Beyond bandlimited sampling of speech spectral envelope imposed by the harmonic structure of voiced sounds

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

[1]Department of Design Information Sciences, Wakayama University, Japan
[2]Faculty of Engineering, University of Yamanashi, Japan
[3]Nara Advanced Institute of Science and Technology, Japan

kawahara@sys.wakayama-u.ac.jp

Abstract

A new spectral envelope estimation procedure is proposed to recover details beyond band limitation imposed by the Shannon's sampling theory when interpreting periodic excitation of voiced sounds as the sampling operation in the frequency domain. The proposed procedure is a hybrid of STRAIGHT, a F0-adaptive spectral envelope estimation and the auto regressive model parameter estimation. Wavelet analyses of these spectral models on the frequency domain enabled objective evaluation of this recovery procedure. The proposed procedure provides better speech quality especially when parameter manipulation is introduced.

Index Terms: Speech analysis, envelope spectrum, sampling theory, speech modification, transfer function

1. Introduction

Flexible modification of speech sounds requires spectral envelope modeling as well as the excitation source modeling. Channel VOCODER [1], linear predictive coding (LPC) [2, 3, 4], cepstrum-based method [5, 6, 7], sinusoidal models [8, 9] and STRAIGHT [10, 11] are representative examples of such spectral models. Estimation of relevant spectral envelopes from voiced sounds is not simple, because vocal tract transfer functions consist of components which violate the band limitation requirement. This requirement is imposed by the equivalent spectral sampling of the periodic excitation of voiced sounds when represented in the frequency domain. The violation of band limitation is due to the fact that vocal tract transfer functions are represented by rational functions having a cosine series as the denominator (in discrete time systems). Because the cosine series is in denominator, even though the number of terms of the cosine series is finite, Fourier series representation of the transfer function has infinite number of terms. However, after over a half century from introduction of Shanonn's sampling theory [12], new views on sampling problem emerge and provide basis to alleviate (or evade) this band-limitation barrier by using side information (in our case F0 and knowledge about vocal tract transfer function) [13, 14].

This article reformulates our previous proposals [15, 16] on spectral shape compensation for improving manipulated speech sound quality and proposes an optimized design of a hybrid procedure of STRAIGHT and LPC.

A vocal tract can be modeled using an one dimensional non-uniform acoustic tube [17, 18]. This approximation yields an all-pole transfer function shown in the upper plot of Fig. 1. The squared absolute value of the denominator of the transfer function of a non-uniform acoustic tube is a cosine series with real valued coefficients. Logarithmic conversion of the denominator has sharp negative peaks and they correspond to sharp formant peaks of the transfer function. Attributes of these peaks (formants) have strong contribution to perceptual identity and quality of speech sounds. Acoustic measurements of the vocal tract transfer function using real humans [19] and three dimensional replica [20] verified that the peak shapes are actually very sharp.

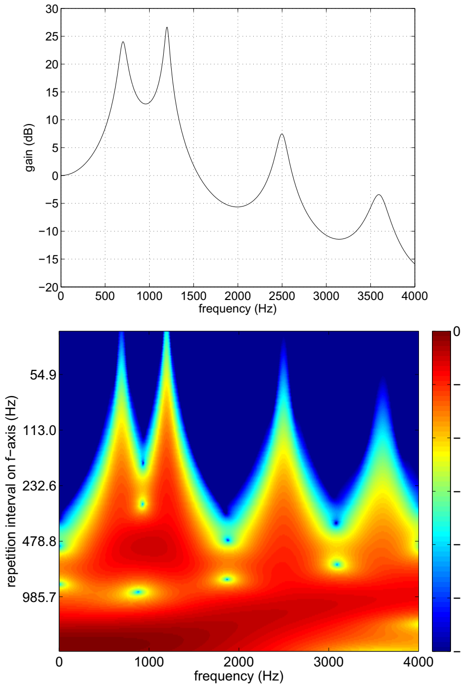

Figure 1: Transfer function example of one dimensional acoustic tube modeling of Japanese vowel /a/ (upper plot) and pseudo color map of the magnitude of its continuous wavelet transform using Morlet wavelet (lower plot).

Continuous wavelet analysis using Morlet wavelet [21] is introduced to clarify the band-unlimited nature of vocal tract transfer functions. The mother wavelet $\psi(f)$ used here has the following form.

$$\psi(f) = \left(e^{jk_0 f} - e^{-\frac{k_0^2}{2}} \right) e^{-\frac{f^2}{2}}, \tag{1}$$

where f represents the frequency axis of Fig. 1. It is a Gaussian wavelet with a small correction factor to satisfy the admissibility condition. The wave number coefficient $k_0 = 4$ is used here.

