A Spectral Envelope Estimation Method Based on F0-Adaptive Multi-Frame Integration Analysis

Tomoyasu Nakano and *Masataka Goto*

National Institute of Advanced Industrial Science and Technology (AIST), Japan
{t.nakano, m.goto}[at]aist.go.jp

Abstract

This paper presents a novel method of spectral envelope estimation and representation. Despite much sophisticated work in this area, estimating an appropriate envelope is still difficult. We therefore propose an F_0*-adaptive multi-frame integration* analysis method for estimating spectral envelopes with appropriate shape and high temporal resolution. The method does not use pitch marks or phoneme labels and can be used with various types of sound (speech, singing, and instruments). The basic idea is to use F_0-adaptive window analysis with a small window length yielding high temporal resolution. The analysis is then extended by using neighboring frames to obtain a stable spectral envelope. In tests using synthesized sound and resynthesized natural sound samples, for 8 of 14 samples the log-spectral distances obtained with the proposed method were smaller than those obtained with well-known previous methods.

Index Terms: spectral envelope, periodic signals, source-filter model, speech/singing, instrument sound

1. Introduction

Source-filter processing [1] is an important way to deal with speech, singing, and instrument sounds. An appropriate spectral envelope reconstructed from an observed spectrum can be useful in flexible sound manipulation (synthesis) and in sound recognition (analysis). The aim of the work reported here was to develop a method for estimating a spectral envelope with appropriate shape and high temporal resolution for high-quality sound synthesis and high-accuracy sound analysis and to do so without using pitch marks[1] and phoneme labels so that the method can be used with any kind of sound (speech/singing/instrument).

Various sophisticated methods for estimating spectral envelopes have been proposed. One of the best known signal modeling techniques is that used in the phase vocoder [2], and many extensions of it have been proposed (*e.g.*, [3]). This technique successfully models and synthesizes harmonic signals with static fundamental frequency (F_0) characteristics but has limitations with regard to the synthesis of sounds with changing F_0. Although PSOLA [1, 4] is a well known signal manipulating method that can be used with F_0-changing sounds, it needs precise pitch marks.

Some other widely known spectral envelope representation methods use linear predictive coding (LPC) [5, 6], line spectral pairs (LSP), or cepstrum. LPC has been extended to mel-generalized cepstral analysis for treating various spectral representations [7], and the cepstrum-based methods have also been improved [8–10] and combined with the LPC method [11]. Although the envelopes estimated using these methods are better (with regard to spectral shape) than those estimated using the simple cepstrum and LPC methods, the detail of the envelope (*e.g.*, the sharpness of its peaks and valleys) is limited by the order of the cepstral/LPC approximation.

The most well known alternatives are the sinusoidal models [12, 13], for which there have been many extensions – such as making distinctions between the sinusoidal and broadband spectrum components or noise in the signal [14, 15], extracting the sinusoids from the spectrogram [16], extracting sinusoids in an iterative way [17, 18], using quadratic interpolation [19], getting higher temporal resolution [20], monaural speech separation [21], and applying sinusoidal models to nonstationary sounds [22, 23]. These sinusoidal approaches can be used for high-quality sound synthesis, and some have high temporal resolution [20, 22]. A sinusoidal model of the harmonic structure, however, has energy only at frequencies corresponding to integer multiples of F_0. It is therefore difficult to identify transfer characteristics between adjacent harmonics.

STRAIGHT [24] is a well known high-quality source-filter processing system (i.e., vocoder) and is widely used in the speech research community. From input speech sound it can estimate an interference-free spectrum based on F_0*-adaptive* analysis. An extension/reformulation called TANDEM-STRAIGHT [25] is based on calculating a temporally stable power spectrum, and either STRAIGHT or TANDEM-STRAIGHT can be used to estimate the envelope of an aperiodic component. The STRAIGHT spectral envelope can be used for high-quality sound synthesis and high-accuracy sound analysis with high temporal resolution without pitch marks and phoneme labels. Although this spectral envelope can be quite accurate, it is obtained by spectral smoothing. This means that even with the STRAIGHT spectral envelope it is difficult to identify transfer characteristics between adjacent harmonics.

This is also true when the spectral envelope is estimated using Gaussian mixture modeling, such as STRAIGHT envelope modeling [26], and when a Gaussian mixture function is used in the joint estimation of the F_0 and the spectral envelope [27].

