Excitation source analysis for high-quality speech manipulation systems based on an interference-free representation of group delay with minimum phase response compensation

Hideki Kawahara[1], *Masanori Morise*[2], *Tomoki Toda*[3], *Hideki Banno*[4]
Ryuichi Nisimura[1], *Toshio Irino*[1]

[1]Faculty of Systems Engineering, Wakayama University, Wakayama, Wakayama, Japan
[2]Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu, Japan
[3]Nara Advanced Institute of Science and Technology, Ikoma, Nara, Japan
[4]Graduate School of Science and Technology, Meijo University, Nagoya, Japan

kawahara@sys.wakayama-u.ac.jp, mmorise@yamanashi.ac.jp, tomoki@is.naist.jp
banno@meijo-u.ac.jp, nisimura@sys.wakayama-u.ac.jp, irino@sys.wakayama-u.ac.jp

Abstract

A group delay-based excitation source analysis and design method is introduced for extension of TANDEM-STRAIGHT, a speech analysis, modification and synthesis system. This introduction makes all components of the system be based on interference-free representations. They are power spectrum, instantaneous frequency and group delay representations. This unification has potential to solve the major weak point of VOCODER architecture for high-quality speech manipulation applications.

Index Terms: speech analysis, speech synthesis, speech modification, VOCODER, interference-free representations

1. Introduction

The proposed group delay representation provides the constant value zero when the input signal is periodic, irrespective to analysis frame positioning. This article introduces the theoretical basis, performance evaluation and application to excitation source design, of this group delay representation.

High-quality speech manipulation systems, which also capable of precise control of physical parameters, need accurate separation of physically meaningful constituent parameters. STRAIGHT [1] and TANDEM-STRAIGHT [2] were designed to fulfill this requirement to provide flexible tools for exploratory research of speech [3]. Continuing expansion of their applications, such as morphing [4, 5, 6, 7, 8], made requirement on speech quality of the manipulated sounds more demanding and clarified weakness of the current representations used. The most crucial issue is excitation source representations, especially non-periodic components [9, 10, 11]. This article introduces a new group delay-based representation to provide a theoretically sound, computationally efficient, and practically useful basis to this issue.

2. Background and related works

A speech signal consists of a *quasi* periodic component and non-periodic components. This led to speech processing techniques based on various signal models. Source filter models provide a basis for VOCODER [12], LPC [13, 14], Cepstrum-based models [15, 16, 17] and STRAIGHT[1] [1, 2]. Sinusoidal models and their extensions [18, 19, 20] also provide bases for high-quality applications. It have been observed that introduction of relevant phase related information (explicitly or implicitly) into these models provides better perceptual quality of processed sounds [21, 18, 22, 17]. For highly flexible manipulations, for example morphing, simple parameterized signal models are desirable. At first glance, quality and flexibility are in trade-off. However, taking into account of perception of temporal fine structures [23, 24, 25], a simple pulse plus time-frequency shaped noise model may provide a counter example, based on the proposed new group delay representation. The proposed representation is applicable to both pulse or epoch [26] based models and sinusoid based models.

3. Interference-free representations

This section briefly summarizes three interference-free representations. Interference-free representation of power spectra of periodic signals [27] enabled separation of filter information and source information of speech sounds and provided the foundation of STRAIGHT. Interference-free representation of instantaneous frequency of periodic signals [28] provided F0 refinement procedure with fine temporal resolution and high-fidelity trajectory tracking [29]. Interference-free representation of group delay of repetitive signals [30], was introduced but was not been effectively used.

This article extends this group delay representation to be dually inteference-free, in other words, it does not have periodic variations both in the time and the frequency domain. Moreover, this extended representation yields constant zero for all frequency range, when the signal is periodic. Since all these representations share the same strategy, power spectral representation is discussed first.

3.1. Power spectrum

Let T_0 represent fundamental period of a periodic signal, the following equation provides power spectral representation

[1]STRAIGHT represents both STRAIGHT [1] and TANDEM-STRAIGHT [2] afterwards. When distinction is necessary, they are represented as legacy-STRAIGHT and TANDEM-STRAIGHT respectively.

