SPEECH SYNTHESIS USING SUBBAND-CODED MULTIBAND SOURCE COMPONENTS AND SINUSOIDS

Nobuyuki Nishizawa and Tsuneo Kato

KDDI R&D Laboratories Inc., Japan
{no-nishizawa, tkato}@kddilabs.jp

ABSTRACT

An improved speech waveform generation method for speech synthesizers using filter banks is proposed where spectral features of synthetic sounds are constructed by amplitude modification and summation of predecomposed source waveforms. In the method, since all operations are performed in the subband-coded domain with a reduced sampling rate, the computational cost can also be reduced. Moreover, to improve the accuracy of spectral reproduction in low frequency domain of voiced sounds, sinusoidal synthesis directly performed on low subbands is also introduced. The result of a subjective test using resynthesized sounds spoken by a male and female narrator indicated that the proposed method was significantly superior to the conventional methods using a mel log spectrum approximation (MLSA) filter and non-maximally decimated filter bank, which was our previously proposed method.

Index Terms— HMM-based speech synthesis, speech waveform generation, filter bank, subband coding, embedded systems

1. INTRODUCTION

HMM-based speech synthesis [1, 2] is suitable even for embedded or mobile systems because it can generate high-quality sounds with only several hundred kilobyte or several megabyte data. However, the computational cost for HMM-based speech synthesizers is still a problem for such systems.

As the filter operation in waveform generation constitutes a large part of the entire processing cost, the focus should be on reducing the calculation cost for the part. In this regard, a filter bank-based speech synthesis method based on a pseudo quadrature mirror filter (QMF) bank [3, 4] was proposed in our previous study [5], where operation with $O(\log N)$ per sample can be achieved by using $O(N \log N)$ fast algorithms such as Chen's discrete cosine transformation (DCT) [6]. In the method, spectral features of speech sounds were constructed by modification of amplitude for each subband the sampling rate of which was reduced by decimation. However, to cancel aliases caused by the modification, maximally-decimated filter banks, which are commonly used for audio coding like MPEG Audio [7], was not applicable to the method. Furthermore, constraints on the amplitude modification between neighboring subbands were also required to cancel aliases.

By contrast, in this study, source waveform decomposition to construct spectral features is separated from the pseudo QMF banks for subband coding. In the proposed method, a speech synthesizer using a filter bank is implemented on a subband coding system based on the maximally decimated pseudo QMF banks. Since amplitude modification to construct spectral features is performed by scaling coded vectors of decomposed source waveforms, the method is free from aliases. Thus, design of the filter bank to construct spectral features of sounds is independent of that of the subband coding system to reduce the sampling rate. Moreover, to improve spectral reproduction, sinusoidal synthesis [8] performed in the subband domain is also adopted to construct mainly low-frequency subband components of voiced sounds.

2. FILTER BANK-BASED SPEECH SYNTHESIS

In this section, speech synthesis using filter banks without sample decimation is introduced. The proposed method is a similar method implemented on a subband coding system.

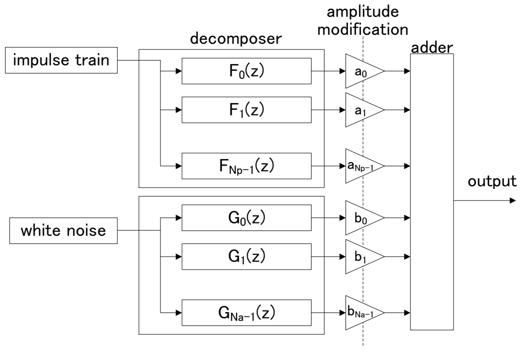
2.1. Speech synthesis using a filter bank

Figure 1 shows a block diagram of the filter bank-based speech synthesizer. In this system, impulse trains and white noise sequences as source waveforms are initially band-decomposed by filter banks. Then, spectral features of synthetic speech sounds are constructed from the band-decomposed waveforms with amplitude modification. This amplitude modification should be controlled with appropriate delays to compensate delays in the filter bank.

For simplicity of the processing, it is desirable that simple summation of the decomposed waveforms without amplitude modification restores the original source waveform. Therefore, cosine modulated filters of the Nth-band filter [9] are used for the band-decomposition. The N-th band filter is a linear-phase low-pass filter where the edge of the stopband is $1/2N$ in normalized frequency, and the magnitude responses at 0 and $1/4N$ in normalized frequency are approximately 1 and $1/2$, respectively. In this study, N equals 32. This base filter in modulation is called the prototype filter. Figure 2 shows the impulse response and magnitude response of the prototype filter used in this study.