Several methods for statistically estimating the envelope from multiple frames have been studied in efforts to deal with this problem [28–30], and this multi-frame approach has also been used in a method estimating spectral envelopes of vocal from music sound mixtures [31]. This *multi-frame integration* approach assumes that additional information, such as phoneme transcriptions, is available for selecting frames at which the spectral envelopes are presumed to be similar to each other. In this approach not only the harmonic components observed at each analyzed frame but also those at other frames are considered so that missing frequency parts can be interpolated statistically. This kind of multi-frame analysis, however, needs phoneme transcriptions and has difficulties dealing with sound varieties and context dependencies.

[1] Time instants for F_0-synchronous analysis [1].

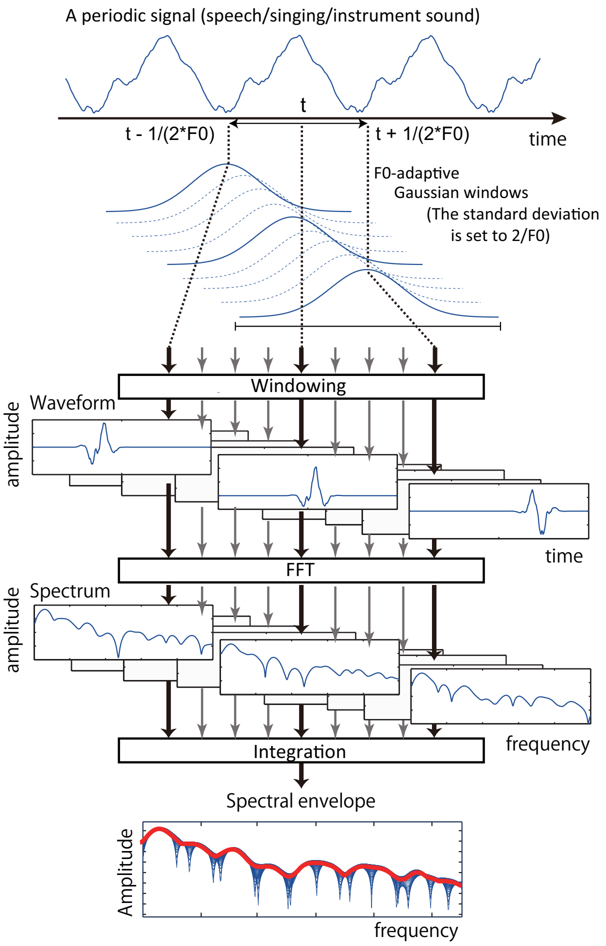

Figure 1: *Overview of F_0-adaptive multi-frame integration analysis.*

This paper presents a novel method of spectral envelope estimation and representation: F_0*-adaptive multi-frame integration analysis*. The basic idea is to use F_0-adaptive analysis with a small window length yielding high temporal resolution. The analysis is extended to a multi-frame integration approach that uses only neighborhood frames. This integration analysis does not need phoneme transcriptions, can estimate stable spectral envelopes without using precise pitch marks, and can determine the *amplitude range* for each frequency bin of the spectral envelope. Like the other methods, our method has difficulty identifying transfer characteristics between adjacent harmonics but this amplitude range information provides a better estimate of the spectral envelope.

2. F_0-adaptive multi-frame integration analysis

Figure 1 shows an overview of the proposed F_0-adaptive multi-frame integration analysis. First, F_0-adaptive Gaussian windows are used for spectrum analysis (F_0*-adaptive* analysis). Second, neighborhood frames are integrated to estimate the target spectral envelope (*multi-frame integration* analysis).

Overlapped windowed waveforms in which F_0-adaptive Gaussian windows are shifted at each wave sample are shown in Figure 2 along with their corresponding short-term Fourier transform (STFT) results (*overlapped spectra*) and their group delays[2]. This figure suggests that we could obtain stable spectral envelopes by integrating these STFT results. When a single window without any integration is used, it often suffers from problematic "valleys" in the spectrum as shown in the middle of Figure 2. These valleys are interference effects caused by the distance between the center of a Gaussian window and the position of the periodically localized energy of the audio signal. Note that these valleys also correspond to positive and negative peaks in the group delays as shown in the bottom of Figure 2, where each peak means that the localized energy is far from the center of the window.