The lower plot of Fig. 1 shows the absolute value of the wavelet transform $W_{\psi,H}(a, f)$. The vertical axis represents the repetition interval of the scaled version of the carrier signal $e^{jk_0 f}$ of each Morlet wavelet. The vertical axis has the dimension of time (from bottom to top, the values roughly correspond to 1, 2, 4, 8 and 16 ms). The pseudo color mapping represents the absolute value of the wavelet transform of the differentiated logarithmic conversion of the vocal tract transfer function $H(f)$ and the numbers indicated on the color bar show the magnitude in terms of dB.

$$W_{\psi,H}(a, f) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \psi^* \left(\frac{\lambda - f}{a} \right) \frac{d \ln |H(\lambda)|}{d\lambda} d\lambda, \tag{2}$$

where * represents complex conjugate and $1/a$ is shown on the vertical axis of Fig. 1. This representation illustrates that higher components violating band limitation requirement imposed by the spectral sampling due to periodic excitation are located around these sharp peaks.

1.1. Spectral smearing by modeling

Spectrum modeling using a set of band-pass filters, such as channel VOCODER [1], smears these sharp peaks, because spectral shape resolution is limited by the bandwidth of the band-pass filters. Cepstrum liftering [5, 6, 7] for removing periodic variations representing harmonic structure also smears these sharp peaks, because it is a low-pass filtering on the frequency axis for removing components varying finer than $2f_0$ period. For example, using excitation of 100 Hz pulse train, details shown in upper than 200 Hz (repetition interval on the *vertical* axis) region in the bottom plot of Fig. 1 are removed.

Spectral smoothing operations used in STRAIGHT (both legacy-STRAIGHT [10] and TANDEM-STRAIGHT [11, 16]) also smear these details, while they are not strictly band-limited, rapidly changing components on the frequency axis are attenuated. These smoothing operations are error-tolerant[1] implementation [10] of piece-wise stair-case interpolation and piece-wise linear interpolation for TANDEM-STRAIGHT and legacy-STRAIGHT respectively. One of our previous proposals [16] reduces this attenuation effect by numerically adjusting amount of enhancement introduced to implement the digital compensation filter for consistent sampling [13, 22].

All these procedures also suffer from spectral smearing due to time windowing when they are implemented using short-term Fourier transform [23, 24]. Frequency domain representation of these time windowing functions provides impulse responses of low-pass filters on the frequency axis.

LPC introduces other smearing. Formulation of LPC by Itakura [2] is originally a maximum likelihood estimation of autoregressive model parameters by assuming speech as a weakly stable stochastic process. Since the model structure, in this case, agrees with the physically correct transfer function, the spectral envelope calculated from the estimated parameters has properly sharp peaks. It is not band-limited. However, since the actual excitation signal of speech sounds is not an independent Gaussian random noise, the true speech spectral envelope (although it is not directly observable) has peaks and dips due to local (short range) correlations. All these details are smeared out in the spectral envelope derived from LPC analyses. Since smearing of these details significantly deteriorates synthesized speech quality, several types of excitation models are proposed to overcome this difficulty of LPC [25, 26].

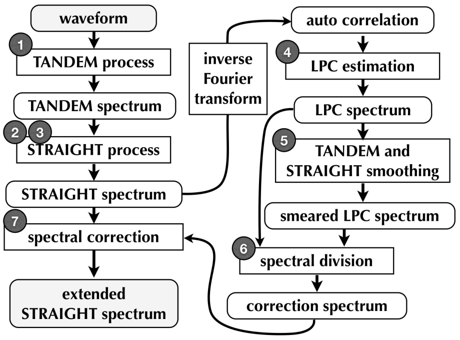

Figure 2: Schematic diagram of the proposed method. Circled numbers represent corresponding steps in 1.2 of the text.

1.2. Outline of the proposed method

By combining merit of LPC with STRAIGHT (also with enhancement based on consistent sampling and numerical tuning), the true spectral envelope is likely to be recovered in the following steps. Figure 2 illustrates a procedure to implement this idea.

