Modulation transfer function design for a flexible cross synthesis VOCODER based on F0 adaptive spectral envelope recovery

Taiki Nishi*, Ryuichi Nisimura†, Toshio Irino† and Hideki Kawahara†
* Graduate School of Systems Engineering, Wakayama University, Sakaedani 930,Wakayama 640-8510, Japan
E-mail: s125038@center.wakayama-u.ac.jp
† Wakayama University, Sakaedani 930,Wakayama 640-8510, Japan
E-mail: {nisimura,irino,kawahara}@sys.wakayama-u.ac.jp

***Abstract*—A new design procedure for flexible cross synthesis VOCODER is proposed based on TANDEM-STRAIGHT framework, a F0 adaptive spectral envelope estimator, and modulation transfer function design. The proposed design procedure enables control of speech intelligibility and timber identity of musical instruments or animal voices. Removal of the averaged and smoothed logarithmic spectrum of speech from the filter reduced the timbre modification effect of filtered sounds and manipulation of cut-off frequencies of modulation transfer function for designing the filter enabled control of trade-offs between intelligibility and timbre preservation.**

I. Introduction

Cross synthesis is a sound effect by mixing features of two sounds to make unique sounds, such as; talking piano, shouting whistle, speaking animal and so on. Current cross synthesis effectors are implemented using a variety of techniques [1], [2], such as Channel VOCODER[3], Phase VOCODER [4] LPC[5], [6] and generalized cepstral analysis[7]. However, current cross synthesis effects tend to deteriorate original characters of musical instruments and usually processed sounds are not very intelligible.

In this report, in order to solve these problems, a new cross synthesis framework, based on an interference-free representation of power spectrum and modulation transfer function design is proposed. The interference-free power spectral representation is estimated using TANDEM-STRAIGHT [8] and is used to calculate global spectral shape, which is the main contributing factor of timber deterioration. The modulation transfer function design is aiming at removing linguistically less important temporal modulation components from the time-frequency representation of speech. The temporally variable filtering operation in cross synthesis is approximately implemented using a series of time invariant FFT-based convolution with minimum phase responses calculated from the modified time-frequency representation. Subjective tests were conducted to evaluate effects of these factors on a) preservation of timbre of musical instruments or animal identity and b) "clearness" impression of linguistic information.

This work is partly supported by Grant in Aid for Scientific Research of JSPA and advanced research project of Wakayama University Japan.

In the following sections, these component procedures are introduced followed by implementation details. Finally, subjective test results and discussions are presented.

II. Spectrum analysis

In the proposed method, the envelope spectrum of speech is extracted using a F0-adaptive procedure called TANDEM-STRAIGHT [8]. The extracted spectrum does not contain any trace of periodic interferences which is inevitably found in conventional short term Fourier transform. In this section, we briefly introduce underlying principles of TANDEM-STRAIGHT.

The spectrum envelope extraction procedure in the TANDEM-STRAIGHT framework contains two sub procedures: the TANDEM and STRAIGHT procedures. The TANDEM procedure yields temporally stable power spectrum of periodic signals. The STRAIGHT procedure removes frequency domain variations from the temporally stable power spectrum extracted in the TANDEM procedure and recovers underlying smooth representation.

A. TANDEM Spectrum

The periodic temporal variations in the power spectrum, which is calculated using the short term Fourier transform, are completely cancelled out by averaging power spectra calculated at two locations a half fundamental period apart. The temporally stable power spectrum (TANDEM spectrum) $P_T(\omega, t)$ is calculated by the following equation.

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

where $P(\omega, t)$ is the power spectrum calculated by the short term Fourier analysis and T_0 is the fundamental period.

B. STRAIGHT Spectrum

The TANDEM spectrum still has periodic variations on the frequency axis. The STRAIGHT procedure removes this variations using a F0-adaptive rectangular smoother and compensate for over smoothing by using a digital filter on the frequency axis to preserves spectral levels at harmonic frequencies by adopting the consistent sampling theory [11].

These two procedures on the frequency axis are implemented using Cepstrum lifter to make resultant spectral envelope positive definite. The STRAIGHT spectrum $P_{TST}(\omega,t)$ is calculated from the TANDEM spectrum $P_T(\omega,t)$.