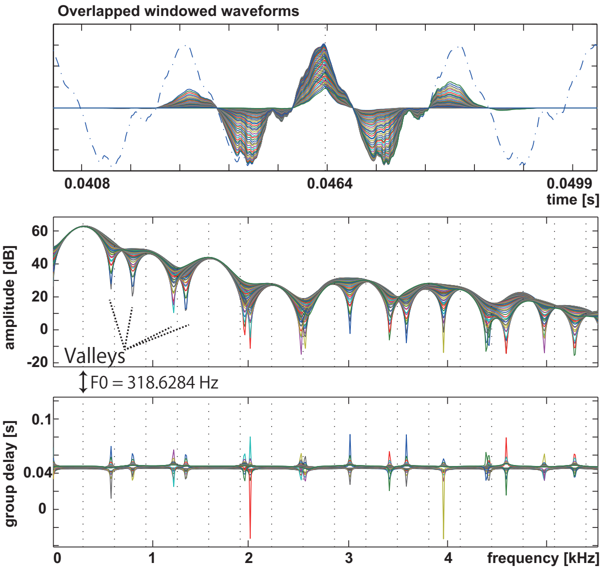

Figure 2: *Overlapped windowed waveforms for F_0-adaptive Gaussian windows (top). Corresponding STFT results (middle) and group delays (bottom).*

Our F_0-adaptive multi-frame integration analysis therefore integrates the STFT results (overlapped spectra) using multiple Gaussian windows to fill in these valleys. As shown in Figure 3, we can obtain the maximum and minimum envelopes from the overlapped spectra. We call the areas between the maximum and minimum envelopes *amplitude ranges*. This amplitude range information is an informative representation because the actual spectral envelope without any problematic valleys should always stay in this area. Our method finally estimates the target spectra envelope by averaging the maximum and minimum envelopes so that the estimated envelope can be in this area.

2.1. Conditions for implementation

To estimate a spectral envelope by F_0-adaptive analysis, we assume that the F_0 of a signal has already been estimated by any appropriate technique. Throughout this paper, sound samples are monaural recordings without accompaniment or noise and are digitized at 16 bit / 44.1 kHz.

[2] The group delay is computed by using a method described in [32].

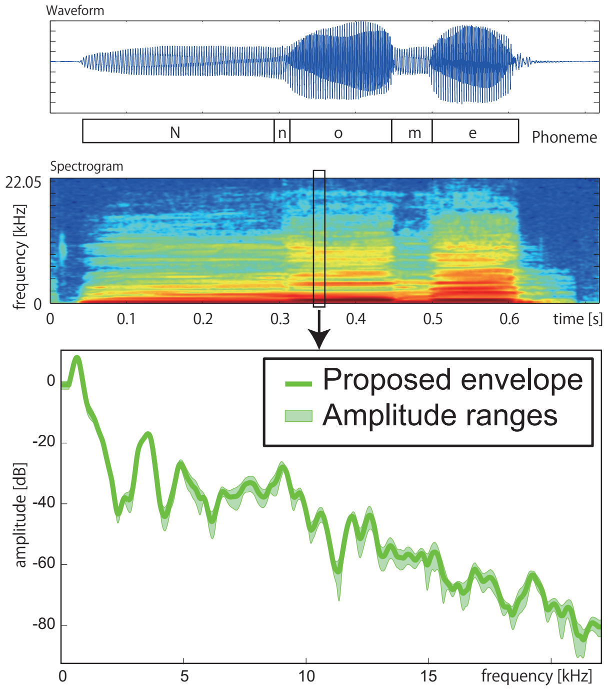

Figure 3: *A waveform of singing (top) and its spectral envelope and amplitude ranges estimated by F_0-adaptive multi-frame integration analysis (middle and bottom).*

2.2. F_0-adaptive analysis

We designed an F_0-adaptive window by using a Gaussian function (Figure 1). Let $w(\tau)$ be a Gaussian window function of time τ defined as follows, where $\sigma(t)$ is the standard deviation of the Gaussian distribution and $F_0(t)$ is the fundamental frequency for analysis time t. The window is normalized by the root mean square (RMS) value calculated with N defined as the FFT length.

