶

* Wakayama University, Wakayama, Japan E-mail: kawahara@sys.wakayama-u.ac.jp
† Health Science University of Hokkaido, Hokkaido, Japan E-mail: kis@hoku-iryo-u.ac.jp
‡ University of Yamanashi, Yamanashi, Japan E-mail: mmorise@yamanashi.ac.jp
§ Meijo University, Aichi, Japan E-mail: banno@meijo-u.ac.jp
¶ Nagoya University, Aichi, Japan E-mail: tomo@icts.nagoya-u.ac.jp

Abstract—We introduce a new fundamental frequency ($f_o$) and aperiodicity estimation method with its certified calibration data on estimation accuracy. The proposed method is based on two independent representations. The first representation is a set of band-wise residuals of a self-tuning process to suppress the prominent periodic component in the band-pass filtered output. The second representation is a set of band-wise deviation measure of the instantaneous frequency of the band-pass filtered output. The combined measure of these representations provides accurate estimates of SNR of each filtered band. This combination significantly reduces the distribution spread from 0.5% to 99.5% probability. The reduction is about 5 dB (for example, 21 dB to 16 dB). A wavelet-based configuration provides the initial estimate of $f_o$. A short term Fourier transform-based configuration provides refinement of $f_o$ and accurate estimate of aperiodicity. A MATLAB implementation of the procedure will be available as an open-source package.

I. INTRODUCTION

Aperiodic components in speech sound play important roles in speech communication (normal, expressive, extreme and so on) and singing expression [1]-[4]. They contribute to the synthesized speech quality and intelligibility [5]-[8]. Despite their importance, reliable estimation of aperiodic components has been a challenging topic [9]-[15]. The major difficulty is the fact that such aperiodic components in speech are much weaker than the periodic components. We recently proposed to use deviations of the prominent periodic components to estimate the level of the aperiodic components [10], instead of directly extracting and measuring them [16]. The first author also proposed to use a self-tuning process to suppress the prominent periodic component to estimate background noise level as residuals [17]. We proposed to combine these two representations to yield more reliable aperiodicity measure. In addition to aperiodicity analysis, the measure is also applied to $f_o$ extraction and tracking.

This article is organized as follows. Background section introduces our motivation, starting with a brief review of $f_o$ trackers and proposes to build a measuring instrument of periodicity/aperiodicity with its certified calibration data. Next section proposes to use a mixed measure based on two sources of information about random components. Following a brief introduction to these two measures, we introduce a new cosine series function used in the second method and a simple implementation of phase derivatives. The third section introduces the architecture of three staged procedure of extracting $f_o$ and aperiodicity. The following section introduces numerical simulation using artificial test signals and examples of natural speech analysis. In the final section, application of the proposed method for generating reliable metadata to speech corpus and a high-quality vocoder are discussed. In summary, our primary contribution is that we built a measuring instrument of periodicity/aperiodicity with its certified calibration data. We attained the proposed goal.

II. BACKGROUND

The value itself and movement in time of $f_o$ play important roles in speech communication [1]-[4]. Because of this importance, reliable estimation of $f_o$ has been investigated based on various paradigm [10], [18]-[22]. In the majority of such investigations, manually assigned reference values and other physiological sources of observing the vocal fold vibration [23] have been used as the ground truth for tuning and evaluation of such procedures.

However, for tuning and evaluating high-precision procedures, these commonly used references are not reliable enough. Instead, we proposed to use a carefully designed test signals [24]. In addition to these, we also introduced a set of antialiased L-F model [25], [26] and Fujisaki-Ljungqvist model [27], [28]. These antialiased models have a strongly suppressed spurious level (less than -170 dB to the highest harmonic component), for precise tuning and evaluation of such procedures. These glottal source models are implemented as a part of a set of interactive tools for education and research on speech science [29], [30]. The tools also consist of a lattice architecture implementation of a one-dimensional acoustic tube, which is capable of parameter update at the audio sampling rate and simulate nonlinear phenomena in

This work was supported by JSPS KAKENHI Grant Numbers JP15H03207, JP15H02726 and JP16K12464.
phonation [31]. These are implemented as MATLAB functions and provided as a part of an open-source software [30].