$$P_{TST}(\omega,t) = \exp(\mathcal{F}[g_1(\tau)g_2(\tau)C_T(\tau,t)]) \quad (2)$$

$$\text{where} \quad g_1(\tau) = \tilde{q}_0 + 2\tilde{q}_1\cos(\frac{2\pi\tau}{T_0}) \quad (3)$$

$$g_2(\tau) = \frac{\sin(\pi f_0\tau)}{\pi f_0\tau} = \mathcal{F}[h_2(\omega)] \quad (4)$$

$$h_2(\omega) = \begin{cases} \frac{1}{\omega_0} & |\omega| \le \frac{\omega_0}{2} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$C_T(\tau,t) = \mathcal{F}^{-1}[\ln(P_T(\omega,t))], \quad (6)$$

where $C_T(\omega,t)$ is the Cepstrum of the TANDEM spectrum $P_T(\omega,t)$ and τ is quefrency. $\mathcal{F}$ represents the Fourier transform. The lifter $g_1(\tau)$ is for compensation of over smoothing and $g_2(\tau)$ is the equivalent lifter of the F0-adaptive rectangular smoother $h_2(\omega)$ where $\omega_0 = 2\pi/T_0$ is the fundamental angular frequency. Since all these are real valued variables and functions, exponential conversion of the processed result is always positive. Details can be found in our previous work [12]. In the latest implementation, the lifter coefficients, $\tilde{q}_0$ and $\tilde{q}_1$ are optimized to improve perceptual quality of the processed speech, based on extensive simulation tests [13].

III. SPEECH SPECTRUM SUITABLE FOR TIMBRE PRESERVATION OF MUSICAL INSTRUMENTS

In the conventional cross synthesis VOCODERs, timbre of the original musical instruments are modified by speech and their identity are not preserved intact. In this section, we systematically remove the primary timbre modifying component ,the global spectral shape, from the speech spectrum. The following two-staged procedure is introduced to calculate the global spectral shape from relatively short speech segment: long term average spectrum estimation and spectral smoothing.

A. Long term averaged spectrum

Generally the long term averaged power spectrum $\overline{P_t(\omega)}$ is calculated using Welch's method [14].

$$\overline{P_t(\omega)} = \frac{1}{T_s}\int_0^{T_s} P(\omega,t)dt, \quad (7)$$

where T_s is the length of the utterance. However, this conventional definition of the long term averaged spectrum is not relevant for perceptual compensation, because our loudness perception is not directly proportional to the power spectrum. Instead of using the power spectrum directly, we introduce a monotonic nonlinear function $g(x)$ and its inverse function $g^{-1}(x)$ to approximate perceptual long term averaged spectrum $\overline{P_{t,g}(\omega)}$.

$$\overline{P_{t,g}(\omega)} = g^{-1}\left(\frac{1}{T_s}\int_0^{T_s} g(P(\omega,t))dt\right). \quad (8)$$

There are two candidates for $g(x)$ to simulate loudness perception: the power law and the logarithmic function. In this article, we selected the logarithmic function because of its conceptual simplicity. The following equation defines the log-averaged power spectrum $\overline{P_{t,\ln}(\omega)}$.

$$\overline{P_{t,\ln}(\omega)} = \exp\left(\frac{1}{T_s}\int_0^{T_s} \ln(P(\omega,t))dt\right). \quad (9)$$

Note that STRAIGHT spectrum $P_{TST}(\omega,t)$, it is a power spectral envelope, is used to calculate the long term averaged power spectrum $\overline{P_t(\omega)}$ and the log-averaged power spectrum $\overline{P_{t,\ln}(\omega)}$ instead of the STFT power spectrum $P(\omega,t)$ in the actual implementation.

B. Global spectral shape

A spectral smoothing procedure in the perceptual frequency domain is introduced to calculate the global spectral shape. It is because of the following reasons. Neither the long term averaged power spectrum nor the log-averaged power spectrum calculated from short utterances are not directly usable for spectral compensation because they consist of noisy details due to statistical fluctuations. There are two factors to design the global spectral shape for perceptual compensation: that the global spectral shape of the excitation source of speech is smooth and that the front end of our auditory system is a set of band-pass filters. Taking these into account, the global spectral shape $\tilde{P}_a(\omega)$ is approximated by smoothing proportionally to the perceptual frequency resolution.

$$\tilde{P}_a(\omega) = \frac{1}{C(\omega)}\int_{\lambda^{-1}(\lambda(\omega)-a/2)}^{\lambda^{-1}(\lambda(\omega)+a/2)} \overline{P_{t,\ln}(q)}dq, \quad (10)$$
$$C(\omega) = \lambda^{-1}(\lambda(\omega)+a/2) - \lambda^{-1}(\lambda(\omega)-a/2)$$

where q is the angular frequency. The function $\lambda(\omega)$ converts the angular frequency ω to the perceptual frequency (specifically, ERB_N number [15] is used) and the inverse function $\lambda^{-1}(\lambda)$ converts the ERB_N number λ to the angular frequency. The parameter a defines smoothing width in the ERB_N number domain.

C. Whitening using the global spectral shape

STRAIGHT spectrum of each frame $P_{TST}(\omega,t)$ is whitened by the global spectral shape $\tilde{P}_a(\omega)$ to yield the compensated spectral envelope $P_C(\omega,t)$.

$$P_C(\omega,t) = \frac{P_{TST}(\omega,t)}{\tilde{P}_a(\omega)}. \quad (11)$$

Figure 1 shows the STRAIGHT spectrogram of a Japanese utterance /koNnitiwa/ ("Hello" or "good day" in English) spoken by a male. Figure 2 shows the whitened STRAIGHT spectrogram of the same utterance shown in Fig. 1.