$P_T(\omega,t)$, which does not have temporally varying component: [27, 2]

$$P_T(\omega,t)=\frac{P\left(\omega,t+\frac{T_0}{4}\right)+P\left(\omega,t-\frac{T_0}{4}\right)}{2}, \quad (1)$$

where $P(\omega,t)$ represent the short term power spectrum using a time window centered at time t. The main idea behind this is that the temporal variation of power spectra caused by the interference of adjacent harmonic components is sinusoid (cosine) of period T_0 and can be cancelled out by the component having the opposite polarity [27].

3.2. Instantaneous frequency

The following average of instantaneous frequencies $\omega_i(\omega,t)$ weighted by power spectra provides an instantaneous frequency representation $\omega_{iT}(\omega,t)$, which does not have temporally varying component: [28]

$$\omega_{iT}(\omega,t)=\frac{P^{(+)}\omega_i\left(\omega,t+\frac{T_0}{4}\right)+P^{(-)}\omega_i\left(\omega,t-\frac{T_0}{4}\right)}{P^{(+)}+P^{(-)}} \quad (2)$$

where $P^{(+)}$ represents $P\left(\omega,t+\frac{T_0}{4}\right)$ and $P^{(-)}$ represents $P\left(\omega,t-\frac{T_0}{4}\right)$. Note that the denominator of Eq. (2) is the interference-free power spectrum $P_T(\omega,t)$ defined by Eq. (1) multiplied by 2. Interference-free behavior is proven [28] by using Flanagan's instantaneous frequency equation [31].

3.3. Group delay: removing frequency interference

Group delay $\tau_d(\omega,t)$ is complementary to instantaneous frequency (for example [32]). This duality led to the following representation of group delay $\tau_{dF}(\omega,t)$, which does not have interferences in the frequency domain caused by multiple (this time two) events: [30]

$$\tau_{dF}(\omega,t)=\frac{P^{(U)}\tau_d\left(\omega+\frac{\omega_0}{4},t\right)+P^{(D)}\tau_d\left(\omega-\frac{\omega_0}{4},t\right)}{P^{(U)}+P^{(D)}}, \quad (3)$$

where $P^{(U)}$ represents $P\left(\omega+\frac{\omega_0}{4},t\right)$ and $P^{(D)}$ represents $P\left(\omega-\frac{\omega_0}{4},t\right)$. Periodicity interval $\omega_0=2\pi/T_0$ on the frequency axis is determined by the temporal interval T_0 between the events. Lengthy derivation of interference-free behavior of $\tau_{dF}(\omega,t)$ is given in [30]. Since group delay is the main topic of this article, outline of the derivation is given below.

The group delay is defined by the negative frequency derivative of the phase of $X(\omega,t)$, the short term Fourier transform of a signal. It is equivalent to calculate the derivative of the imaginary part of the log-converted short term spectrum $\log(X(\omega,t))$. It leads to the following computationally efficient equation:

$$-\tau_g(\omega,t)=\frac{\Re[X(\omega,t)]\Im[X_d(\omega,t)]-\Im[X(\omega,t)]\Re[X_d(\omega,t)]}{|X(\omega,t)|^2}, \quad (4)$$

where $X(\omega,t)$ and $X_d(\omega,t)$ are defined below:

$$X(\omega,t)=\int_{-\infty}^{\infty}w(\tau)x(\tau-t)e^{-j\omega\tau}d\tau \quad (5)$$

$$X_d(\omega,t)=\frac{dX(\omega,t)}{d\omega}=-j\int_{-\infty}^{\infty}\tau w(\tau)x(\tau-t)e^{-j\omega\tau}d\tau. \quad (6)$$

Note that the weights $P^{(U)}$ and $P^{(D)}$ in Eq. (3) cancel out with the denominator of Eq. (4) and that the denominator of Eq. (3) does not have periodic variation on the frequency axis. These make inspection on the denominator unnecessary. Substituting Eq. (4) to Eq. (3) and using the identity $(\sin^2\theta+\cos^2\theta=1)$ shows that the periodic variation of group delay on the frequency axis caused by multiple excitation effectively vanishes [30]. However, unlike power spectrum and instantaneous frequency, the proposed interference-free representation of group delay $\tau_{dF}(\omega,t)$ was not very successful in speech applications [30]. This inefficacy is caused by the huge dynamic range of speech spectra, because interference suppression requires that the denominator $P^{(U)}+P^{(D)}$ is changing smoothly and gradually in terms of ω. This is not the case for vowels.