In the first step, temporally stable power spectrum (TANDEM spectrum) is calculated. In the next step, periodic spectral variations due to periodic excitation is selectively suppressed by using an F0-adaptive triangular smoother having width of $2f_0$ on the frequency axis. In the third step, over-smoothing due to smearing caused by time windowing and the triangular smoother are compensated by the digital filter on the frequency axis designed based on consistent sampling theory and numerically tuned afterwards. Let name the spectral envelope obtained at this point as STRAIGHT spectrum. (It is a power spectrum.) In the fourth step, LPC spectral envelope is estimated from autocorrelation coefficients calculated from the inverse Fourier transform of the STRAIGHT spectrum. In the fifth step, the smeared LPC spectrum is calculated by smoothing it using the frequency domain representation of the time window and the triangular smoother. In the sixth step, the correction spectrum is yielded as the LPC spectrum divided by the smeared LPC spectrum. In the final step, the extended STRAIGHT spectrum is calculated by multiplying the correction spectrum to the STRAIGHT spectrum.

This is a reformulation and extension of our proposals [15, 16]. The next section starts from brief descriptions of TANDEM and STRAIGHT, followed by detailed explanations of the proposed method with numerical examples.

2. The proposed method

The proposed method shares the following TANDEM and STRAIGHT procedures. A brief introduction to these procedures and examples are presented here.

2.1. TANDEM

The TANDEM procedure eliminates temporal variation due to periodicity by F0-adaptive window design and F0-adaptive averaging [27, 11]. Let $P(\omega, t)$ represent the power spectrum around time t of a windowed voiced speech and $T_0 = 1/f_0$ represent the fundamental period. Then, TANDEM spectrum $P_T(\omega, t)$, which does not have temporally varying component

[1] Error here refers fundamental frequency estimation error and background noise as additive power spectrum error.

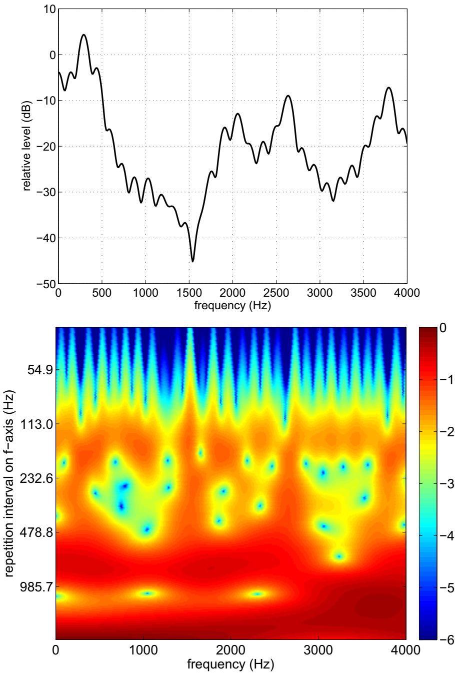

Figure 3: TANDEM spectrum (upper plot) and its wavelet analysis result (lower plot). A segment around vowel /i/ in a Japanese utterance /arigatou/ ("thank you" in English) spoken by a male is used. The horizontal orange blob in the lower plot corresponds to the harmonic structure ($f_0 = 145$ Hz) and cyan periodic spikes correspond to the negative periodic dips in the upper plot.

due to excitation periodicity is calculated by the following equation.

$$P_T(\omega, t) = \frac{P(\omega, t - \frac{T_0}{4}) + P(\omega, t + \frac{T_0}{4})}{2} \tag{3}$$

This procedure yields temporally stable representation, but the spectrum still has periodic variations due to excitation periodicity. The variation is a single cosine and multiplied to power spectral envelope. However, the variation represented in the log-power spectrum has harmonic components in the cepstrum domain, because of the logarithmic nonlinearity. The wavelet transform shows this nonlinear effects and periodic variation due to periodic excitation clearly as shown in Fig. 3.

2.2. STRAIGHT (with compensation and enhancement)

These interfering variations due to periodicity are selectively eliminated by the F0-adaptive smoothing in STRAIGHT. Smearing caused by the over-smoothing mentioned before is compensated by the digital filter in the frequency domain. This digital filter is also utilized to enhance spectral details for improving perceptual quality of the manipulated and resynthesized speech. Figure 4 shows STRAIGHT spectrum and its wavelet analysis results.