The compensated spectral envelope $P_C(\omega,t)$ still has the other timbre modifying components in the modulation frequency domain. The following section introduces filtering operation in the modulation frequency domain to remove such components.

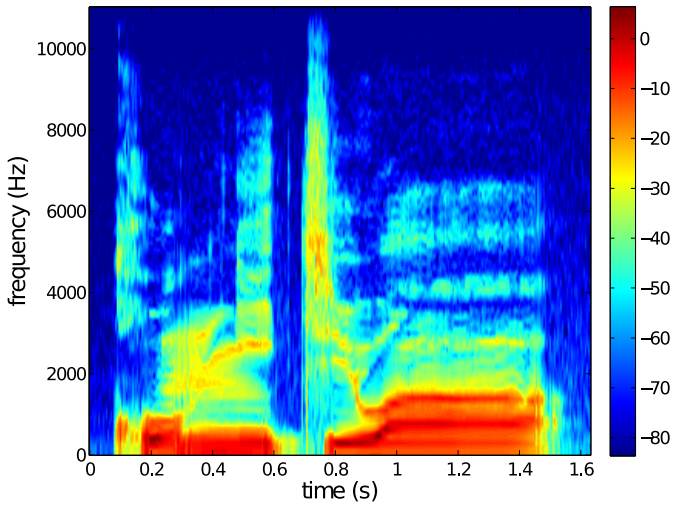

Fig. 1. STRAIGHT spectrogram of a Japanese utterance /koNnitiwa/ (“Hello” or “good day” in English) spoken by a male.

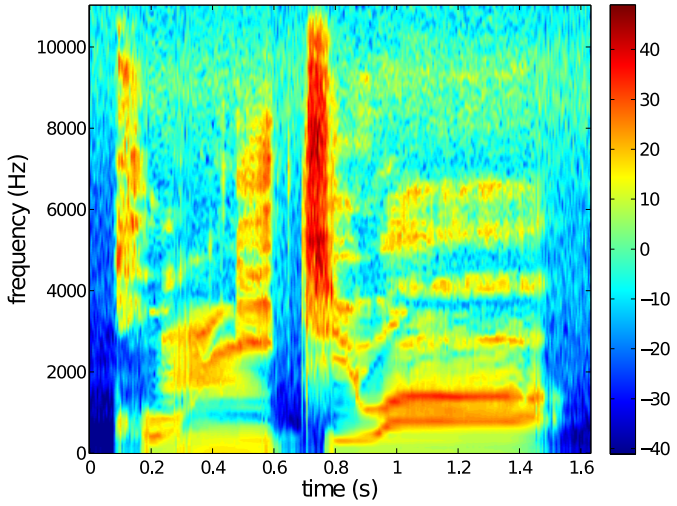

Fig. 2. Whitened STRAIGHT spectrogram of the same utterance shown in Fig. 1

IV. Band-pass filtering in the modulation frequency domain

To remove the remaining timbre modifying components from the whitened STRAIGHT spectrum, the band-pass filtering in the modulation frequency domain is introduced. This is our original contribution to the cross synthesis applications. This idea is motivated by findings that the linguistic information of speech is distributed in a limited modulation frequency region [16], [17], [18]. We thought that by removing the modulation frequency components which is not important for the linguistic contents of speech sounds can reduce interference to the identity of the musical instruments and/or the animal voices. Note that more general two dimensional filtering in the time-frequency representation is probably not necessary for the perceptual compensation in cross synthesis. Please refer to the literature [19].

