Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Evaluation and optimization of F0-adaptive spectral envelope estimation based on spectral smoothing with peak emphasis

**Hayato Akagiri (1), Masanori Morise (2), Ryuichi Nisimura (1)
Toshio Irino (1) and Hideki Kawahara (1)**

(1) Department of Design Information Sciences, Wakayama University, Wakayama, Japan
(2) College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan

PACS: 43.72.-p ,43.72.Ar, 43.72.Ja, 43.75.Rs, 43.71.Gv, 43.70.Jt

ABSTRACT

A new spectral estimation method which improves processed sound quality of STRAIGHT, a speech analysis, modification and re-synthesis framework widely used for high-quality speech and singing manipulations, is proposed. Application of the proposed method to TANDEM-STRAIGHT, a completely reformulated version of STRAIGHT, yielded the best spectral envelope approximation among conventional methods such as LPC, cepstrum and legacy-STRAIGHT. TANDEM-STRAIGHT consists of two parts, a temporarily stable power spectrum estimation method of periodic signals (TANDEM) and a spectral envelope calculation method based on consistent sampling theory. The proposed method uses F0-adaptive smoothing and compensation of logarithmic power spectrum, for improving approximation accuracy of spectral peaks, which effects on the quality of re-synthesized sound. A series of simulations was conducted to optimize internal parameters of the proposed method. The optimized system was evaluated and compared with conventional methods using stylized spectra and simulated speech spectra. The evaluation was based on a spectral distance measure proposed by Itakura and Saitou with modification to perceptually relevant ERB_Nnumber frequency axis. The following set of spectra were used. Power spectra calculated from vocal tract area functions measured using MRI data with LF-model excitation spectra were used as the grand truth and spectral distances between this target and the estimated spectra were evaluated. A set of periodic pulse train was used for excitation signal in this case. These evaluation results indicated that the proposed method yields the smallest spectrum distance among conventional methods such as LPC, cepstrum and legacy-STRAIGHT.

INTRODUCTION

This paper presents a new spectral estimation method which improves processed sound quality of TANDEM-STRAIGHT (H.Kawahara et al. 2008b), a speech analysis, modification and re-synthesis framework.TANDEM-STRAIGHT supersedes its former implementation STRAIGHT (H.Kawahara, I.Masuda, and A.Cheveigné 1999) (legacy-STRAIGHT), which is widely used for high-quality speech and singing manipulations. However, reproduced speech sounds by TANDEM-STRAIGHT sometimes had lesser quality than the legacy-STRAIGHT. The proposed method solves this last defect and provides TANDEM-STRAIGHT the optimized performance.

The following section clarifies the problem to be solved. Then, based on a brief outline of TANDEM-STRAIGHT, the proposed method is introduced. The proposed method is then optimized by using a set of simulated spectral envelopes calculated based on a vocal tract shape data and a glottal source model. Discussion on the relevant spectral distance measure is also given. Finally, the optimized method is compared with other spectral envelope estimation methods and revealed that the proposed method yields the best performance.

BACKGROUND

The legacy-STRAIGHT and TANDEM-STRAIGHT share the common underlying principle, that periodic excitation of voiced sounds is a built-in sampler of the smooth time-frequency surface formed mainly by the vocal tract transfer function. Since this time-frequency surface is a collection of spectral envelopes, the problem to be solved is a discrete-to-analog conversion in the frequency domain, where discrete information is the levels of harmonic components. The legacy-STRAIGHT solved this problem using a collection of various heuristics, which were carefully tuned in a trial-and-error fashion. TANDEM-STRAIGHT replaced many of these heuristics with more theoretically sound procedures which do not require such tuning.

However, there still remains a fundamental problem because the spatial frequency contents of vocal tract transfer functions generally exceed Nyquist limit of the equivalent sampling rate determined by the fundamental frequency. Especially, in the vicinity of spectral peaks, higher-than-Nyquist spatial frequency components are indispensable to make peaks sharp . Important issue is that the sharpness of peaks contributes to make processed speech sound better in quality.