$$w(\tau) = \frac{\hat{w}(\tau)}{\sqrt{(1/N)\sum_{\tau=0}^{N-1}\hat{w}(\tau)^2}} \quad (1)$$

$$\hat{w}(\tau) = \exp(-\frac{\tau^2}{2\sigma(t)^2}) \quad (2)$$

$$\sigma(t) = \frac{1}{F_0(t)} \times \frac{1}{3} \quad (3)$$

The Gaussian window's $\sigma(t) = 1/(3 \times F_0(t))$ means the analysis window length ($2 \times 3\sigma$) is set to two fundamental periods ($2 \times 3\sigma = 2/F_0(t)$, see Figure 1). This length is known to give a good approximation of the local spectral envelope [1]. The results of F_0-adaptive window analysis have F_0-related fluctuation along the time axis as shown in the bottom part of Figure 4.

Since the Gaussian windows are shifted at each wave sample, the discrete time-shift of the F_0-adaptive analysis is set to one wave sample ($1/44100$ s).

2.3. Multi-frame integration analysis

This multi-frame integration analysis first overlaps neighborhood spectra as shown in Figure 2. The range of the overlapping is from $-1/(2 \times F_0)$ to $1/(2 \times F_0)$, which can cover a period of harmonic signals (Figure 1). The results of our preliminary experiment showed that a smaller range is not sufficient to fill in the spectral valleys.

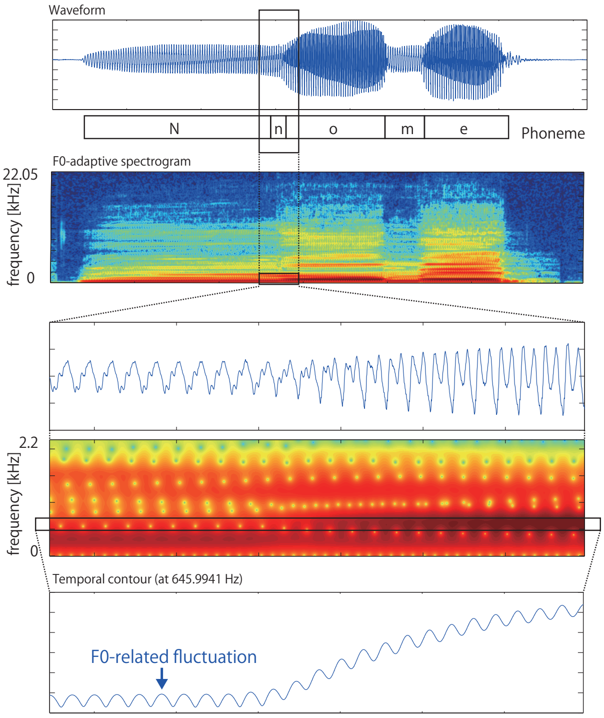

Figure 4: *A waveform and its F_0-adaptive spectrogram without pitch marks (top). A close-up view of the waveform and the spectrogram (middle), and a temporal contour of the spectrogram at 645.9961 Hz (bottom).*

By integrating the neighborhood spectra, one obtains the maximum and minimum envelopes of those spectra as the upper and lower bounds of the target spectral envelope. Our method then estimates the target envelope by averaging the maximum and minimum envelopes. Before this averaging, however, the valleys in the minimum envelope should be eliminated because frequency components around the valleys are not appropriately observed. For this elimination, we transform the maximum envelope into the *new* valley-free minimum envelope that touches positive peaks of the *old* (original) minimum envelope with valleys as shown in Figure 5. After the ratio of this transformation at each peak (at its frequency bin) is set, the ratio of the transformation at the other frequency bins is linearly interpolated and adjusted so that the transformed envelope can be higher than the old minimum envelope.

Before the maximum and new minimum envelopes are averaged, they are smoothed along both the time and frequency axes by using an 2-dimensional FIR low-pass filter. As shown in Figure 6, for example, the temporal contour of the maximum envelope is much smoother than the F0-adaptive spectrum before the integration, but still has a step-like contour that can be further smoothed by the low-pass filter.

Since the estimated envelope of frequency bins under F_0 is known to be unreliable, we finally smooth the envelope of those bins so that it can be same with the value of the envelope at F_0 as shown in Figure 5.