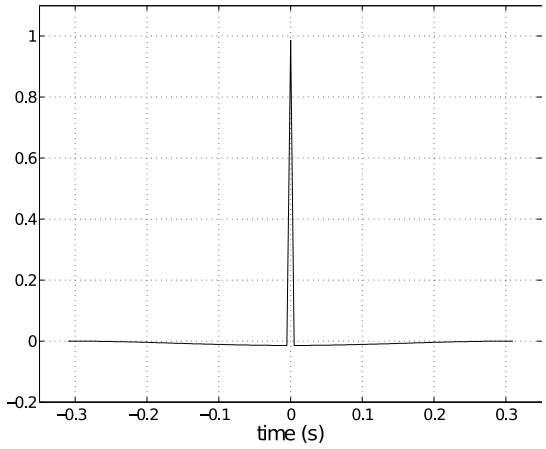

Fig. 3. Impulse response of the high-pass filter

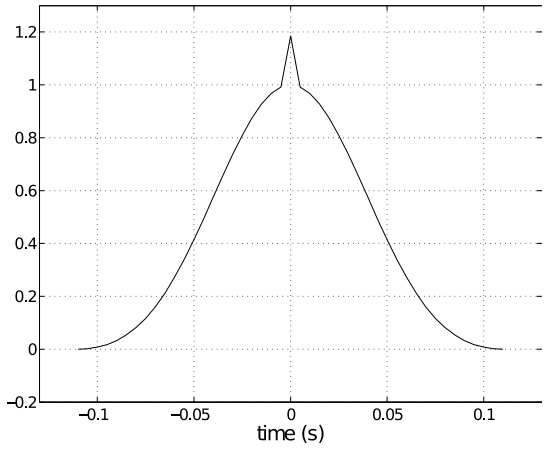

Fig. 4. Impulse response of the low-pass filter

A. *filtering in the modulation frequency domain*

Filtering in the modulation frequency domain is performed on the logarithmic version of the compensated STRAIGHT spectrum $L_C(\omega,t) = \ln(P_C(\omega,t))$ and converted back to the power spectrum $P_B(\omega,t)$ for designing the time varying filter for cross synthesis.

$$L_H(\omega,t) = \int_{-T_H}^{T_H} h_H(\tau) L_C(\omega, t-\tau) d\tau \quad (12)$$

$$P_B(\omega,t) = \exp\left(\int_{-T_L}^{T_L} h_L(\tau) L_H(\omega, t-\tau) d\tau \right), \quad (13)$$

where $h_H(\tau)$ and $h_L(\tau)$ are impulse responses of the high-pass and low-pass filters in the modulation frequency domain, respectively. The impulse responses are temporally bounded in $(-T_H, T_H)$ for the high-pass filter and $(-T_L, T_L)$ for the low-pass filter. Filtering on logarithmic spectra assures that the resulted spectrum $P_B(\omega,t)$ is always positive.

Linear phase FIR filters are used in this procedure in order to avoid phase distortion artifacts pointed out by the literature [18]. The FIR filters are constructed using a time