On the other hand, cosine modulation corresponds to shifting of the magnitude response along the frequency axis. The cosine modulation of the filter coefficients for the n-th filter $f_n(i)$ for band-decomposition is performed by the fol-

978-1-4799-0356-6/13/$31.00 ©2013 IEEE



$F_n(z)$: n–th band–decomposition filter for voiced components
$G_{n'}(z)$: n' –th band–decomposition filter for unvoiced components
a_n, $b_{n'}$: amplitude modification factors for at the n–th and n'–th band
N_p, N_a: the numbers of bands for voiced and unvoiced sounds

Fig. 1. Block diagram of the filter bank based speech synthesizer.

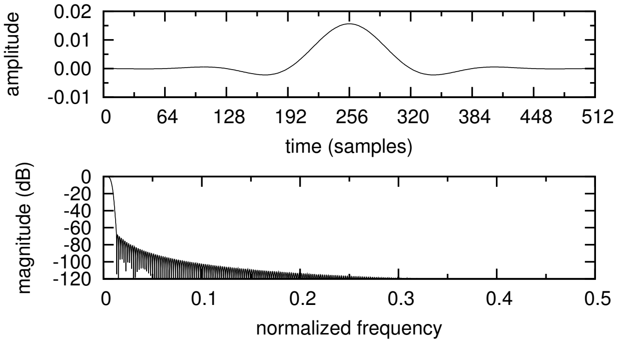


Fig. 2. Impulse response and magnitude-frequency response of a 32nd-band filter as the prototype filter for the decomposer.

lowing equation:

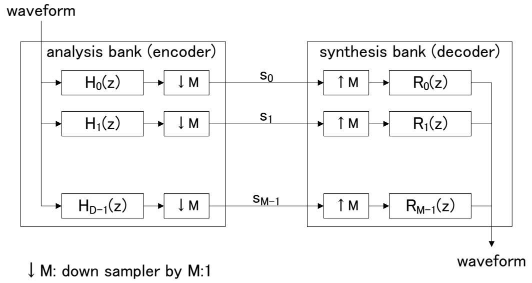
$$f_n(i) = 2f(i)\cos\Big(\frac{\pi}{2N}(2n+1)i\Big)\ (0 \le n \le N-1) \quad (1)$$

where $f(i)$ denotes the impulse response of the prototype filter. In this configuration, the edges of the n-th passband are $(2n-1)/4N$ and $(2n+1)/4N$ in normalized frequency.

In this study, the decomposition banks for impulse trains and white noise sequences are the same (i.e., $N_p = N_a$ and $F_n(z) = G_n(z)$ in Fig. 1). The amplitude modification factors are given by the magnitudes of the target spectra at the central frequencies of the bands, similar to our previous study [5].

2.2. Sinusoidal synthesis for low bands of voiced sounds

In our previous study, insufficiency of the resolution along the frequency axis caused degradation of the quality of the synthetic sounds. Although increasing the number of bands of the decomposition improves the accuracy of spectral feature reproduction, it makes the length of the filters longer. Therefore, speech synthesis by sinusoids [8] is also introduced to improve the accuracy of the spectral feature reproduction in the proposed method. Since intervals of harmonic components are usually narrower than those of the bands, errors in the spectral feature reproduction can be reduced.



↓M: down sampler by M:1
↑M: up sampler by 1:M
s_m: subband samples at the m–th subband

Fig. 3. Block diagram of the subband coding system.

In this study, for the sake of simplicity, a formulation is introduced where waveforms are built by summation of cosine functions rather than sine functions:

$$x(t) = \sum_k A_p(\omega_0)\cos\big(\omega_0 k(t-t_0)\big) \quad (2)$$

where ω_0, $A_p(\omega_0)$ and t_0 are angular frequency of fundamental vibration, amplitude of the target sound at ω_0 and the position of the corresponding impulse, respectively.

However, the sinusoidal synthesis limited for the low-frequency domain for voiced sounds may be preferable because the cost for the sinusoidal synthesis increases due to increase of the number of sinusoids especially for low fundamental frequency sounds. Actually, steep magnitude characteristics in the frequency domain are observed mainly in the low-frequency domain of voiced sounds. In such hybrid systems, the phases of the sinusoids should be controlled with synchronization of those of the impulse trains for the excitation source.