The test signals mentioned and these antialiased glottal source models, together with time varying filter models, also provide a hierarchical set of ground truth for tuning and evaluating aperiodicity estimator. Our previous implementation of the antialiased L-F model [26] was extensively used to design an open source \( f_o \) tracker [17]. These preparation enable us to design \( f_o \) extractors and aperiodicity estimators and to provide certified calibration and accuracy data. We propose to revisit the goal of \( f_o \) extractors and aperiodicity estimators design, taking into account of complex nature of the phonation process [2], [31]–[33] and rapid advances in machine learning including deep learning [34]. Our answer is to change the design goal from building classifiers to building accurate measuring instruments with certified data sheets on calibration and performance test results.

III. MIXED MEASURE OF TWO SOURCES OF INFORMATION

This section proposes a mixed measure made from two different measures of the random component. The first measure is based on residuals of a self-tuning periodic component canceller [17]. The second measure is based on deviations of phase derivatives [16].

A. Residuals of periodicity detector

Assume a complex-valued band-pass filter which has a quadrature signal \( h(t) \) as its impulse response. Assume an input signal \( x(t) \) which has a prominent sinusoidal component \( x_p(t) \) and a random component \( x_r(t) \). When the periodic component is inside the pass-band of the filter, the absolute value of the output \( y_1(t) \) of the filter is mainly determined by the absolute value of the filtered sinusoidal component \( h(t) * x_p(t) \), where \(*\) represents convolution. Then, by normalizing \( y_1(t) \) using its absolute value yields approximately a complex exponential \( \exp(j\omega_c t) \), where \( \omega_c \) represents an unknown angular frequency of the periodic component \( x_p(t) \). Also by normalizing \( y_2(t) \), which is the filtered output of \( y_1(t) \), yields another approximate signal of \( \exp(j\omega_c t) \). Then, the variance of the residual \( r(t) \) defined below is proportional to the relative variance of the random component \( x_r(t) \), when \( V[x_r(t)] \ll V[x(t)] \).

\[
r(t) = \frac{y_1(t)}{|y_1(t)|} - \frac{y_2(t)}{|y_2(t)|} \\
\approx \frac{h(t) * x(t)}{|h(t) * x_p(t)|} - \frac{\hat{h}(t) * (h(t) * x(t))}{|\hat{h}(t) * (h(t) * x_p(t))|},
\]

where the real part of \( h(t) \) has even symmetry and the imaginary part has odd symmetry and bounded inside of the region \((-t_w/2, t_w/2)\). Note that this procedure does not require the prior information about the frequency of the prominent sinusoidal component to suppress the component.

The estimate of the relative level of the random component \( \eta_r(t) \) is as follows:

\[
\eta_r^2(t) = c_r \langle |r(t)|^2 \rangle,
\]

where \( c_r \) represents the calibration coefficient and \( \langle Q \rangle \) represents the average of the representation \( Q \).

This framework is general enough to apply to Fourier transform-based methods as well as wavelet transform-based methods. The analysis part of the YANG vocoder [17] uses a wavelet-based method for its initial \( f_o \) estimation stage and uses a Fourier transform-based method for its refinement and aperiodicity estimation stage. Note that the envelope of \( h(t) \) used in YANG vocoder is the 11-th item of Table II of Nutall’s reference [35]. We have conducted a systematic test including a six-term cosine series defined by Eq. 5. Comparison details are shown in APPENDIX-B.

B. Deviation measure of phase derivatives

Similar to the first method, assume a complex-valued band-pass filter which has a quadrature signal \( h(t) \) as its impulse response. Assume an input signal \( x(t) \) which has a prominent sinusoidal component \( x_p(t) \) and a random component \( x_r(t) \). When the random component is negligibly small, the instantaneous frequency of the filter output is equal to the frequency of the prominent sinusoidal component, as far as the prominent component resides inside of the pass-band. In other words, the filter center frequency derivative of the output instantaneous frequency is zero.