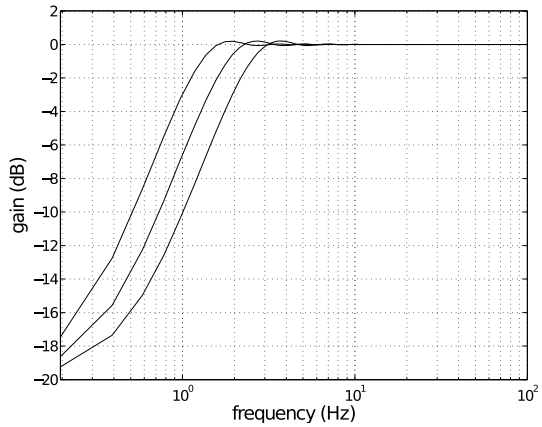

Fig. 5. Modulation transfer function of the high-pass filters

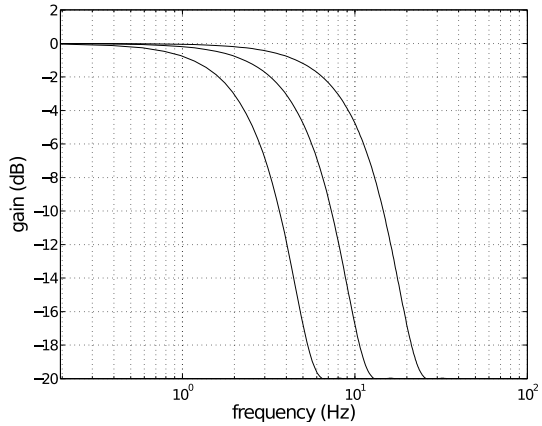

Fig. 6. Modulation transfer function of the low-pass filters

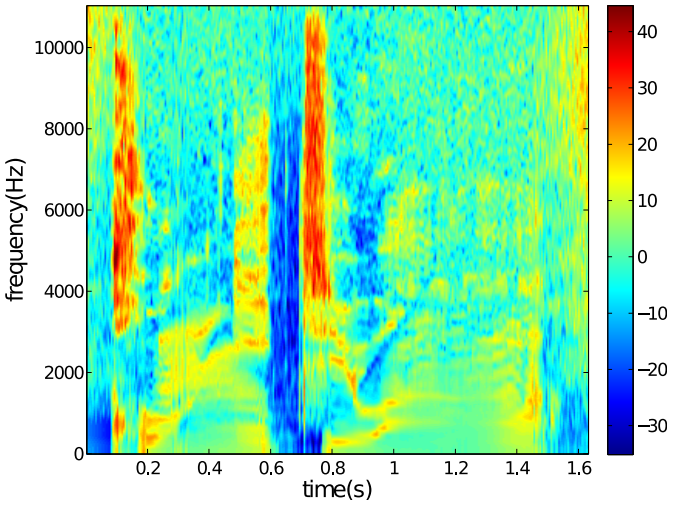

Fig. 7. Spectrogram after high-pass modulation filtering with 2 Hz cut-off

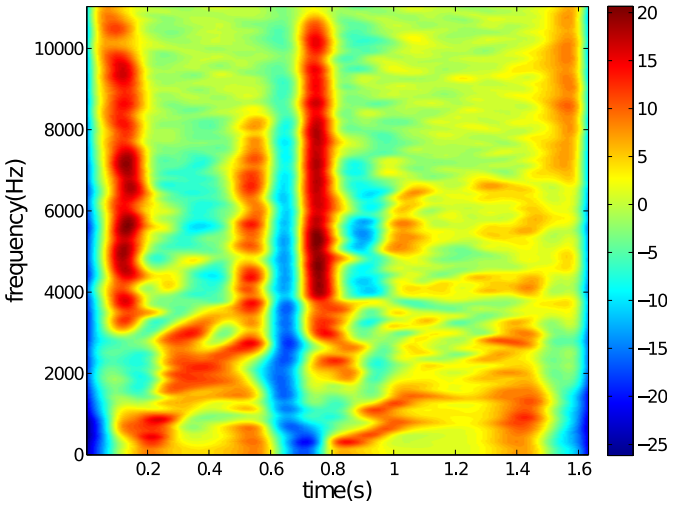

Fig. 8. Spectrogram after low-pass and high-pass modulation filtering. The low-pass cut-off is 4 Hz and the high-pass cut-off is 2 Hz

windowing function (specifically Hann window in the current implementation) in order to reduce temporal oscillation, which is commonly found in the impulse response of the filters having sharp cut-off characteristics. Figures 3 and 4 show impulse responses of one of the high-pass filters and the low-pass filters used in the experiments.

Figures 5 and 6 shows their frequency responses. The nominal cut-off frequencies were defined at -3 dB points. The high-pass nominal frequencies were set to 1, $\sqrt{2}$ and 2 Hz and the low-pass nominal frequencies were set to 2, 4, and 8 Hz. The attenuation level outside of the pass-band is set to 20 dB. Note that the components outside of the nominal pass-band still have significant power because of the filters' less steep cut-off.

Figure 7 shows the compensated spectrogram after high-pass modulation filtering ($L_H(\omega, t)$ represented in dB) . Note that the temporal transitions are enhanced. Figure 8 shows the compensated spectrogram after low-pass and high-pass modulation filtering ($P_B(\omega, t)$ represented in dB) . Note that the temporal details are smeared.

V. Implementation of Cross Synthesis

FFT-based high-speed convolution is used to implement the time varying filtering of cross synthesis. It is necessary to properly design this filtering because the FFT-based convolution is cyclic and is defined only for the time invariant systems.

A. Cyclic convolution and linear convolution

For implementing the linear convolution using the cyclic convolution, it is necessary for the FFT buffer length to be longer than the sum of the lengths of two signals to be convolved [20]. In case of cross synthesis, the effective length of the impulse response calculated from the compensated (and modulation frequency band-limitted) spectrogram has to be carefully investigated. The subdivided length of the musical instruments and/or animal voices can be set arbitrary and cannot be a problem for implementing the linear convolution using the cyclic convolution.

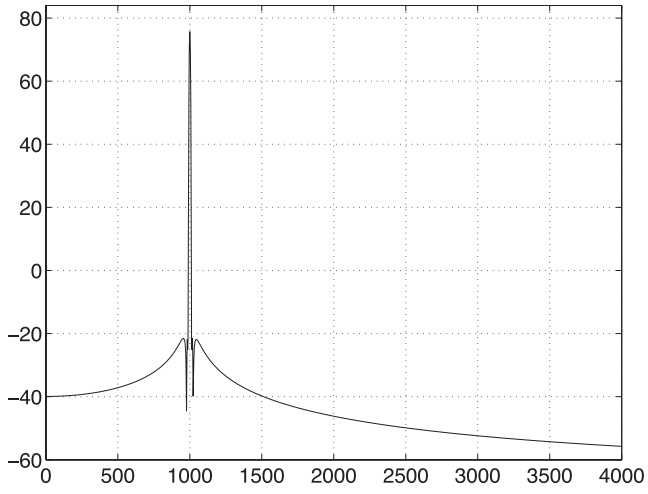

Fig. 9. Output power spectrum of a time invariant filter with exclusive subdivision of the input signal (1000 Hz sinusoid)