Whole procedures are approximately implemented as liftering in the cepstrum domain [22]. The following equation represents the whole process.

$$P_{TST}(\omega) = \exp\left(\mathcal{F}\left[g_A(\tau) g_C(\tau) C_T(\tau)\right]\right), \tag{4}$$

where $C_T(\tau)$ represents the cepstrum calculated from the TANDEM spectrum. The symbol $\mathcal{F}$ represents Fourier transform. The lifter $g_C(\tau)$ is the approximate implementation of the compensation digital filter, which is initially designed based on the consistent sampling, followed by the numerical tuning to reduce spectral smearing around formant peaks [16], based on a set of simulations using vocal tract area functions of eleven English vowels [28] and glottal waveform model [29] with random parameter perturbations.

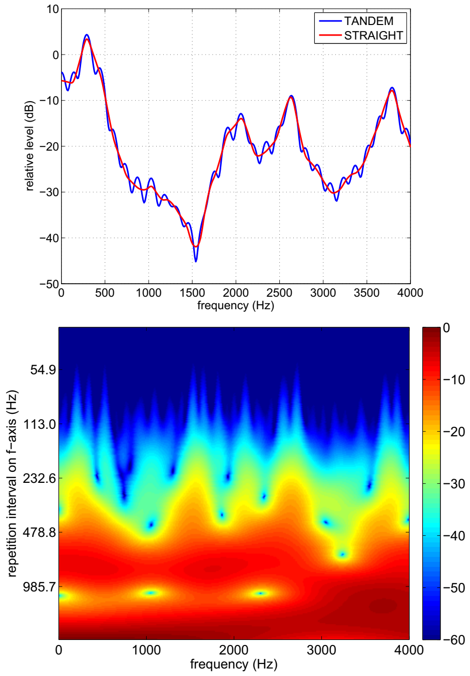

Figure 4: STRAIGHT spectrum (blue line) and TANDEM spectrum (red line) (upper plot) and wavelet analysis result of STRAIGHT spectrum using $g_C(\tau)$ and $g_2(\tau)$ lifters (lower plot). The horizontal orange blob and sharp cyan spikes found in Fig. 3 are suppressed. However, large sharp peaks corresponding to formants are not salient due to smoothing.

$$g_C(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_0}\right). \tag{5}$$

The lifter $g_A(\tau), A \in \{1, 2\}$ is F0 adaptively designed to eliminate periodic variations due to the harmonic structure. The lifter $g_1(\tau)$ is used in TANDEM-STRAIGHT implementation.

$$g_1(\tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_0 \tau} = \mathcal{F}[h_1(\omega)], \tag{6}$$

$$h_1(\omega) = \left\{ \begin{array}{cc} 0 & |\omega| \geq \frac{\omega_0}{2} \\ \frac{1}{\omega_0} & \text{otherwise} \end{array} \right. . \tag{7}$$

The lifter $g_2(\tau)$ is used in legacy-STRAIGHT implementation.

$$g_2(\tau) = \left(\frac{\sin(\pi f_0 \tau)}{\pi f_0 \tau}\right)^2 = \mathcal{F}[h_2(\omega)], \tag{8}$$

$$h_2(\omega) = \left\{ \begin{array}{cc} 0 & |\omega| \geq \omega_0 \\ \frac{1}{\omega_0}\left(1 - \left|\frac{\omega}{\omega_0}\right|\right) & \text{otherwise} \end{array} \right. \tag{9}$$

2.3. LPC from STRAIGHT spectrum

TANDEM spectrum and STRAIGHT spectrum are derived from the corresponding power spectra and can be converted to autocorrelation coefficients $r_k, (k = 0, 1, \ldots, N)$ by inverse Fourier transform. A Toeplitz matrix $\boldsymbol{R}$ and an autocorrelation vector $\boldsymbol{r}$, composed from the autocorrelation coefficients r_k are used to estimate the LPC coefficients $a_n, (n = 1, 2, \ldots, p)$, where p is the assumed number of poles. Let $\boldsymbol{a}$ represent the prediction coefficient vector consisting of a_k. The following equation yields the least square solution and is the maximum likelihood estimate of the underlying autoregressive process, when the process is excited by a Gaussian white noise.

$$\boldsymbol{a} = \boldsymbol{R}^{-1}\boldsymbol{r}. \tag{10}$$

This Gaussian assumption does not hold for voiced speech and the estimates may consist of bias due to model mismatch. However, this provides the best available estimate and the function form of the transfer function around the spectral peaks agrees with the physically relevant model (although parameter values are biased). The next step is designed to use this physically correct shape to compensate for smearing caused by over smoothing introduced by the time windowing and periodicity suppression of STRAIGHT procedure.