We once proposed a model-based approach to recover these higher-than-Nyquist spatial frequency components by using AR (auto regressive) type spectral model(H.Kawahara et al. 2008a). Although the proposed model enabled to supply those missing high spatial frequency components, it was not robust enough to be used in various applications, partly because of inflexible LPC parameter estimation procedure. The goal of this article is to provide a procedure which does not depend on AR model while enabling recovery of higher-than-Nyquist spatial

frequency components. The answer is to use logarithmic nonlinearity combined with its inverse, exponential function.

TANDEM-STRAIGHT

TANDEM-STRAIGHT consists of two parts, a temporarily stable power spectrum estimation method of periodic signals, called TANDEM(M.Morise et al. 2007) and a spectral envelope calculation method based on consistent sampling(M.Unser 2000). This section briefly outlines the spectral envelope estimation method used in TANDEM-STRAIGHT. Optimization of this envelope estimation procedure is the target of this article.

TANDEM spectrum

Power spectrum of a windowed periodic signal temporally varies due to the interference of harmonic components within an equivalent pass-band of a short-term Fourier transform. This temporal variation is effectively cancelled out by averaging two power spectra calculated using a pair of time windows temporally separated by half of the fundamental period. The principle of operation is given below(M.Morise et al. 2007).

Let $H(\omega)$ represent the Fourier transform of a time-windowing function. Assume a simplest case where two harmonic components $\delta(\omega)$ and $\alpha e^{j\beta}\delta(\omega-\omega_0)$ are located within an equivalent bandpass filter $H(\omega)$. Where $\omega_0 = 2\pi f_0$ represents fundamental angular frequency. The power spectrum of the windowed signal located at τ yields the following :

$$\begin{aligned} P(\omega,\tau) =& H^2(\omega) + \alpha^2 H^2(\omega-\omega_0) \\ &+ 2\alpha H(\omega)H(\omega-\omega_0)\cos(\omega_0\tau+\beta). \end{aligned} \tag{1}$$

The third term represents the temporal variation of the calculated power spectrum. This term is canceled out by averaging two power spectra calculated $T_0/2$ apart. Therefore, the average :

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

has no time-dependent term. Where $T_0 = 1/f_0$ represents the fundamental period. This simple method was named TANDEM (M.Morise et al. 2007) and the calculated temporally stable power spectrum $P_T(\omega)$ is called TANDEM spectrum.

STRAIGHT spectrum

TANDEM spectrum is a smoothed version of the sampled original spectral envelope. The spectral sampling interval is the fundamental frequency f_0. The problem to be solved is to recover the original spectral envelope from this sampled and smoothed representation.

This problem is solved in two steps, conceptually. In the first step, periodic spectral variations due to spectral sampling are completely suppressed by smoothing using smoothing functions having zero gain at spatial frequency $1/f_0$. The simplest smoother of this type is a rectangular function with smoothing width f_0. (The legacy-STRAIGHT uses a triangular function for smoothing. This function is convolution of this rectangular smoothing function and itself.)

The second step is to restore spectral values at harmonic frequencies. A digital filter for implementing this restoration is designed based on consistent sampling (M.Unser 2000). Consistent sampling only requires re-sampled values to be restored. This is a strong contrast with the Shannon's sampling theory, where complete restoration of whole signal is required. Due to this freedom in consistent sampling, restored spectral shape between harmonic frequencies is highly dependent on the smoothing function. The rectangular function used in TANDEM-STRAIGHT tends to smears out spectral peaks. This smearing makes perceptual quality of the processed speech degraded. This is the specific problem to be solved by the proposed method.

PROPOSED METHOD

Exponential function emphasizes peaks. Logarithmic conversion suppresses peaks. Logarithmic conversion followed by exponential expansion yields identity mapping because they are inverse of each other. Applying F0 adaptive spectral smoothing mentioned above on the logarithmic spectra in the middle of this compression and expansion process, instead of applying directly on power spectra, makes recovered spectral envelopes have sharper peaks. This is the simplest explanation of the proposed method.