The variance of this derivative is proportional to the variance of the random component. However, because the derivative is a real-valued time series with zero-mean, it vibrates around zero. Roughly speaking, the absolute value of the time derivative of the frequency derivative has a minimum when the frequency derivative is maximum or minimum. It has a maximum value when the frequency derivative crosses zero. The mean squared value of the frequency derivative and the time derivative of the frequency derivative represents the magnitude of deviations, which is proportional to the background random component level on average.

The estimate of the relative level of the random component \( \eta_f(t) \) is as follows:

\[
\eta_f^2(t) = c_f \left( \frac{1}{\omega_c} \left| \frac{d\omega_c(t)}{dt} \right|^2 \right)^2 + c_{mix} \left( \frac{1}{\omega_c} \left| \frac{d^2\omega_c(t)}{dt^2} \right|^2 \right)^2,
\]

where \( c_f \) represents the calibration constant for SNR estimation, and \( c_{mix} \) represents the mixing-constant to equalize distribution. Note that this measure is insensitive to amplitude modulation (AM) because it starts from the frequency derivative of the instantaneous frequency. AM introduces deviation with even symmetry and does not change the first order coefficient of Taylor expansion.

The practical implementation uses power spectrum weighted averaging. The next subsection introduces a new cosine series for designing the envelope of the filter impulse response. It has strong attenuation (-114.24 dB) of the highest sidelobe level and steep decay rate (54 dB/oct) of sidelobes. It is because the time derivative in the instantaneous frequency definition makes the most localized windowing functions (prolate spheroidal wave function [36] and its approximation, Kaiser window [37]), which have slow decay rate (6 dB/oct), irrelevant.
APPENDIX-A details on the importance of using the appropriate windowing function, the proposed \( w(t) \).

1) A cosine series window: We introduced a new cosine series function for designing the impulse response of the band-pass filter [16]. The function was initially designed for anti-aliasing glottal source models [39]. The following equation defines the function:

\[
w(t) = \sum_{k=0}^{5} h_k \cos \left( \frac{k \pi t}{t_w} \right)
\]

(4)

where \( t_w \) represents the half length of the windowing function and for \( |t| \geq t_w \), the value of \( w(t) \) is zero. The procedure for designing this function is based on Nuttall’s cosine window design [35]. Note that the residual-based method uses the 11-th item in Table II of the reference [35].

2) Implementation of phase derivatives: The following equations provide a simpler implementation of the instantaneous frequency of a complex valued discrete signal \( x[n] \).

\[
\omega_i[n] = \angle \left[ \frac{x[n+1]}{x[n]} \right] f_s,
\]

(6)

where \( f_s \) represents the sampling frequency.

Similarly, the following equation provides a simple implementation of group delay \( \tau_g[k] \) of a discrete spectrum \( X[k] \).

\[
\tau_g[k] = -\angle \left[ \frac{X[k+1]}{X[k]} \right] \frac{1}{\omega_s}
\]

(7)

where \( \omega_s \) represents the angular frequency difference of neighboring discrete frequency bins.

The Flanagan’s equation [40] and the similar group delay equation can be used when the calculation of inverse trigonometric functions are slow.

C. Mixed measure

Two measures for estimating the level of random components use different principles. The residual based measure depends on power directly while the deviation based measure depends on phase deviations, which are not directly depending on power. Therefore, it is possible to provide more reliable measure \( \eta_{\text{mix}}(t) \) by mixing these two measures: It is defined as follows.

\[
\eta_{\text{mix}}^2(t) = \frac{1}{2} (\eta_r^2(t) + \eta_i^2(t)),
\]

(8)

where no adjustable mixing coefficient is needed because each measure is already calibrated. Note that this mixed measure is proportional to the local SNR and therefore proportional to the standard deviation of the observed instantaneous frequency of the prominent component, in the case of the fundamental component, the standard deviation of \( f_o \). This highly-linear estimator is the main contribution of this article.