3.4. Group delay: removing time-frequency interference

The interference-free representation of group delay $\tau_{dF}(\omega,t)$ defined by Eq. (3) still has periodic interference in the time domain when periodic signals are analyzed. Similar to the interference-free power spectra and instantaneous frequencies, calculating weighted average of $\tau_{dF}(\omega,t)$ calculated at two points $T_0/2$ apart may suppress the temporal interferences in $\tau_{dF}(\omega,t)$. A group delay representation $\tau_{dD}(\omega,t)$ that is interference-free in the both time and frequency domains is defined below:

$$\tau_{dD}(\omega,t)=\frac{P^{B+}\tau_{dF}\left(\omega,t+\frac{T_0}{4}\right)+P^{B-}\tau_{dF}\left(\omega,t-\frac{T_0}{4}\right)}{P^{B+}+P^{B-}}, \quad (7)$$

where P^{B+} represents $P\left(\omega+\frac{\omega_0}{4},t+\frac{T_0}{4}\right)+P\left(\omega-\frac{\omega_0}{4},t+\frac{T_0}{4}\right)$ and P^{B-} represents $P\left(\omega+\frac{\omega_0}{4},t-\frac{T_0}{4}\right)+P\left(\omega-\frac{\omega_0}{4},t-\frac{T_0}{4}\right)$. When the signal is periodic, $\tau_{dD}(\omega,t)=0$ effectively holds. This equation is conceptually simple and computationally efficient.

Unfortunately, this dually interference-free representation $\tau_{dD}(\omega,t)$ does not suppress both interferences perfectly. Numerical optimization was conducted using Hann windowing function $w(t)$ defined below:

$$w(t)=\begin{cases}1+\cos\left(\frac{2\pi t}{\beta T_0}\right) & 2|t|<\beta T_0\\ 0 & 2|t|\geq\beta T_0\end{cases}, \quad (8)$$

where β is the tuning parameter. The cost function L for this tuning is defined below:

$$L^2=\frac{1}{S(\Omega,T)}\int_{\Omega}\int_{T}|\tau_{dD}(\omega,t)|^2dtd\omega, \quad (9)$$

where $S(\Omega,T)$ represents the area defined by the temporal region T and frequency region Ω.

3.4.1. Parameter tuning

Two types of test signals were used to test this interference free group delay $\tau_{dD}(\omega,t)$. First one is a random phase harmonic signal $x_p(t)$ defined below:

$$x_p(t)=\sum_{k=0}^{\left\lfloor\frac{f_s}{2f_0}\right\rfloor}\cos\left(2\pi kf_0t+\varphi_k\right), \quad (10)$$

where f_s represents the sampling frequency (44,100 Hz in this test) and f_0 represents the fundamental frequency of the test signal (100 Hz in this test). The highest harmonic number is

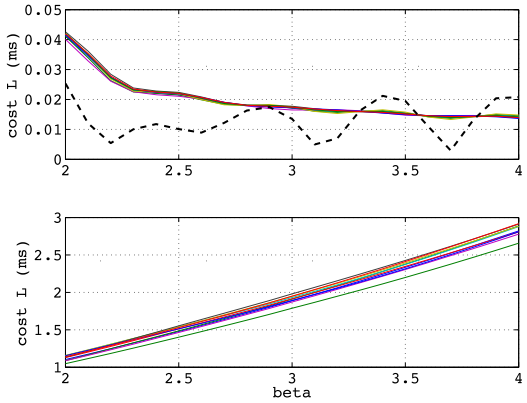

Figure 1: Dependency of the cost function L on the tuning parameter β. (Upper plot) F0 of the signal is 100 Hz. Sampling frequency is 44,100 Hz. The thick dashed line shows the result for the zero phase cosine series. (Lower plot) Test signal is Gaussian white noise samples.