The proposing method is described as :

$$L_s(\omega) = \int h(\lambda)\log(P_T(\omega-\lambda))d\lambda \tag{3}$$

$$P_{TST}(\omega) = \exp\Big(L_s(\omega) + \bar{q}_1(L_s(\omega-\omega_0)+L_s(\omega+\omega_0))\Big). \tag{4}$$

where $P_T(\omega)$ represents the TANDEM spectrum, $P_{TST}(\omega)$ represents the STRAIGHT spectrum, $h(\omega)$ represents the smoothing function, and $\bar{q}_1$ represents the first coefficient of the digital filter for restoration of the harmonic value in the smoothed spectrum. Note that the process is efficiently implemented using Cepstrum, instead of directly calculating convolution in Eq. 3.

This coefficient $\bar{q}_1$ has to be optimized because the original compensating digital filter has infinite number of coefficients and is defined in the power spectral domain. It is necessary to define relevant spectral distance measure and test sets to be used in the optimization process.

OPTIMIZATION

This section describes the evaluation method for optimizing this coefficient $\bar{q}_1$. The goal of this optimization is to minimize spectral distance between "true" spectral envelope and the estimated spectral envelope. The "true" spectral envelopes used in this optimization are calculated taking into account of vocal tract transfer functions, glottal source waveform and radiation characteristics. Vocal tract shape data measured by magnetic resonance imaging (MRI) (B.H.Story 2008) is used to calculate transfer functions. The LF-model (G.Fant and J.Liljencrants 1985) is used for calculating glottal source waveforms because it provides parametric representation of glottal source waveform with radiation characteristics.

Vocal tract transfer function

This section describes how to calculate vocal tract transfer functions used in the following optimization process. The one dimensional vocal tract area functions derived from 3D measurement of vocal tract shapes using MRI (B.H.Story 2008) are used for the base data. The reference provides a table of 44 cross sectional areas acquired for 11 American English vowels [i, ɪ, e, ɛ, æ, ʌ, ɑ, ɔ, o, ʊ, u] spoken by a male speaker. The length of each section varies with vowel types and are also tabulated.

The upper panel of Fig. 1 shows an example of the measured vocal tract area function. LPC coefficients are calculated from reflection coefficients between each uniform tube section and assuming the termination condition at the glottis opening 0.25cm^2 in this figure. The magnitude transfer function shown in the lower plot is calculated using this LPC coefficients. In this calculation, acoustic impedance of the vocal tract wall is not taken

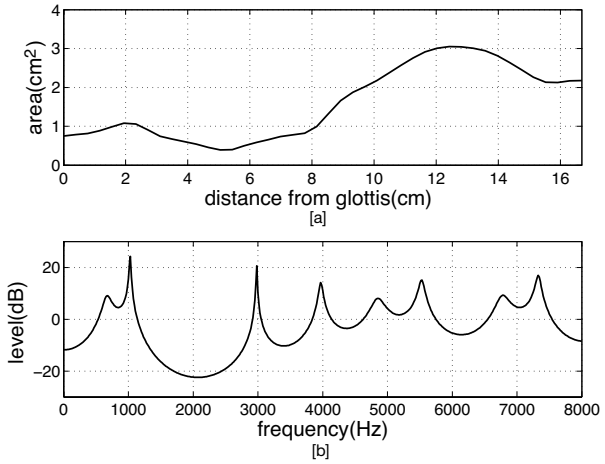

Figure 1: (a) Vocal tract area function and (b) calculated vocal tracts transfer function. (vowel [ʌ], glottis opening = 0.25cm^2)

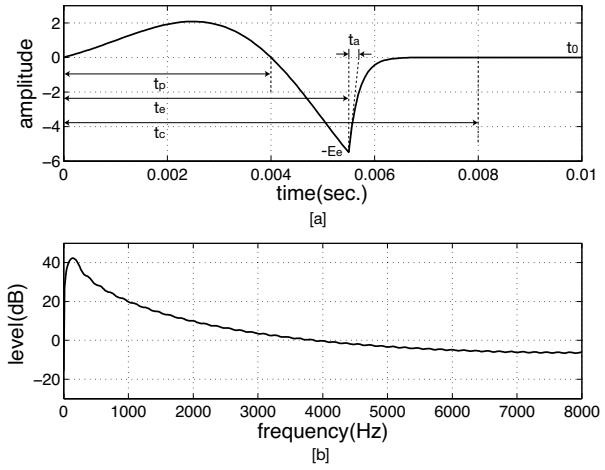

Figure 2: (a) the differentiated glottal flow waveform $E(t)$ calculate by the LF-model. (b) Spectral representation of $F(t)$. (f_0=100Hz, $t_p = 0.004$, $t_e = 0.0055$, $t_c = 0.008$, $t_a = 0.0002$(sec.))

into account.