IV. Architecture

The proposed system for \( f_o \) and aperiodicity estimation consists of three subsystems, similar to the analysis part of YANG vocoder [17]. The first stage calculates a periodicity map using wavelet arrangement of band-pass filters. The second stage extracts the initial \( f_o \) estimate using the periodicity map. The final stage refines \( f_o \) using \( f_o \) adaptive time warping recursively and estimates the aperiodicity as a map in the final iteration. The following subsections introduce these subsystems.

A. Periodicity map

This stage extracts a set of \( f_o \) candidate frequencies \( f_k[n] \) and associated standard deviations \( \sigma_k[n] \) for each analysis frame \( n \). Each \( f_o \) candidate corresponds to the local dip of the mixed measure \( \eta_{\text{mix}} \). All filter output instantaneous frequency and the mixed measure inside each basin around each \( f_o \) candidate are averaged using the following equation:

\[
f_k[n] = \frac{\sum_{j \in B_k} \beta_j[n] \log(f_j[n])} {\sum_{j \in B_k} \beta_j[n]}
\]

(9)

\[
\sigma_k[n] = \left( \frac{1}{\#(B_k)} \sum_{j \in B_k} \beta_j[n] \right)^{-\frac{1}{2}}
\]

(10)

where \( B_k \) represents the set of channel indices of the \( k \)-th basin of the \( n \)-th frame and \( \#(B_k) \) represents the cardinal number of the set \( B_k[n] \). We defined as \( \beta_j[n] = 1/\sigma_j^2[n] \) for simplicity.

The probability map representation visualizes these candidates. The channel probability of \( j \)-th channel represents the sum of the probability of each candidate to reside in the \( j \)-th channel. We use sigmoids to approximate error functions for computational efficiency.

B. Initial \( f_o \) estimation

Tracking the best candidates using the standard deviation of each candidate and the transition cost matrix provides the best candidate trajectory. The tracking uses the Viterbi algorithm with some heuristic extension. Each frame of this best candidate sequence provides the frequency and the SNR at each frame. This procedure yields the initial estimate of the \( f_o \) trajectory.

C. Refinement and aperiodicity

In the final stage, the frequency information of the initial \( f_o \) trajectory provides stretching coefficients for adaptive time axis warping for making apparent \( f_o \) constant. The FFT-based procedure combined with down sampling provides efficient calculation. The final stage uses the original sampling frequency to calculate the full-range aperiodicity information.

V. NUMERICAL EXAMPLES

This section uses examples to illustrate the behavior of component systems. First section tests effect of mixing using simple test signals.
A. Effect of mixing

The mixture of periodic pulse train signals and Gaussian white noise are used here. The sampling frequency is 44,100 Hz, and two \( f_o \) frequencies, 120 Hz, and 240 Hz are tested. SNR levels from 0 dB to 80 dB in 10 dB steps are used. For making simulation tests efficient, we used an FFT-based implementation and gathered results from all harmonic frequencies. The test signals have the same local SNR for all harmonic frequencies, and the output instantaneous frequencies and residuals showed the same statistical behavior. Test results using wavelet-based filter allocation also showed the same behavior for the same center frequency filter. Instantaneous frequencies of wavelet allocation are shown in APPENDIX-A.

Figure 1 shows how mixing makes the width of the distribution of the estimated SNR narrow. The horizontal axis represents the SNR of the test signal. The vertical axis represents the difference of the estimated SNR between given reference points on the cumulative distribution. Mixing two measures reduces the distance about 5 dB for the distance between 0.5% and 99.5% points.