The power spectral representation of LPC spectral envelope $P_A(\omega)$ is defined by the following equation.

$$P_A(\omega) = \frac{1}{\left|1 - \sum_{n=1}^{p} a_n \exp(-j2\pi\omega/\omega_s)\right|^2}, \tag{11}$$

where $\omega_s = 2\pi f_s$ represents the sampling angular frequency. Let name $P_A(\omega)$ the LPC spectrum in Fig. 2.

2.4. Spectral correction

Applying spectral smoothing using the frequency domain representation of time windowing function followed by cepstral liftering defined by Eq. 4 yields smeared version of the LPC spectrum $P_{Sim}(\omega)$. The correction spectrum is defined by the ratio of the LPC spectrum and the smeared LPC spectrum. The final extended STRAIGHT spectrum $P_E(\omega)$ is calculated by the following equation.

$$P_E(\omega) = \frac{P_A(\omega)}{P_{Sim}(\omega)} P_{TST}(\omega). \tag{12}$$

The coefficient $P_A(\omega)/P_{Sim}(\omega)$ is the correction spectrum in Fig. 2. As shown in Fig. 5, the difference between STRAIGHT spectrum and the extended STRAIGHT spectrum is visible only around formant peaks. The modification introduced by the correction spectrum is highly localized. This localized correction enables to combine merit of non-parametric spectrum model (STRAIGHT) and parametric spectrum model (LPC spectrum) without introducing wrong side-effects. Preliminary listening tests of the proposed method indicated improvement in perceptual quality and intelligibility of the manipulated speech. Manipulated speech samples are linked to our web page [30].

3. Discussion

LPC coefficients generally suffer from estimation bias due to periodic excitation of voiced sounds [31]. The proposed spectral correction also suffers from this bias. Less biased methods [31, 32, 33, 34] should be tested for the LPC spectrum calculation process in the proposed method. The proposed procedure uses slightly excessive number of coefficients based on preliminary tests. An automatic determination algorithm should be implemented. Since, various representations of LPC related parameter are mutually converted uniquely [35], they are essentially identical to auto-correlation parameter (excluding covariance method). In this respect, applying generalized cepstrum method [36] to calculate less biased autocorrelation is an interesting possibility.

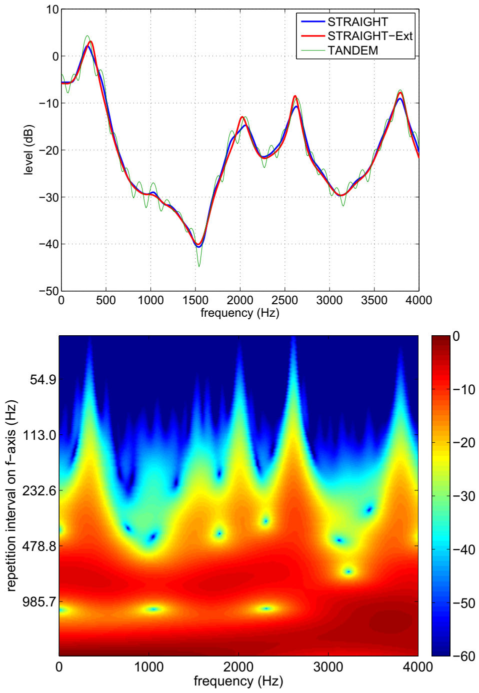

Figure 5: The extended STRAIGHT spectrum (read thick line) is overlaid on STRAIGHT spectrum (blue thin line) and TANDEM spectrum (green thin line) (upper plot) and wavelet analysis result of extended STRAIGHT spectrum is also shown (lower plot). Sharp peaks are recovered in the upper plot and sharp cyan peaks represent recovered band-unlimited components.

One of the authors proposed to use statistical approach for recovering smeared spectral peaks by making use of distributed information from the other parts of utterances [37]. Since the current proposed method is frame-based, it is interesting to integrate with such statistical approaches. These are interesting issues for further study.

It is important to note that the actual speech production process involves nonlinear interactions between vocal fold vibration and resulted acoustic impedance variations within one pitch period [38]. These detailed effects should be tested subjectively, since temporal aspects of masking also significantly affects perceived SNR and timbre [39].

4. Conclusion

A new spectral envelope recovery procedure is introduced as a hybrid of STRAIGHT and LPC to overcome band-limitation imposed by spectral sampling due to periodic excitation in voiced sounds. Parameter optimization based on objective as well as subjective tests are underway.