Instead of using those details, the terminal condition at glottis are adjusted for the resultant transfer functions to have relevant formant band widths. The calculated formant band widths are compared with the acoustic measurement data of formant band widths measure using sweep-tone (O.Fujimura 1971) and are adjusted.

Glottal source and radiation characteristics

The LF-model 1 (G.Fant and J.Liljencrants 1985) is used to generate glottal air flow waveform. The differentiated version of the air flow was used to simulate radiation transfer function from lip opening. Figure 2 shows the differentiated glottal air flow example and its spectral representation. Parameters t_p, t_e, t_c and t_a shown in the upper plot determines the glottal air flow waveform implicitly through a, ε and ω_g in the following equation.

$$E(t) = \begin{cases} E_0 e^{at} \sin \omega_g t, & 0 \le t \le t_e \\ -\dfrac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right], & t_e \le t \le t_c < t_0 \end{cases} \quad (5)$$

where t_0 represents the fundamental period. The parameter t_p marks the position of peak glottal flow ($t_p = \pi/\omega_g$). The parameter t_e is the instant of the maximum glottal closing rate. The parameter t_c marks the instant of complete glottal closure. The parameter t_a is the time constant of the exponential recovery. It also represents abruptness of glottal closure. Because the waveform $E(t)$ is a differentiated periodic signal, integration of one fundamental period has to obey the following constraint.

$$\int_0^{t_0} E(t) = 0 \quad (6)$$

Spectrum distance measure

Based on our preliminary study on correlation between subjective speech quality data and objective spectral distance measures, Itakura-Saito distance measure on warped frequency axis (H.Akagiri et al. 2009) were used in the following optimization experiment and comparative test with other F0 estimation algorithms.

Let $P(\lambda)$ represent estimated spectral envelope and $P_{\mathrm{ref}}(\lambda)$ represent the "true" spectral shape. Then, the distance $D(P, P_{\mathrm{ref}}; f_0)$ is defined by the following equation.

$$D(P, P_{\mathrm{ref}}; f_0) = c \int_{\lambda_L(f_0)}^{\lambda_H} \left(\log \frac{P(\lambda)}{P_{\mathrm{ref}}(\lambda)} + \frac{P_{\mathrm{ref}}(\lambda)}{P(\lambda)} - 1 \right) d\lambda, \quad (7)$$

where λ represents perceptually relevant warped frequency axis defined by the following equation. The region of integration from $\lambda_L(f_0)$ to λ_H is for eliminating artifacts.

$$\lambda = 21.4 \log_{10}\left(\frac{4.37}{1000} f + 1\right), \quad (8)$$

This axis is called ERB$_N$number rate (B.C.J.Moore 2003). In implementation of this measure, variable conversion using the following equation is used.

$$\frac{d\lambda}{df} = h(f) = \frac{9.294}{0.00437 f + 1}. \quad (9)$$

Substituting all these, the following equation is derived and used for writing test codes of optimization experiments.

$$D(P, P_{\mathrm{ref}}; f_0) = \frac{1}{\int_{2f_0}^{0.45 f_s} h(f) df} \int_{2f_0}^{0.45 f_s} \left(\log \frac{P(f)}{P_{\mathrm{ref}}(f)} + \frac{P_{\mathrm{ref}}(f)}{P(f)} - 1 \right) h(f) df \quad (10)$$

where f denotes the frequency, f_0 the fundamental frequency given by $f_0 = \omega_0/2\pi$, and f_s the sampling frequency.

Evaluation method of spectral envelope

Figure 3 shows an example of the "true" spectrum (reference spectrum) and the estimated spectral envelope by TANDEM-STRAIGHT. Integrated difference between the solid line and the dotted line yields a single entry $D(P, P_{\mathrm{ref}}; f_0)$ for evaluation of this specific example.