Figure 2 shows the aperiodicity level \( AP[n] \) of the \( n \)-th frame defined by the following equation:

\[
AP[n] = 10 \log_{10} \left( \frac{V[x_r[n]]}{V[x_p[n]] + V[x_{sp}[n]]} \right),
\]

where \( V[x] \) represents the variance of \( x \). The vertical axis represents the probability of the estimated value \( AP[n] \) is smaller than the value on the horizontal axis \( AP_0 \) for each given input SNR:

\[
P_i(AP[n] \leq AP_0; SNR_{input}),
\]

where from right to left, \( SNR_{input} \) levels are 0 dB to -80 dB in -10 dB steps. The plot shows results for the YANG vocoder and the deviation-based method and the mixed measure.

Figure 1 and Fig. 2 illustrate the main contribution of our work. The proposed method provides significantly fewer outliers and narrower distribution of estimated value. Because the proposed method is a fundamental building block, it is straightforward to add statistical methods [41] on top of this procedure.

B. Natural speech example

We use the same speech sample used in YANG vocoder’s analysis subsystem to illustrate how each component works. It is a Japanese vowel sequence /aiueo/ spoken by a male talker. The sampling frequency was 22,050 Hz and an omnidirectional condenser microphone was used for recording.

1) Instantaneous frequency and aperiodicity map: Figure 3 shows the results of the initial wavelet analysis. The upper image shows instantaneous frequency of each filter output. The filter allocation density was 12 channels/oct, while the image shows maps with 48 channels/oct, for avoiding disturbing “jaggy-ness.” Each image is associated with the time-aligned waveform. The aperiodicity map has a big blue area from 0.1 s to 0.7 s, which corresponds to the region where the fundamental component is prominent.

2) Fundamental component candidates: Figure 4 shows the extracted candidates of the fundamental component. The upper plot shows the time-frequency allocation of the candidates with color coding of the associated standard deviation of the estimated frequency. The color coding is as follows; yellow for \( \sigma_k[n] \geq -20 \), cyan for \( \geq -30 \), dark green for \( \geq -40 \), blue for \( \geq -50 \), and black for \( < -50 \) (dB). The lower plot shows the estimated standard deviation \( \sigma_k[n] \) of each candidate in terms of dB, where 0 dB corresponds to 100% relative error in the instantaneous frequency value. In this example, the most likely candidates from 0.15 s to 0.65 s have the estimated RMS (root mean squared) error less than 1%. Practically, this level of error does not need further refinement.

3) Probability map for summary visualization: Figure 5 shows the visualization of local probability. The value in each bin represents the sum of the probability of each candidate
to be located inside the bin. Also for visualization, 48 channels/oct and 1 ms frame shift are used. The information in this representation is a summary of the two figures shown in Fig. 4. This probability visualization is relevant for interactive investigations on vocalization. The discrete set of candidates is relevant for tracking and refinement of \( f_0 \) and aperiodicity analysis.

4) Aperiodicity map based on deviation measure: Figure 6 shows the aperiodicity map using deviation-based measure. This time, the initial candidate \( f_0 \) values are used to warp the time axis. The target \( f_0 \) of this warping was set to 40 Hz.

VI. DISCUSSION

We proposed an accurate method of \( f_0 \) and aperiodicity estimation based on two different underlying principles, one is based on residuals of a periodicity detector the other is based on derivatives of instantaneous frequency. These two sources are based on independent aspects of the background noise. Consequently, combining these two sources of information
makes the resultant aperiodicity measure more reliable than each component.

Note that the proposed method is general enough to apply to any signal. For example, using the derivative of the Hilbert envelope of a band-pass filtered speech signal for the input to this procedure will provide a noise robust \( f_0 \) extractor. Harmonic frequencies associated with their estimated SNR also can be used for Kalman-filter based \( f_0 \) tracker [41].

We did not discuss the temporal distribution of the random component within one pitch period, but it is an important issue [11], [14]. It has to be properly estimated because the temporal location of a noise burst has a significant impact on the perceived noise level [42]. A weighted average of group delay with minimum phase compensation [43] can apply to this issue by revisiting derivation of the group delay equation and windowing function design. As we pointed out in our previous paper [16], this is the next research target of aperiodicity analysis.