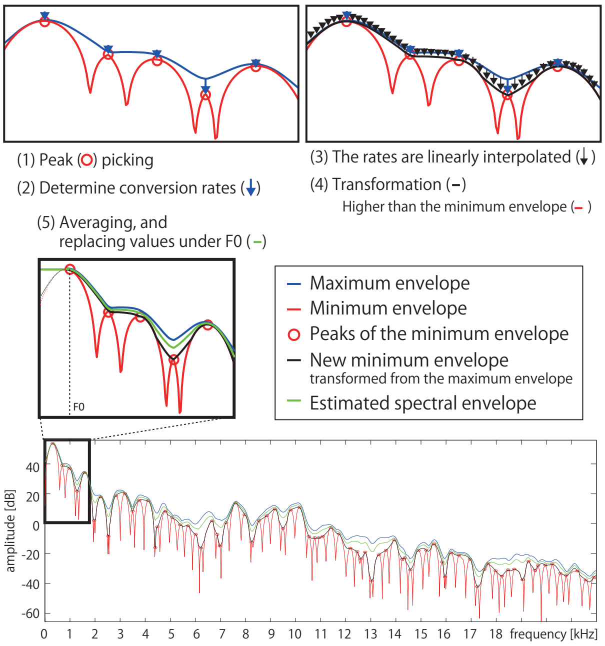

Figure 5: *The target spectral envelope (green line) is estimated by averaging the maximum envelope (blue line) and the new minimum envelope (black line).*

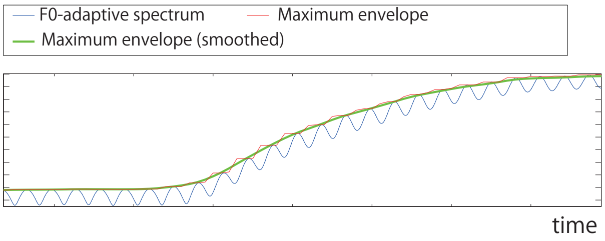

Figure 6: *Temporal contours of the maximum envelope before and after the low-pass filter. Integration smoothes the* F_0*-adaptive spectrum (blue line, same as that in the bottom of Figure 4) to the maximum envelope (red line), which still has steps that can be eliminated (green line).*

3. Experimental evaluations

In two experiments the proposed method was compared with two previous methods, STRAIGHT [24] and TANDEM-STRAIGHT [25]. An unaccompanied male singing sound (solo vocal) was taken from the RWC Music Database[3], a female spoken sound was taken from the AIST Humming Database (E008) [34], and two kinds of instrument sounds were taken from the RWC Music Database[4].

All spectral envelopes were represented by 2049 frequency bins (4096 FFT length) and had a temporal resolution of 1 ms, which was the discrete time step in the analyses.

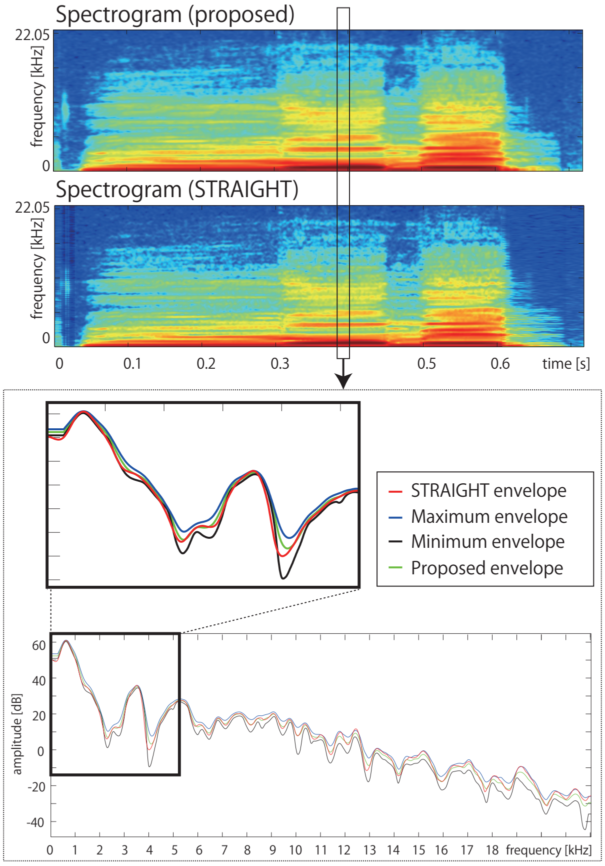

Figure 7: *Comparison of the spectrogram estimated by the proposed method (top) and the STRAIGHT spectrogram (middle). The corresponding spectral envelopes at time* 0.4 *sec can also be compared (bottom).*

3.1. Experiment A: comparison

By using natural sound samples, we compared spectral envelopes estimated by the proposed method with spectral envelopes estimated by STRAIGHT [24], which are known to be highly accurate.