Since our goal is to optimize the filter coefficient $\bar{q}_1$ for compensation, averaged behavior of the distance over various variations of voiced sounds has to be evaluated. Taking into this goal, a total cost is defined as the average of $D(P, P_{\mathrm{ref}}; f_0)$ over 11 vowel types with parametric perturbations and over relevant F0 range. Perturbation is introduced on the vocal tract length and opening of glottis area.

Although the proposed TANDEM-STRAIGHT based method does not depend on time window positioning, phase randomization was introduced to simulate window positioning effects.

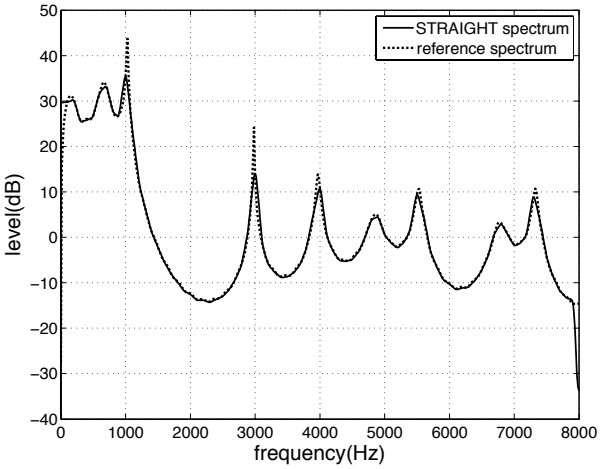

Figure 3: STRAIGHT spectrum (solid line) and reference spectrum (dashed line) of a vowel [ʌ].

Table 1: Test condition for optimization

parameters	range
Vocal tract shape	11 English vowels
Vocal tract length	variation at s.d. $\pm 10\%$ of V.T.L.
Glottal opening	$0.09 \sim 0.25 \mathrm{cm}^2$ (17 steps)
Fundamental frequency	40∼800Hz (1/12 octave step)

The input speech signals to test spectral estimators are generated using the reference spectrum $P_{\mathrm{ref}}(f)$.

$$x(t;f_0) = \sum_{k=1}^{N} \sqrt{P_{\mathrm{ref}}(kf_0)} \cos(2\pi k f_0 t + \varphi_k), \qquad (11)$$

where φ_k represents the randomized initial phase.

PARAMETER OPTIMIZATION

The evaluation score described in the previous section (averaged spectrum distance) was used to optimize the filter coefficient $\bar{q}_1$ for compensation. Vocal tract parameter perturbation is introduced to simulate talker and gender differences. Note that this perturbation is not aiming at accurately simulate variabilities of natural speech samples. Randomization introduced here is intended to avoid over-fitting to a specific test configuration. Perturbed parameters for vocal tract are listed on Tables 1 with their perturbation range.

Table 2 shows reported statistical data of the LF-model parameters (D.G.Childers and C.Ahn 1995). The LF-model parameters used in this optimization are randomized based on this statistical data of these parameter distributions. Only mean and standard deviations are adjusted based on the simulation results.

Test results

Figure 4 shows dependence of the evaluation score to the target coefficient $\bar{q}_1$. The evaluated results (averaged spectral distance) are represented in dB. Because the parameter to be tuned is only one in this case, exhaustive search is used to evaluate the effects. The parameter $\bar{q}_1$ was tested for the range $(-0.3 < \bar{q}_1 < 0.2)$ in 0.05 steps.

The score has the minimum value 1.55 dB at $\bar{q}_1 = -0.2$. This optimized coefficient is used in the following evaluations.

Table 2: Mean values and standard deviations for LF-model parameters (D.G.Childers and C.Ahn 1995). (Each parameter is represented in terms of percentage of the fundamental period (t_0).)