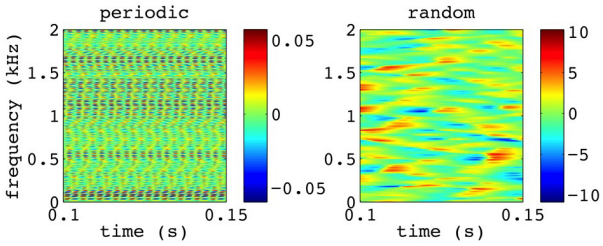

Figure 2: Time frequency representation of the interference-free group delay $\tau_{dD}(\omega, t)$ for a periodic signal (left image) and a random signal (right image). Note the scale of the color bars (the numbers are ms). ($\beta = 3$).

defined by flooring $\left\lfloor \frac{f_s}{2f_0} \right\rfloor$. The initial phase φ_k of the k-th component uniformly distribute in $[0,\ 2\pi)$. Second type signal is a Gaussian random noise $x_n(t)$. In this test, frequency region was set from 0 to $0.45 \times f_s$ and the width of the temporal region was $3 \times T_0$.

Figure 1 shows results for ten test runs. Upper plot is for the periodic test signals and the lower plot is for the random signals. In the upper plot, the black thick dashed line represents the result using a zero phase cosine series. In the lower plot, the same F0 is assumed for analyzing random signals for making comparison with periodic signals. These results indicate that remaining cancellation error is less than 1% of the standard deviation of random signals, when $\beta = 3$ is used. Based on this and other practical considerations, $\beta = 3$ is used afterwards.

Figure 2 shows the time frequency representations of the interference-free group delay $\tau_{dD}(\omega, t)$ for a periodic signal with phase randomization (left image) and a Gaussian random signal (right image). The image shows 0 Hz to 2 kHz. Image of higher frequency regions look similar to these examples. Note the huge magnitude difference between two color bar scales. The numbers on the color bar is represented in terms of ms.

3.4.2. Performance evaluation

Upper plot of Fig. 3 shows tolerance to harmonic amplitude variations as a function of standard deviation σ_{amp} represented by dB. Amplitude of each harmonic distributes in Gaussian with zero mean and variance of σ^2_{amp} (dB^2). The vertical axis represents the cost L defined by Eq. (9).

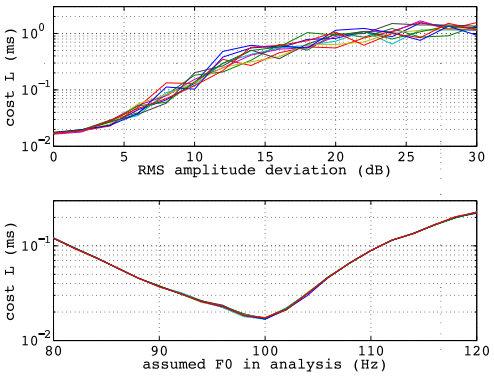

Figure 3: (Upper plot) Tolerance to harmonic amplitude variations. (Lower plot) Tolerance to F0 estimation error in analysis. F0 of the test signal is 100 Hz. ($\beta = 3$).

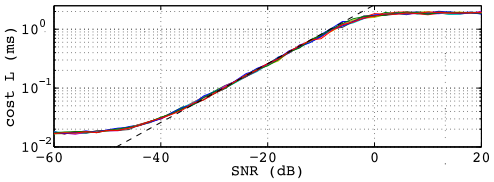

Figure 4: SNR linearity test. Gaussian white noise was added. The black dashed line represents $L = 2.6 \times 10^{\frac{SNR}{20}}$. ($\beta = 3$).

Lower plot of Fig. 3 shows the tolerance to F0 estimation error in analysis. Horizontal axis represents the assumed F0 value in calculating the interference-free group delay $\tau_{dD}(\omega, t)$. The true F0 value of the test signals is 100 Hz. The vertical axis represents the cost L. This figure suggests that upside 3% and downside 8% F0 estimation error can be negligible.