VII. CONCLUSIONS

We introduced an accurate fundamental frequency (\( f_0 \)) and aperiodicity estimation method using two representations based on independent underlying principles. The method applies to both wavelet and Fourier transform based analysis front ends. The wavelet-based implementation is useful for extracting the fundamental component candidate. The Fourier transform-based implementation is useful for \( f_0 \) refinement and aperiodicity analysis. The proposed method will be applied to provide reference \( f_0 \) information to CSJ corpus [44]. After tests and refinement, we are planning to introduce the proposed method into our open-source systems [30], [45].

REFERENCES


Fig. 6. Estimated aperiodicity using the mixed measure.
Fig. 7. Transfer functions of the band-pass filters, which are designed using the following functions for their envelope. The functions are; from higher sidelobe levels around 300 Hz in decreasing order: Hamming, Blackman, Kaiser, prolate spheroidal wave function, the proposed six-term cosine series $w_4(t)$. (Excerpted from [46] with permission)

APPENDIX-A: SELECTION OF WINDOWING FUNCTION

This appendix illustrates the effect of windowing function $w(t)$ on calculating the instantaneous frequency. This appendix is an excerpted and translated version of the first author's short review [46]. Assume a quadrature signal, which uses the windowing function $w(t)$ for the envelope of a complex exponential $\exp(j\omega_c t)$, where $\omega_c$ represents the center frequency of the band-pass filter. The pass-band of the filters are designed to separate the fundamental component with enough margin. The bandwidth of each windowing function is adjusted to match each other by setting its duration defined using the second moment the same.

Figure 7 shows transfer functions of band-pass filters, designed using typical windowing functions. There are three group of windowing functions regarding the decay rate of sidelobes. The first group has a decay rate of 6 dB/oct. The members are; Hamming, Kaiser, and prolate spheroidal wave function. The second group has a decay rate of 18 dB/oct. The member is Blackman. The third group has a decay rate 54 dB/oct. The member is the proposed six-term cosine series. The prolate spheroidal wave function was adjusted to make the maximum sidelobe level match with the proposed six-term cosine series. The parameter of Kaiser window was set to have the same sidelobe level to the prolate spheroidal wave function. We prepared three impulse responses. The original and its dilated versions, $2^{-\frac{7}{6}}$ times and $2\pi$ times in length.

A.1: Calculation of instantaneous frequency

We generated a test signal using a periodic 100 Hz pulse train with a sampling frequency 44,100 Hz. The test signal was fed to each filter and calculated the instantaneous frequency of the filter output. For the actual calculation of the instantaneous frequency, we used the Flanagan’s equation here. The simple
discrete implementation by Eq. 6 provides effectively the same results.

Figure 8 shows the instantaneous frequency of each filter output. The filter output using the original impulse response yields the same instantaneous frequency to \( f_o \). Filter outputs using dilated impulse responses yield deviating instantaneous frequency from \( f_o \). Except for the proposed function, all functions show deviations with sharp edges. The magnitude of deviations is higher for the functions having smaller sidelobe level attenuation. However, the magnitude of deviations of Kaiser and the prolate spheroidal was function were approximately ten times larger than the proposed function, even though they have comparable highest sidelobe levels to the proposed function. It is because their sidelobe decay, 6 dB/oct is canceled by the derivation in time, which defines the instantaneous frequency. These sharp edges are not caused by their implementation. Instead, they are the result of the definition of instantaneous frequency and are inevitable. Note that Blackman window, which has 18 dB/oct sidelobe decay rate still has sharp edges.

### APPENDIX-B Envelope functions for Residual-based method

This section revisits the residual-based method [17].

#### B-1 Discrete time formulation

Assume that signals are discrete time signals and take Discrete Fourier transform of the procedure.

\[
R[k] = \frac{H[k]X[k]}{H[k]} - H[k]X[k],
\]

where \( X[k], H[k] \) and \( R[k] \) are the Discrete Fourier transform of \( x[n], h[n] \) and \( r[n] \), respectively. Assume \( x[n] = x[n+N] \), where \( N \) represent the period of the cycle. It also holds for \( r[n] \). For using FFT, the envelope of the impulse response \( w[n] = |h[n]| \) is used.