LF-model parameters	t_p (%)	t_e (%)	t_a (%)	t_c (%)
mean values	41.35	55.31	0.81	58.19
(standard deviations)	(5.50)	(7.78)	(0.06)	(8.85)

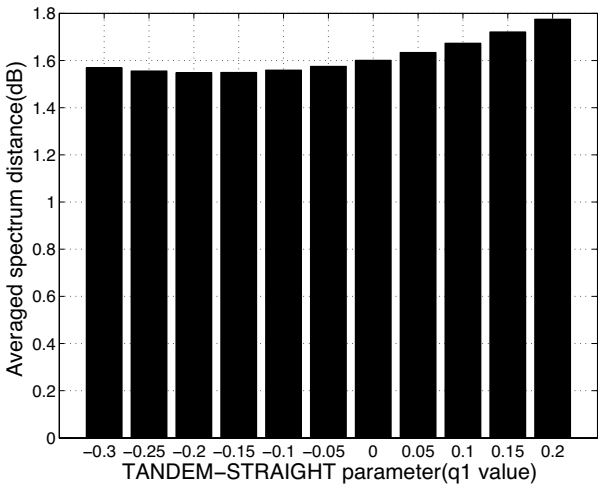

Figure 4: Averaged spectral distance vs. compensation coefficient $\bar{q}_1$.

COMPARATIVE TEST

In this section, the proposed method is compared with conventional methods such as LPC, cepstrum, legacy-STRAIGHT, and plain-TANDEM-STRAIGHT (without parameter optimization proposed here). The same test conditions used in the previous section are also used. The averaged spectral distance of the proposed method and the conventional methods are shown in Table 3. The results indicates that the proposed method yields the best (smallest) spectral distance.

CONCLUSION

A new spectral estimation method which uses F0-adaptive smoothing and compensation of logarithmic power spectrum, for improving approximation accuracy of spectral peaks is proposed. The propose method, applied to TANDEM-STRAIGHT, yielded the best spectral envelope approximation among conventional methods such as LPC, cepstrum and legacy-STRAIGHT.

Table 3: Minimum averaged spectral distances of TANDEM-STRAIGHT and conventional methods

estimation method	Minimum averaged spectral distance (dB)
TANDEM-STRAIGHT (proposed optimization with $\bar{q}_1 = -0.2$)	1.55
plain-TANDEM-STRAIGHT ($\bar{q}_1 = -0.1$)	1.83
legacy-STRAIGHT	1.80
LPC(order=24)	1.82
cepstrum(order=24)	3.07
FFT (blackman,window length=30ms)	9.11

Acknowledgement This study is partially supported by Grants-in-Aid for Scientific Research (A) 19200017 by JSPS and the CrestMuse project by JST.

REFERENCES

B.C.J.Moore (2003). "An Introduction to the Psychology of Hearing". *Academic Press, fifth edition*.

B.H.Story (2008). "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002". *J. Acoust. Soc. Am.* 123.1, pp. 327–335.

D.G.Childers and C.Ahn (1995). "Modeling the glottal volume-velocity waveform for three voice types". *J. Acoust. Soc. Am.* 97.1, pp. 505–519.

G.Fant and J.Liljencrants (1985). "A four-parameter model of glottal flow". *STL-QPSR* 26.4, pp. 1–13.

H.Akagiri et al. (2009). "Effects of spectral envelope representations on resynthesized speech quality". *Tech. Rep. IEICE* SP-109.99, pp. 63–68.

H.Kawahara, I.Masuda, and A.Cheveigné (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". *Speech Communication* 27, pp. 187–207.

H.Kawahara et al. (May 2008a). "Spectral Envelope Recovery beyond the Nyquist Limit for High-Quality Manipulation of Speech Sounds". *Interspeech 2008*, pp. 650–653.

H.Kawahara et al. (2008b). "TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation". *ICASSP2008*, pp. 3933–3936.

M.Morise et al. (2007). "Power Spectrum Estimation Method for Periodic Sigals Virtually Irrespective of Time Window Positioning". *Trans. IEICE* J90-D.12, pp. 3265–3267.

M.Unser (2000). "Sampling–50 Years After Shannon". *Proceedings of the IEEE* 88.4, pp. 569–587.

O.Fujimura (1971). "Sweep-Tone Measurements of Vocal-Tract Characteristics". *J. Acoust. Soc. Am.* 49.2B, pp. 541–558.