Figure 4 shows the cost L as a function of the input SNR (random noise RMS divided by periodic signal RMS). Note that from -40 dB to -5 dB SNR, the cost L is almost proportional to SNR. This wide range linearity allows to covert the cost L to band wise pulse to noise ratios which are useful for designing mixed-mode excitation signals for synthesis.

All the test results presented here are conducted using a single F0 value, 100 Hz. The test results regarding the cost L are proportional to fundamental period T_0 and these findings hold for other F0 values.

4. Application to speech

This section illustrates how to apply this proposed group delay representation to actual voice analysis. Figure 5 shows the waveform and extracted F0 of an isolated and sustained Japanese vowel /a/ spoken by a male speaker. F0 was extracted using TANDEM-STRAIGHT with 5 ms frame rate.

Figure 6 shows time frequency representations of the costs L of two interference-free group delays. The absolute value of the interference-free group delay $|\tau_{dD}(\omega, t)|$ is converted to equivalent SNR. Upper plot shows direct application of Eq. (7) and lower plot shows application of Eq. (7) after spectral flattening by using FIR inverse filters which are designed by square root of the reciprocal of spectral envelope estimated by TANDEM-STRAIGHT at each frame. Note that the horizontal trace of low SNR region around 500 Hz is disappeared in the lower plot.

Figure 7 shows temporal standard deviation of the interference-free group delay represented in terms of equivalent SNR. Upper plot shows that the spectral flattening is very effec-

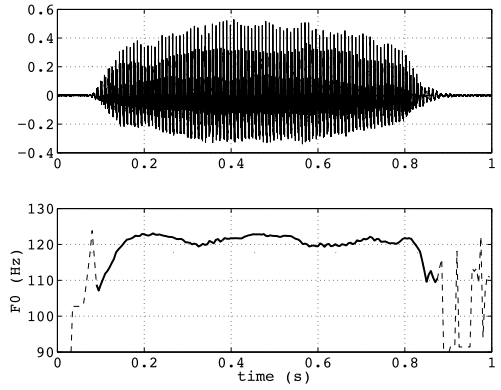

Figure 5: Waveform and F0 of an isolated Japanese vowel /a/ spoken by a male. Solid line indicates voicing.

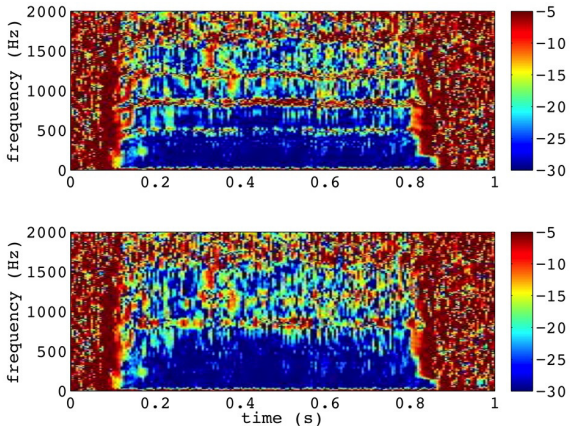

Figure 6: Time frequency representation of the interference-free group delay of Japanese vowel /a/. The group delay is converted to equivalent SNR.

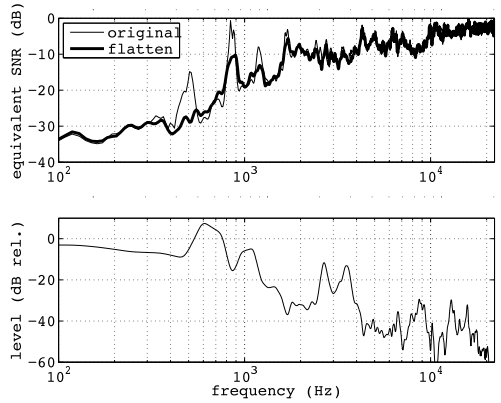

Figure 7: Temporal standard deviation of the interference-free group delay of Japanese vowel /a/. The lower plot shows averaged spectral envelope.

tive for suppressing SNR peak around 500 Hz as well as peaks around 850 Hz, 1200 Hz and 1800 Hz. This spectral flattening effects corresponds to the reduction of the horizontal trace around 500 Hz found in the lower plot of Fig. 6 and consistent with the tolerance test shown in the upper plot of Fig. 3.