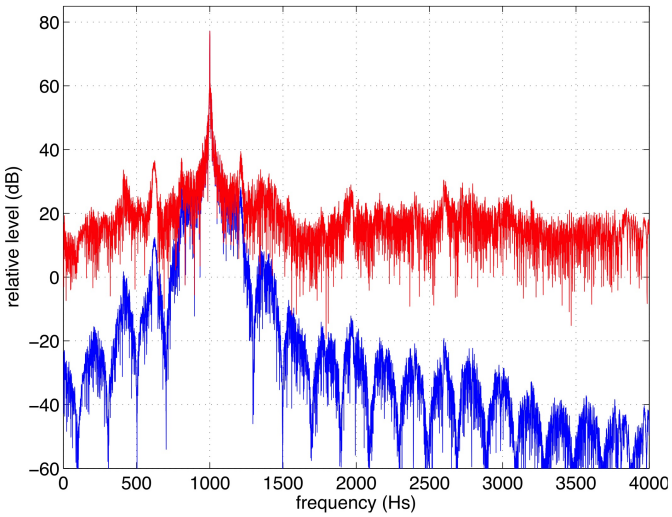

Fig. 10. Output power spectrum of a time varying filters with exclusive subdivision of the input signal (red line) and a 50% overlapping Hann windowing function (blue line). The input signal is 1000 Hz sinusoid

In the current implementation, the minimum-phase impulse response [20] is calculated form the spectral slice of the modulation-filtered compensated STRAIGHT spectrogram, $P_B(\omega, t)$. This decision is based on our observations that it sounds better than the linear phase FIR filter calculated from the same spectral slice and that the effective length of the minimum phase response is generally shorter than the linear phase response.

B. Approximation of time varying convolution

A 50% overlapping subdivision using the Hann windowing function is used to approximate the time varying filtering using the FFT-based high-speed convolution in our implemenation [21]. For the time invariant systems, mutually exclusive subdivision of the input signal is the most computationally efficient implementation [20]. However, the exclusive subdivision is not relevant for approximating the time varying convolution using segmentally time invariant subsystems. It is because the sudden change of the impulse response at frame boundaries effectively introduces the excessive step excitation in case of the mutually exclusive subdivision. This artifact is alleviated by using smoother subdivision methods. The 50% overlapping subdivision using the Hann windowing function is a practical solution to alleviate this problem.

Figure 9 shows the power spectrum of a time invariant filter output calculated using the FFT-based convolution with exclusive input subdivision. The input signal is a 1000 Hz sinusoid. Virtually, only the input sinusoidal component is visible in this output spectrum. Figure 10 shows the responses to the same sinusoidal input using different subdivision methods. The red line shows the output using the exclusive subdivision and the blue line shows the output using a 50% overlapping Hann windowing function-based subdivision. Note that the side-band levels, which is negligible when approximation is perfect, are about 40 dB higher when the exclusive subdivision is used.

TABLE I
THE SINUSOIDS TO TOTAL SIDE-BAND RATIOS (SNR) AS A FUNCTION OF FRAME UPDATE INTERVAL

	frame update interval (ms)					
	5	10	20	40	80	160
SNR (dB)	45.6	56.1	64.6	74.6	86.5	93.1

The power spectra were calculated using a four term cosine window with a very low side lobe level (lower than -90 dB) and fast side lobe decay (-18 dB/oct). The window is the 12th item of Table II in the literature [22]. The coefficients for zero through third cosines are 0.355768, 0.487396, 0.144232 and 0.012604, respectively.

In our implementation, 5 ms frame update is used in subdivision using the Hann windowing function. Although the side-band levels decrease by increasing frame update interval (Table I), accuracy of the approximation deteriorates. The rms errors from the output calculated using the 5 ms frame subdivision are shown in Table II. The input signal was a 1000 Hz sinusoid.

VI. SUBJECTIVE EVALUATION

Preservation of the linguistic information and the identity of the musical instruments and animal voices as a function of cut-off modulation frequencies were evaluated using subjective tests.

TABLE II
THE RMS ERROR VS. FRAME UPDATE INTERVAL

	frame update interval (ms)				
	10	20	40	80	160
rms error (dB)	-35.0	-32.9	-26.5	-18.7	-13.9

TABLE III
INPUT SIGNALS

Symbol	explanation
piano	Combination of C3, E3 and G3 of the piano sounds in No.1
guitar	Combined code C3 using the guitar sounds in No.13
cow	No.15 cow
horse	No.17 horse

TABLE IV
COMBINATIONS OF CUT-OFF FREQUENCIES OF THE HIGH-PASS AND LOW-PASS MODULATION FILTERS

low-pass cut-off frequencies (Hz)	high-pass cut-off frequencies (Hz)		
	1	$\sqrt{2}$	2
2	a	d	g
4	b	e	h
8	c	f	i

A. Test materials

An utterance /ohayoH gozaimasu/ (“Good morning” in English) spoken by a male is used to design the time varying filters. For the input signals, the musical instrument sounds were excerpted from “RWC music database: instrument sound” [23] and were edited. The animal voices were also excerpted from the CD titled “sound effects Complete ②: Animal, Bird, frog”. Table III shows input signals and their descriptions. All materials were sampled at 44,100 Hz with 16 bit resolution.