In Figure 7 the frequency spectrogram obtained using our proposed method is shown along with the corresponding STRAIGHT spectrogram. The spectral envelope obtained using STRAIGHT lies between our maximum and minimum envelopes and is similar to the proposed envelope. We also found that sounds resynthesized using the spectrogram obtained with our method can give an impression of naturalness comparable to that given by sounds resynthesized using the STRAIGHT spectrogram.

3.2. Experiment B: simulation

We evaluated the accuracy of the spectral envelope estimation by comparing the proposed method with STRAIGHT [24] and TANDEM-STRAIGHT [25]. To prepare the ground truth for evaluating its accuracy, we used sounds resynthesized from natural sound samples by using STRAIGHT and also used sounds synthesized by a Klatt synthesizer.

For sounds resynthesized by using STRAIGHT, we first analyzed the natural sound samples (singing, speech, and instrument sounds) to estimate the F_0 and the STRAIGHT spec-

[3] Music Genre [33]: RWC-MDB-G-2001 No. 91.

[4] Musical Instrument Sound [33]: RWC-MDB-I-2001 No. 01 011PFNOM as a piano sound, and No.16 161VLGLM as a violin sound.

Table 1: *Control parameters for cascade-type Klatt synthesizer [35] in experiment B.*

Symbol	Name	Value (Hz)
F0	Fundamental frequency	125
F1	First formant frequency	250–1250
F2	Second formant frequency	750–2250
F3	Third formant frequency	2500
F4	Fourth formant frequency	3500
F5	Fifth formant frequency	4500
B1	First formant bandwidth	62.5
B2	Second formant bandwidth	62.5
B3	Third formant bandwidth	125
B4	Fourth formant bandwidth	125
B5	Fifth formant bandwidth	125
FGP	Glottal resonator frequency	0
BGP	Glottal resonator bandwidth	100

Table 2: *F1 and F2 values for cascade-type Klatt synthesizer [35] in experiment B.*

Sample No.	F1 (Hz)	F2 (Hz)	Sample No.	F1 (Hz)	F2 (Hz)
K01	250	750	K04	1000	1500
K02	250	1500	K05	1000	2000
K03	500	1500	K06	500	2000

trogram. We then resynthesized sound samples from the STRAIGHT spectrogram with different F_0s. This STRAIGHT spectrogram can be used as the ground truth for these resynthesized sound samples.

We synthesized six voice-like sounds by using a cascade-type Klatt synthesizer [35] with the parameters listed in Table 1. F1 and F2 for the six sounds were set as shown in Table 2. Note that we just obtained spectral envelopes from the Klatt synthesizer for the ground truth. We then used our own sinusoidal synthesis implementation using the obtained spectral envelopes to generate sound samples having the F_0 of 125 Hz.

The spectral envelopes (spectrogram) estimated by the three methods were evaluated in terms of the log-spectral distance LSD defined by

$$LSD = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{F}\sum_{f=F_L}^{F_H}\left|20\log_{10}\frac{S_g(t,f)}{\alpha(t)\cdot S_e(t,f)}\right| \quad (4)$$

$$\alpha(t) = \frac{\sum_{f=F_L}^{F_H} S_g(t,f)S_e(t,f)}{\sum_{f=F_L}^{F_H} S_e(t,f)^2} \quad (5)$$

$$\epsilon^2 = \sum_{f=F_L}^{F_H}\left(S_g(t,f) - \alpha(t)S_e(t,f)\right) \quad (6)$$

where T is the number of voiced frames, F is the number of frequency bins ($= F_H - F_L + 1$), (F_L, F_H) is the frequency range for the evaluation, $S_g(t,f)$ and $S_e(t,f)$ are respectively the ground-truth and estimated spectral envelopes, and $\alpha(t)$ is a normalization factor determined by minimizing an error defined as a square error ϵ^2 between $S_g(t,f)$ and $\alpha(t)S_e(t,f)$ at each frame t. The log-spectral distances obtained in this experiment are listed in Table 3, and an estimated result is shown in Figure 8.