Figure 8 shows smoothed equivalent SNR, converted from $\tau_{dS}(\omega, t)$, the smoothed interference-free group delay with spectral flattening. The smoothing is a weighted smoothing using the spectral envelope $P_{TST}(\omega, t)$ (STRAIGHT spectrum:

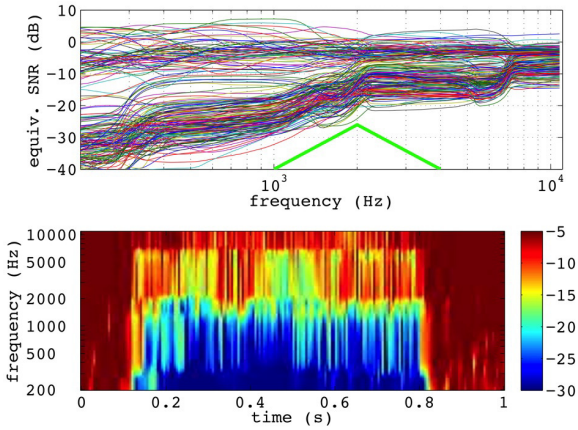

Figure 8: Equivalent SNR representation of Japanese vowel /a/ with two octave triangular smoother $h_\tau(\omega)$ (thick green line in the upper plot). Upper plot shows overlaid frequency characteristics. Lower plot shows the time frequency representation.

physical dimension is power spectrum) for weighting.

$$\tau_{dS}(\omega,t)=\left(\frac{\int_{\omega/2}^{2\omega} h_\tau(\nu)P_{TST}(\nu,t)\tau_{dD}^2(\nu,t)g(\nu)d\nu}{\int_{\omega/2}^{2\omega} h_\tau(\nu)P_{TST}(\nu,t)g(\nu)d\nu}\right)^{\frac{1}{2}}, \quad (11)$$

where $g(\nu)$ represents weighting for frequency axis conversion.

This smoothing reduces random fluctuation by increasing the TB (time bandwidth) product [33]. Upper plot of Fig. 8 shows the overlaid plot of smoothed group delay. The smoother shape is shown in the plot with thick green line. The lower plot shows the time frequency map of the same information.

All the procedures are implemented using Matlab [34] for feasibility study (computational efficacy is not considered here). Even without computational optimization, elapsed time of whole process (spectral flattening with FIR filter design at each frame, interference-free group delay calculation, group delay smoothing in the frequency domain) was 1.4 s for processing 1 s data with 44,100 Hz sampling on a PC (MacBook Pro, 2.6 GHz Intel Core i7, 16GB memory, OS X 10.9.2).

5. Discussions

A dually interference-free group delay representation is introduced. Tests using artificial signals and natural speech analysis examples indicate that the proposed representation has desirable feature for designing spectral shaping of mixed mode excitation signals [9, 11, 20] for high-quality speech manipulation systems. Spectral flattening can be improved in terms of efficiency and precision by introducing STRAIGHT-based LPC estimation [35] for spectral peak recovery [36]. Also group delay compensation for excitation extraction [37] based on minimum phase response [38] provides additional independent information, concentration of excitation energy around GCIs (Glottal Closure Instance) [26], for designing the temporal shape of each excitation event. This temporal shaping has huge perceptual effects especially in low-pitched voice [24]. Integration to STRAIGHT and perceptual evaluation of these extensions are currently understudy.

6. Acknowledgements

This work is partly supported by Kakenhi (Aids for Scientific Research) of JSPS 24300073 and 24650085.

7. References

[1] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[2] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," in *Proc. ICASSP 2008*, 2008, pp. 3933–3936.

[3] H. Kawahara, "STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustic Science & Technology*, vol. 27, no. 5, pp. 349–353, 2006.

[4] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Proc. ICASSP 2003*, vol. I, Hong Kong, 2003, pp. 256–259.