B. subjects

Nine subjects (six male and three female students aged from 21 to 24) were participated in the subjective evaluation tests. All subjects were native Japanese with normal hearing. Prior to the experiments, subjects were informed about the purpose and procedures of the experiments. All subjects agreed to participate in the experiments and recorded the written consent form.

C. Experimental conditions

Test stimuli were presented monaurally to the both ears using the headphones (SENNHEISER HD-580) in a sound proof room (YAMAHA AVITECS). The rms levels of the test stimuli were normalized to -26 dB from the 16 bit full scale range. The sound pressure level of the presented stimuli were set comfortable for subjects and measured afterwards using the HATS (B&K 4128). They were ranged from 60 dB to 70 dB in the A-weighting. The condition of cut-off frequencies and their symbols are shown in Table IV.

D. Experimental procedure

Two separate experiments were conducted for evaluating linguistic preservation and sound identity preservation respectively. Each experiment uses the paired comparison with forced choice of two alternatives. In the linguistic information preservation test, the original utterance and the stimulus S1 followed by the stimulus S2 were presented sequentially. The subjects were instructed to select the stimulus which preserves the original linguistic information better. In the timbre identity preservation test, the original sound (musical instruments or animal voices) and the stimulus S1 followed by the stimulus S2 were presented sequentially. The subjects were instructed to select the stimulus which preserves the original timbre better. The stimuli ordering was counter balanced to prevent the bias.

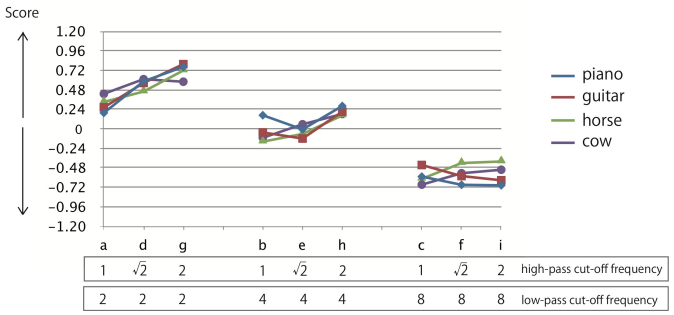

Fig. 11. Results of the timbre identity preservation test

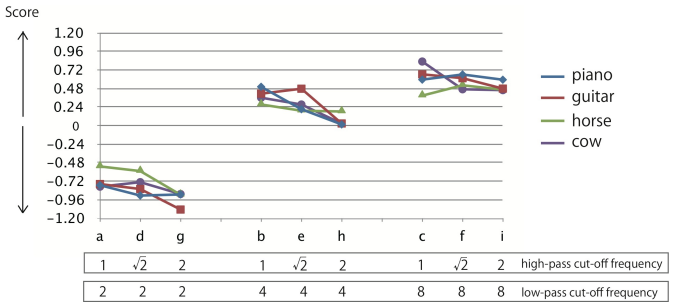

Fig. 12. Results of the linguistic information preservation test

E. test results

The test results were summarized using Thurstone’s case V procedure to yield scores on an interval scale. Figure 11 shows the results of the timbre identity preservation test. The vertical axes represents the score of identity preservation on an interval scale. The higher score indicates better preservation. The results shows that higher high-pass cut-off preserves timbre better and the lower low-pass cut-off preserves timbre better. It is interesting to observe that the scores seem to be irrespective to the original sounds.

Figure 12 shows the results of the linguistic information preservation test. The vertical axes represents the score of linguistic information preservation on an interval scale. The higher score indicates better preservation. The results shows that lower high-pass cut-off preserves linguistic information better and the higher low-pass cut-off preserves linguistic information better. It is interesting to observe that the scores seem to be irrespective to the original sounds also in this experiment.

Figure 13 shows the summary of two experiments. The line annotated “linguistic” is the mean score of the linguistic preservation test. The line annotated “instruments” is the mean score of the identity preservation test. The results suggest that there is a trade-off relation between linguistic information

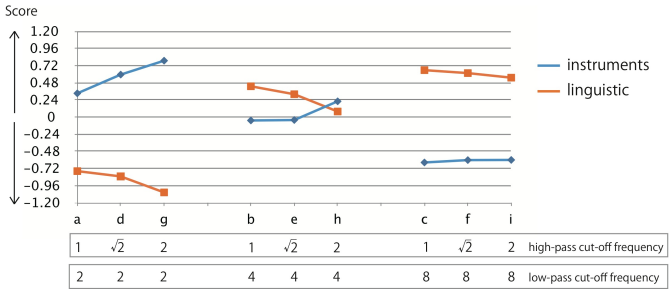

Fig. 13. Summary of the test results of the two experiments

and timbre identity. However, it should be noted that in the intermediate conditions, such as b, e and h, the both information were reasonably preserved perceptually, since the primary timbre modifying component, global spectral shape, is already removed in these conditions.