The envelope estimation accuracies obtained using our proposed method are sometimes better than those obtained using STRAIGHT and TANDEM-STRAIGHT. This suggests that our method can be used for high-quality sound synthesis and high-accuracy sound analysis.

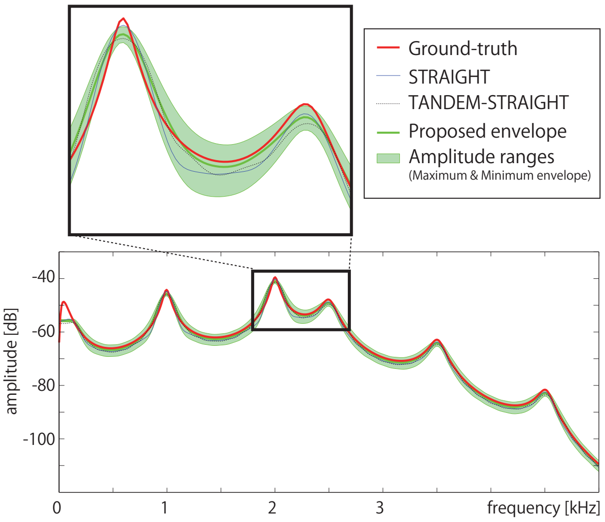

Figure 8: *Comparison of Klatt (K05) spectral envelopes estimated by the proposed method and two previous methods.*

4. Conclusion

This paper proposed a simple, highly accurate, and flexible spectral envelope estimation method called F_0-adaptive multi-frame integration analysis. The method was evaluated by using synthesized/resynthesized sound samples and comparing its results with those of two well-known methods for spectral envelope estimation. The proposed method can estimate accurate spectral envelopes with amplitude range information useful for flexible representation. The evaluation results suggest that envelopes estimated by the proposed method can be used for sound synthesis/analysis.

In future work we expect to improve the envelope estimation method, to investigate a method for estimating aperiodicity, phase information, and excitation, and to incorporate statistical techniques and additional information (*e.g.*, phoneme transcriptions).

5. Acknowledgements

This research utilized the RWC Music Database “RWC-MDB-G-2001” (Music Genre), “RWC-MDB-I-2001” (Musical Instrument Sound), and AIST Humming Database (E008). This research was supported in part by OngaCREST project (CREST, JST).

6. References

[1] U. Zölzer and X. Amatriain, *DAFX - Digital Audio Effects*. Wiley, 2002.

[2] J. Flanagan and R. Golden, “Phase vocoder,” *Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[3] D. W. Griffin, *Multi-Band Excitation Vocoder*. Technical report (Massachusetts Institute of Technology. Research Laboratory of Electronics), 1987.

[4] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[5] F. Itakura and S. Saito, “Analysis synthesis telephony based upon the maximum likelihood method,” in *Proc. 6th ICA*, 1968, pp. C17–20.

[6] B. S. Atal and S. Hanauer, “Speech analysis and synthesis by linear

Table 3: *Log-spectral distances for spectrograms obtained in experiment B when using two previous methods and the proposed method. The smallest values are* underlined*, and the second smallest values are indicated by* **bold-faced type***.*

sound	length	F_L	F_H	LSD (log-spectral distance) [dB]		
type	[s]	[kHz]	[kHz]	STRAIGHT	TANDEM	Proposed
singing (male)	6.5	0	6	1.0981	1.9388	**1.4314**
singing (male)	6.5	0	22.05	**2.0682**	2.3215	2.0538
speech (female)	4.6	0	6	**2.1068**	2.3434	2.0588
speech (female)	4.6	0	22.05	2.7937	**2.7722**	2.5908
instrument (piano)	2.9	0	6	3.6600	**3.4127**	3.1232
instrument (piano)	2.9	0	22.05	4.0024	**3.5951**	3.3649
instrument (violin)	3.6	0	6	1.1467	1.7994	**1.3794**
instrument (violin)	3.6	0	22.05	**2.2711**	2.3689	2.1012
Klatt (K01)	0.2	0	5	2.3131	1.6676	**1.9491**
Klatt (K02)	0.2	0	5	3.8462	1.5995	**2.8278**
Klatt (K03)	0.2	0	5	**1.6764**	1.4700	2.2954
Klatt (K04)	0.2	0	5	1.7053	**1.2699**	1.1271
Klatt (K05)	0.2	0	5	1.5759	**1.2353**	1.0643
Klatt (K06)	0.2	0	5	1.1712	**1.2662**	1.8197

prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 4, pp. 637–655, 1971.

[7] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," in *Proc. ICSLP1994*, 1994, pp. 1043–1045.

[8] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *IEICE Trans. A (in Japanese)*, vol. J62-A, no. 4, pp. 217–223, 1979.

[9] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. DAFx2005*, 2005, pp. 30–35.

[10] F. Villavicencio, A. Röbel, and X. Rodet, "Extending efficient spectral envelope modeling to mel-frequency based representation," in *Proc. ICASSP2008*, 2008, pp. 1625–1628.

[11] ——, "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation," in *Proc. ICASSP2006*, 2006, pp. 869–872.

[12] R. McAulay and T.Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. ASSP*, vol. 34, no. 4, pp. 744–755, 1986.

[13] J. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. ICMC 1987*, 1987, pp. 290–297.

[14] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[15] Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*.

[16] P. Depalle and T. Hélie, "Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows," in *Proc. WASPAA1997*, 1997.

[17] E. George and M. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, no. 6, pp. 497–515, 1992.

[18] Y. Pantazis, O. Rosec, and Y. Stylianou, "Iterative estimation of sinusoidal signal parameters," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 461–464, 2010.

[19] M. Abe and J. O. Smith III, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," in *Proc. AES 117th Convention*, 2004.

[20] J. Bonada, "Wide-band harmonic sinusoidal modeling," in *Proc. DAFx-08*, 2008, pp. 265–272.

[21] H. Kameoka, N. Ono, and S. Sagayama, "Auxiliary function approach to parameter estimation of constrained sinusoidal model for monaural speech separation," in *Proc. ICASSP 2008*, 2008, pp. 29–32.

[22] M. Ito and M. Yano, "Sinusoidal modeling for nonstationary voiced speech based on a local vector transform," *J. Acoust. Soc. Am.*, vol. 121, no. 3, pp. 1717–1727, 2007.

[23] A. Pavlovets and A. Petrovsky, "Robust HNR-based closed-loop pitch and harmonic parameters estimation," in *Proc. INTERSPEECH2011*, 2011, pp. 1981–1984.

[24] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency based on F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

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

[26] P. Zolfaghari, S. Watanabe, A. Nakamura, and S. Katagiri, "Modelling of the speech spectrum using mixture of gaussians," in *Proc. ICASSP 2004*, 2004, pp. 553–556.

[27] H. Kameoka, N. Ono, and S. Sagayama, "Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency," vol. 18, no. 6, pp. 2502–2505, 2006.

[28] M. Akamine and T. Kagoshima, "Analytic generation of synthesis units by closed loop training for totally speaker driven text to tpeech system (TOS Drive TTS)," in *Proc. ICSLP1998*, 1998, pp. 1927–1930.

[29] Y. Shiga and S. King, "Estimating the spectral envelope of voiced speech using multi-frame analysis," in *Proc. EUROSPEECH2003*, 2003, pp. 1737–1740.

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

[31] H. Fujihara, M. Goto, and H. G. Okuno, "A novel framework for recognizing phonemes of singing voice in polyphonic music," in *Proc. WASPAA2009*, 2009, pp. 17–20.

[32] H. Banno, L. Jinlin, S. Nakamura, K. Shikano, and H. Kawahara, "Efficient representation of short-time phase based on group delay," in *Proc. ICASSP1998*, 1998, pp. 861–864.

[33] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. of ISMIR 2003*, 2003, pp. 229–230.

[34] M. Goto and T. Nishimura, "AIST humming database: Music database for singing research," in *IPSJ SIG Notes (in Japanese)*, ser. 2005-MUS-61, 2005, pp. 7–12.

[35] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971–995, 1980.