3. SUBBAND CODING FOR REDUCTION OF SAMPLING RATE

Figure 3 shows a block diagram of a subband coding system with maximally decimated pseudo QMF banks where the number of subbands is M. In the system, input waveforms are equally decomposed into subbands in the analysis bank, and then the decomposed waveforms are composed in the synthesis bank. This coding system can be easily integrated into the speech synthesis system described in Section 2, where only a synthesis bank is necessary since the subband coding system consists only of linear processing. Figure 4 shows the structure of the proposed speech synthesizer schematically. Similar to our previous study, coded decomposed components of the source waveform are pre-stored. Thus, no decomposition or coding is performed on the speech synthesizer.

3.1. Predecomposition and pre-encoding of source waveforms

Commonly, analysis and synthesis filters of pseudo QMF banks are also made by cosine modulation of the impulse response of a prototype filter. In this study, the impulse responses of the filters of the analysis and synthesis banks are given by:

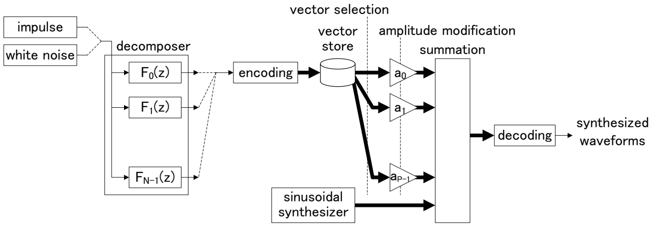


Fig. 4. Block diagram of the proposed system. Bold lines correspond to coded vectors.

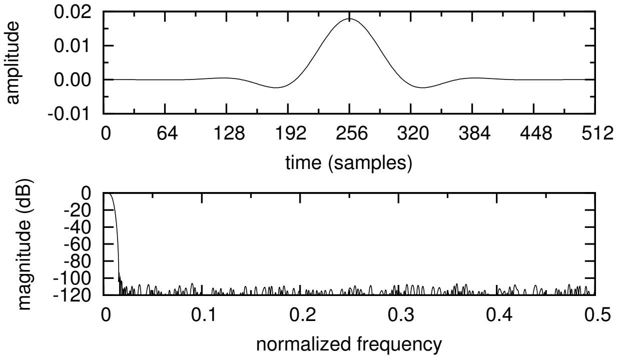


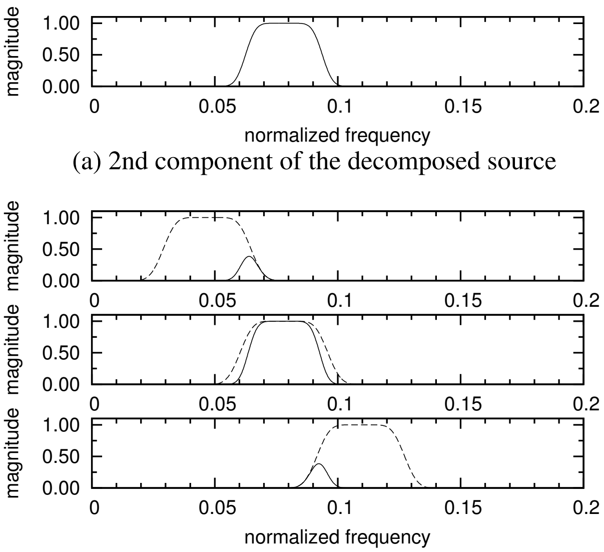
Fig. 5. Impulse response and magnitude-frequency response of the prototype filter for a 32-band pseudo QMF bank.

$$h_m(i) \quad = \quad 2h(i)\cos\Big(\frac{\pi}{64}(2m+1)(i-16)\Big) \qquad (3)$$

$$r_m(i) \quad = \quad 2h(i)\cos\Big(\frac{\pi}{64}(2m+1)(i+16)\Big) \qquad (4)$$

where $h(i)$ is the impulse response of the prototype filter. These equations are from the MPEG Audio specification [7]. In this study, the prototype filter defined in the MPEG Audio specification, which is for 32-band filter banks, is also used. The length of the filters is 512. Thus, filter banks that are components of highly optimized MPEG Audio decoders like [10] are directly applicable to systems for the proposed method. Figure 5 show plots of coefficients and magnitude-frequency response of the prototype filter, respectively.