5. Acknowledgements

The authors appreciate Keisuke Saito of our group for testing preliminary implementation of our proposal. They also appreciate Tomoyasu Nakano and Masataka Goto for their inspiring discussions and proposals on spectral estimation [40]. Noboru Morimoto made the authors revisit wavelet analysis. This work is supported by the Aid for Scientific Research No.24300073, 23700221 and 24650085 from JSPS.

6. References

[1] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169–177, 1939.

[2] 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].

[3] 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.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[5] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, no. 2, pp. 221 – 226, 1968.

[6] 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.

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

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

[9] H. Kenmochi, "Singing synthesis as a new musical instrument," in *ICASSP2012*, March 2012, pp. 5385 –5388.

[10] 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.

[11] 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," *ICASSP2008*, pp. 3933–3936, 2008.

[12] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, 1949.

[13] M. Unser, "Sampling–50 years after Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, 2000.

[14] Y. C. Elder and T. Michaeli, "Beyond bandlimited sampling," *IEEE Signal Processing Magazine*, pp. 48–68, May 2009.

[15] H. Kawahara, M. Morise, H. Banno, T. Takahashi, R. Nisimura, and T. Irino, "Spectral envelope recovery beyond the Nyquist limit for high-quality manipulation of speech sounds," in *Proc. Interspeech 2008*, 2008, pp. 650–653.

[16] H. Akagiri, M. Morise, T. Irino, and H. Kawahara, "Evaluation and optimization of F0-adaptive spectral envelope estimation based on spectral smoothing with peak emphasis," *Trans. IEICE*, vol. J94-A, no. 8, pp. 557–567, 2011, [in Japanese].

[17] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[18] H. Wakita, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 3, pp. 281 – 285, 1979.

[19] O. Fujimura and J. Lindqvist, "Sweep ‐ tone measurements of vocal ‐ tract characteristics," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 541–558, 1971.

[20] T. Kitamura, H. Takemoto, S. Adachi, and K. Honda, "Transfer functions of solid vocal-tract models constructed from ATR MRI database of japanese vowel production," *Acoustical Science and Technology*, vol. 30, no. 4, pp. 288–296, 2009.

[21] R. Kronland-Martinet, J. Morlet, and A. Grossmann, "Analysis of sound patterns through wavelet transforms," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 01, no. 02, pp. 273–302, 1987.

[22] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *SADHANA*, vol. 36, no. 5, pp. 713–722, 2011.

[23] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[24] A. H. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. Audio Speech and Signal Processing*, vol. 29, no. 1, pp. 84–91, 1981.

[25] M. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in *ICASSP'85*, 1985, pp. 937–940.

[26] Z. Wen, J. Tao, and H.-U. Hain, "Pitch-scaled spectrum based excitation model for HMM-based speech synthesis," in *ICSP2012*, 2012, pp. 609–612.

[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] B. H. Story, "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," *J. Acoust. Soc. Am.*, vol. 26, no. 1, pp. 327–335, 2008.

[29] G. Fant and J. Liljencrants, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[30] H. Kawahara *et.al.*, "Spectral recovery." [Online]. Available: http://www.wakayama-u.ac.jp/%7ekawahara/exBandIS2013/

[31] H. Kawahara, K. Tochinai, and K. Nagata, "On the linear predictive analysis using a small analysis segment and its error evaluation," *Acoustical Society of Japan*, vol. 33, no. 9, pp. 470–479, 1977, [in Japanese: covariance method on closed vocal fold period, with trace of inverse covariance matrix for evaluation].

[32] D. Wong, J. Markel, and J. Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 4, pp. 350 – 355, 1979.

[33] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *Signal Processing, IEEE Transactions on*, vol. 39, no. 2, pp. 411 – 423, Feb. 1991.

[34] M. Thomas, J. Gudnason, and P. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 82 –91, 2012.

[35] S. Sagayama and F. Itakura, "Duality theory of composite sinusoidal modeling and linear prediction," in *ICASSP '86.*, vol. 11, April 1986, pp. 1261 – 1264.

[36] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis –a unified approach to speech spectral estimation," in *Proc. ICSLP1994*, vol. 3, 1994, pp. 1043–1046.

[37] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory hmm," in *ICASSP2008*, April 2008, pp. 3925 –3928.

[38] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, May 2008.

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

[40] T. Nakano and M. Goto, "A spectral envelope estimation method based on F0-adaptive multi-frame integration analysis," in *SAPA workshop*, Portland Oregon, Sept. 2012.