[5] S. R. Schweinberger, C. Casper, N. Hauthal, J. M. Kaufmann, H. Kawahara, N. Kloth, and D. M. Robertson, "Auditory adaptation in voice perception," *Current Biology*, vol. 18, pp. 684–688, 2008.

[6] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G. Rousselet, H. Kawahara, and P. Belin, "Vocal attractiveness increases by averaging," *Current Biology*, vol. 20, no. 2, pp. 116–120, 2010.

[7] H. Kawahara, H. Morise, Banno, and V. G. Skuk, "Temporally variable multi-aspect N-way morphing based on interference-free speech representations," in *ASPIPA ASC 2013*, 2013, p. 0S28.02.

[8] S. R. Schweinberger, H. Kawahara, A. P. Simpson, V. G. Skuk, and R. Zäske, "Speaker perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.

[9] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *Proc. MAVEBA*, pp. 13–15, 2001.

[10] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J. C. Williams, "Noh voice quality," *Logopedics Phoniatrics Vocology*, vol. 34, no. 4, pp. 157–170, 2009.

[11] H. Kawahara, M. Morise, T. Takahashi, H. Banno, R. Nisimura, and T. Irino, "Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems." in *Proc. Interspeech 2010*, 2010, pp. 38–41.

[12] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[13] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencie," *Electro. Comm. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970, [in Japanese].

[14] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 637–655, 1971.

[15] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 5, pp. 1087 – 1089, Oct. 1984.

[16] V. Villavicencio, A. R obel, and X. Rodet, "Applying improved spectral modeling for high quality voice conversion," *ICASSP2009*, pp. 4285–4288, 2009.

[17] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[18] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech and Signal Processing,*, vol. 34, no. 4, pp. 744–754, Aug 1986.

[19] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proc. Eurospeech 95*, Madrid, 1995, pp. 451–454.

[20] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2085–2095, Oct 2013.

[21] B. S. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. ICASSP'82*, vol. 7. IEEE, 1982, pp. 614–617.

[22] A. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct 1994.

[23] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Am.*, vol. 82, no. 5, pp. 1560–1586, 1987.

[24] J. Skoglund and W. Kleijn, "On time-frequency masking in voiced speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 361 –369, Jul. 2000.

[25] S. Uppenkamp, S. Fobel, and R. D. Patterson, "The effect of temporal asymmetry on the detection and perception of short chirp," *Hearing Research*, vol. 158, no. 1-2, pp. 71–83, 2001.

[26] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1602–1613, 2008.

[27] M. Morise, T. Takahashi, H. Kawahara, and T. Irino, "Power spectrum estimation method for periodic signals virtually irrespective to time window position," *Trans. IEICE*, vol. J90-D, no. 12, pp. 3265–3267, 2007, [in Japanese].

[28] H. Kawahara, T. Irino, and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," in *Proc. ICASSP 2011*, May 2011, pp. 5420–5423.

[29] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," in *Proc. ICASSP 2013*, 2013, pp. 6797–6801.

[30] ——, "An interference-free representation of group delay for periodic signals," in *Proc. APSIPA ASC 2012*, Dec 2012, pp. 1–4.

[31] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, pp. 1493–1509, November 1966.

[32] L. Cohen, *Time-frequency analysis*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[33] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, April 1967.

[34] Matlab, *version 8.2.0.701 (R2013b)*. Natick, Massachusetts, USA: The MathWorks Inc., 2013.

[35] A. Arakawa, Y. Uchimura, H. Banno, F. Itakura, and H. Kawahara, "High quality voice manipulation method based on the vocal tract area function obtained from sub-band lsp of straight spectrum," in *Proc. ICASSP 2010*, March 2010, pp. 4834–4837.

[36] H. Kawahara, M. Morise, T. Toda, R. Nisimura, and T. Irino, "Beyond bandlimited sampling of speech spectral envelope imposed by the harmonic structure of voiced sounds," in *Proc. Interspeech 2013*, 2013, pp. 34–38.

[37] H. Kawahara, Y. Atake, and P. Zolfaghari, "Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay." in *ICSLP 2000*, 2000, pp. 664–667.

[38] A. V. Oppenheim and R. W. Shafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall,, 1987.