The windowing function in the discrete time domain \( w[n] \) is an even function. It means \( w[n] \) is defined in \([-n_w, n_w]\). It has to be placed in an FFT buffer on \([0, \ldots, n_w]\) and \([N-n_w, \ldots, N-1]\). However, this is sometimes confusing and inconvenient for implementation.

Common practice is to locate it from 0 to \( 2n_w \). Fourier transform of this \( W_5[k] \) introduces phase rotation (b stands for bias in phase). It can be compensated by dividing it by the Fourier transform of the unit impulse located at \( n_w \).

#### B-2 Performance determining parameters

The output of the procedure is the residual component level \( \sigma_r \), which is defined as follows:

\[
\sigma_r^2[k_p] = \langle |X[k]|^2 |R[k]|^2 \rangle_{k=k_p},
\]

where \( \langle |X[k]|^2 |R[k]|^2 \rangle_{k=k_p} \) represents weighted averaging in the discrete frequency domain around one of harmonic frequencies \( k_p \).

Specifically, it is written as follows:

\[
\langle |X[k]|^2 |R[k]|^2 \rangle_{k=k_p} = \frac{W_5[k] o |X[k]|^2 |R[k]|^2}{W_5[k] o |X[k]|^2} |_{k=k_p},
\]

where \( W_5[k] \) represents the frequency domain smoother defined below:

\[
W_5[k] = 0.5 + 0.5 \cos \left( \frac{\pi k}{c_{mag}k_m} \right),
\]

where \( k_m \) represents the modified target fundamental frequency and \( c_{mag} \) represents the magnification coefficient which controls the width of this smoother.

The other parameter is the stretching coefficient \( c_{dil} \) of the impulse response. In this implementation, the half length \( n_w \) is defined as follows:

\[
n_w = \text{round}\left[ c_{dil}n_o \right],
\]

where \( n_o \) represents the fundamental period, and \( \text{round}[a] \) is a rounding function to output the nearest integer of \( a \). The modified fundamental frequency is defined as follows:

\[
k_m = \frac{f_o}{n_w},
\]
B.3 Numerical tests

Figure 9 shows maps using two measures for optimization, calculated using the six-term cosine series. The left map shows deviations from the linear relation between the given SNR and median of the resulted residual level. The right plot shows the map of the distance between 0.5% to 99.5% probability points of the resulted residual level for the test signal with 30 dB SNR. The horizontal axis represents the smoother width $c_{\text{mag}}$ and the vertical axis represents the temporal stretching factor $c_{\text{dil}}$.

After preliminary parameter tuning and calibration, the cumulative probability distribution of the estimated residual level (RLV (dB)) was tested. The test signals are a pulse train plus Gaussian white noise signals with different SNR. The sampling frequency was 44,100 Hz, and the SNR levels were 0 dB to -80 dB in 10 dB steps.

Figure 10 shows linearity comparison of different envelope functions for the residual-based method. The original proposal used Nuttall’s 11-th window. These results indicate that the six-term cosine series used for deviation-based method also provides the best linearity. Currently, we are revising all procedures using this cosine series. The natural speech analysis example uses this revised implementation.

Fig. 9. Performance measure maps for the six-term cosine series. Left map shows deviations from the linear relation between the given SNR and median of the resulted residual level. The right plot shows the map of the distance between 0.5% to 99.5% probability points of the resulted residual level for the test signal with 30 dB SNR.

Fig. 10. Comparison of envelope functions using cumulative probability distribution of estimated residual levels (RLV (dB)). Envelope functions are: from top to bottom; Kaiser window, Blackman window, Nuttall’s 11-th item, and the newly designed six-term cosine series. The green lines represent the distribution before frequency smoothing. The black lines represent the distribution after smoothing.