With this formulation, the coded vectors of band-decomposed source waveforms have many zero components. Figure 6 explains the reason schematically. The figure shows (a) magnitude-frequency characteristic of the 2nd component of decomposed white source waveform, and (b) responses of the 1st, 2nd and 3rd filters in the analysis bank of the subband coding system. The figure (b) indicates only the 1st, 2nd, 3rd elements of the coded vector have non-zero values. The other elements can be zero since the component is cut by the analysis filters. In general, where the number of subbands equals the number of bands in the band-decomposition for the source waveforms (i.e., $M = N$), only three components of the coded vector of the decomposed source waveforms have non-zero values. It reduces not only the size of pre-stored source waveform components but also the computational cost for the amplitude modification. Consequently, the computational complexity of the system can be still $O(\log N)$ per sample, which is similar to the proposed method in our previous study.



(a) 2nd component of the decomposed source

(b) 1st, 2nd and 3rd analysis filter responses for (a)

Fig. 6. Magnitude-frequency characteristic of the 2nd component of the decomposed white source, and those filtered by the 1st, 2nd and 3rd filters of the analysis bank in the subband coding system. Hatched lines correspond to responses of the analysis filters solely.

In this study, a periodic pseudo noise sequence where the period is 4096 samples is used as the noise source. On the other hand, impulse trains are generated by summation of impulse waveforms. In speech synthesis, predecomposed and pre-encoded vector sequences are used, similar to our previous study.

3.2. Sinusoidal synthesis in subband domain

Since magnitude-frequency and phase-frequency responses of the filters of the analysis bank are given, encoded results of sinusoids can be easily obtained by calculations in frequency domain. Where $|H_m(\omega)|$ and $\arg H_m(\omega)$ are magnitude-frequency and phase-frequency responses of the analysis filter for the m-th subband, respectively, m-th element of the subband vector of encoded cosine waveform with angular frequency ω is given given as follow:

$$x_{\omega,m}(t) = |H_m(\omega)|A_p(\omega)\cos\big(\omega(t-t_0)+\arg H_m(\omega)\big) \quad (5)$$

where $A_p(\omega)$ denotes the magnitude of the spectrum at ω.

Referring to Eq. (3), $\arg H_m(\omega) = -\pi(2m+1)/4$. On the other hand, $|H_m(\omega)|$ can be easily obtained with a pre-calculated table. For example, a 4096-entry table for between 0 and $\pi/32$ in angular frequency was used with shift and reverse operations along the frequency axis and linear interpolation to synthesize sounds for the following evaluation.

Basically, one sinusoid encodes into two subbands due to the overlap structure of the analysis bank. By these operations, the sinusoidal synthesis can be performed in the subband domain with the reduced sampling rate.

4. SUBJECTIVE EVALUATION

To examine the practicality in quality of sounds by the proposed method, a mean opinion score (MOS) test of the synthetic sounds was conducted. Similar to the experiment in our previous study, use of the proposed method in the conventional HMM-based speech synthesizers using melcepstrum was intended in this study. Therefore, the original spectral features were modeled using melcepstrum. Source components of voiced and unvoiced sounds consisted only of impulse trains (and sinusoids in the proposed method) and noise sequences, respectively. It also corresponds to the source model of the target speech synthesizer. However, to simulate an ideal condition in speech modeling, the synthesis target parameters were extracted from natural speech sounds.

In the test, 10 subjects listened to synthetic speech sounds of a male and female narrator and scored them on a 5-point discrete scale (1: very poor, 2: poor, 3: fair, 4: good, 5: very good) to express their preferences. The speech analysis method was basically similar to that of the listening test of the previous study with Speech Processing Toolkit (SPTK) version 3.5 [11]. In this test, the sampling rate was 16 kHz. In speech analysis, Blackman window was applied first, then 39th-order melcepstral coefficients where frequency warping coefficient $\alpha = 0.42$ were extracted with the *mcep* tool in the SPTK. The fundamental frequency (F_0) was also extracted by the *pitch* tool in the SPTK with the "sawtooth waveform inspired pitch estimator" (SWIPE) [12] algorithm. Different from the previous study, the frame period was always fixed to 2 ms (32 samples) because the number of subbands was fixed to 32. Subband amplitude modification factors were determined from the extracted melcepstrum through power spectra in mel-scale with linear interpolation, where the number of the dimensions for mel-spectrum was 64 for all conditions.