These results suggest that the proposed method provides means to control the linguistic contents and the timbre identity of the cross synthesized sounds. It opens interesting possibilities for sound engineers and designers.

VII. Conclusion

In this paper, we proposed a new cross-synthesis VOCODER, which enables control of the linguistic information and the timbre identity in the synthesized sounds. The proposed method is based on a F0-adaptive spectral envelope estimation (TANDEM-STRAIGHT), which provides close-to-natural synthetic speech in a VOCODER framework. In the whitening stage, the temporally averaged and smoothed logarithmic spectrum is removed for eliminating the primary timbre modifying component, global spectra shape. The flexible control of the linguistic information and the timbre identity is implemented by designing cut-offs of band-pass filtering of the whitened utterance spectrogram in the modulation frequency domain. This flexibility enables customization of the proposed method to various applications.

Acknowledgment

The authors thank our colleague Kokeguchi for providing the Thurstone's case V program. The authors also appreciate anonymous reviewers' comments for improving readability of this article and fixing mistakes.

References

[1] C. Roads: The Computer Music Tutorial, The MIT Press, 2002.
[2] P. Lanchantin, S. Farner, C. Veaux and G. Degottex, N. Obin, G. Beller, F. Villavicencio, S. Huber, G. Peeters, A. Roebel and X. Rodet: VIVOS VOCO: A survey of recent research on voice transformations at IRCAM, Proc. DAFx-11, pp.19–23, 2011.
[3] H. Dudley: Remaking speech, J.Acoust.Soc.Am., vol.11, no.2, pp.169–177, 1939.
[4] J. L. Flanagan: Phase vocoder, the Bell System Technical Journal, pp.1493–1509, 1966.
[5] F. Itakura and S. Saito: A statistical method for estimation of speech spectral density and formant frequencies, Trans. IEICEJ, Vol.53-A, No.1, pp.35–42, 1970. [in Japanese]
[6] B. S. Atal, S. L. Hanauer: Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, J. Acoust. Soc. Am., vol.50, 2B, 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, Proc. ICSLP-94, Vol.3, No.4, pp.1043–1046, 1994.
[8] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," Proc. ICASSP 2008. Las Vegas., pp.3933–3936, 2008.
[9] H. Kawahara, I. Masuda, 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.
[10] M. Morise, T. Takahashi, H. Kawahara, T. Irino, "Speech Analysis Using Temporally Stable Power Spectrum Estimation Method for Periodic Signals," Institute of Electronics, Information, and Communication Engineers, vol.J 92-A, No.3, pp.163–171, 2009.
[11] M. Unser, "Sampling–50 years after Shannon," Proceedings of the IEEE, vol.88, no.4, pp.569–587, 2000.
[12] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," Sadhana, Vol. 36, Part 5, October 2011, pp. 713–727
[13] H. Akagiri, M. Morise, T. Irino, H. Kawahara, "Evaluation and Optimization of F0-Adaptive Spectral Envelope Extraction Based on Spectral Smoothing with Peak Emphasis," A - Abstracts of IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, Vol.J94-A No.8, pp.557–567 (In Japanese)
[14] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," IEEE Trans. Audio and Electroacoustics, vol. AU-15, no. 2, pp. 70–73, 1967.
[15] B. R. Glassberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," Hearing Research, vol.47(1–2), pp.103–138, 1990.
[16] R. Drullman ,J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am., vol.95, no.2, pp.1053–1064, Feb,1994.
[17] R. Drullman ,J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am., vol.95, no.5, pp.2670–2680, May,1994.
[18] N. Kanedera,T. Arai, T. Funada, "Robust Automatic Speech Recognition Emphasizing Important Modulation Spectrum," The Institute of Electronics, Information and Communication Engineers(D-II), Vol.J84-D-2, No.7, pp.1261–1269, 2011.
[19] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am., vol.130, Issue 3, PP.1475–1487, 2011.
[20] A. V. Oppenheim and R. W. Shafer, "Discrete-Time Signal Processing," Prentice-Hall, Englewood Cliffs, NJ, 1987.
[21] H. Kawahara, Y. Wada, T. Nishi, M. Morise, R. Nisimura, T. Irino, "Time varying filter implementation and excitation source representation for speech analysis, modification and synthesis systems," Technical Committee of Psychological and Physiological Acoustics the Acoustical Society of Japan, Vol.41, No.7, pp.561–566, 2011. [in Japanese]
[22] 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.
[23] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Database of Copyright-cleared Musical Pieces and Instrument Sounds for Research Purposes," Journal of Information Processing Society of Japan, Vol.45, No.3, pp.728–738, March 2004.