For comparison, speech sounds synthesized using a mel log spectrum approximation (MLSA) filter [13] and our previous filter bank-based method with non-maximal decimation were also prepared. As the MLSA filter and excitation source for speech synthesis, the SPTK were also used. On the other hand, in the previous filter bank-based method, the number of subband was 128, which is the best condition of the listening test of the previous study. Only in this case, the frame period is 4 ms (64 samples), which is half the number of subbands.

Consequently, there were 8 conditions per narrator. For each condition, stimuli for 10 sentences that were similar to those from J01 to J10 of the ATR503 corpus [14] were prepared. These sentences were also the same as those used in the listening test of the previous study. Thus, 160 stimuli in total were presented to each subject. The stimuli were randomly ordered for each subject and presented to both ears through closed-ear headphones in a silent room.

Figure 7 shows the results of the test. The scores that were less than 4 for all conditions might be caused by the down-sampling to 16 kHz and voiced sounds excited only by impulse trains. Lack of noise components in voiced sounds might degrade MOS scores especially in female sounds. In comparison to the conventional methods, the proposed method with 4 or more sinusoidal subbands is significantly superior to the conventional method. Although it has been reported spectral errors of the MLSA filter are small enough for practical use, the filter, which is an infinite impulse response system, can be unstable temporarily in time-varying systems in practice. This is because no time-varying characteristic of MLSA filter parameters is taken into account regarding the stability of the system. This time-unvarying assumption can lead to the generation of noise-like sounds in the actual systems. On the other hand, the proposed method, which consists only of finite impulse response systems, is always stable theoretically. This might be the reason for the difference in the scores.

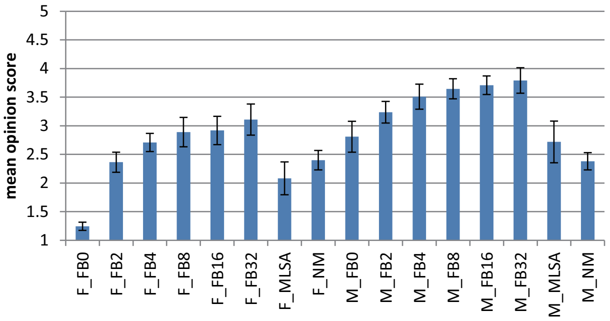


Fig. 7. Result of the subjective evaluation. In the conditions, prefix F and M correspond to the female and male narrator, and suffix FBx, MLSA and NM correspond to the proposed filter bank-based methods where x is the number of low subbands synthesized by the sinusoidal synthesis, the MLSA filter-based method and the non-maximally decimated filter bank-based method, respectively. The error bars indicate the 95% confidence intervals.

Although increasing the number of subbands synthesized by sinusoids improved the MOS scores, no significant difference between 4 and 32 in the number was observed in both male and female sounds. In practice, the optimal condition depends on not only the MOS scores but also the computational costs. However, accurate and practical cost evaluation especially for comparison among coded sinusoidal waveform generation, which depends on the range of the fundamental frequency of the synthetic sounds, amplitude modification for each component and decoding of subband-coded vectors requires highly optimized implementations. A cost evaluation under practical conditions will be performed in future work.

5. CONCLUSION

This paper presented an improved filter bank-based speech synthesis on subband coding. To improve the accuracy of spectral restoration for voiced sounds, sinusoidal synthesis on the subband coding was also introduced. In the system, decomposition of source waveforms and subband coding are completely separated. The results of a subjective test using resynthesized sounds indicated that the method was superior to conventional methods using an MLSA filter and non-maximally decimated filter bank in terms of the quality of sounds.

In future work, we will examine the performance of the proposed method under practical conditions in terms of computational cost.

6. REFERENCES

[1] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S., "Speech synthesis using HMMs with dynamic features," in Proc. of ICASSP '96, vol. 1, pp. 389–392, May 1996.

[2] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in Proc. Eurospeech '99, pp. 2347–2350, Sep. 1999.

[3] Rothweiler, J. H., "Polyphase quadrature filters – A new subband coding technique," in Proc. ICASSP '83, Boston, MA, U.S.A., vol. 3, pp. 1280–1283, Apr., 1983.

[4] Princen, J. P., Johnson, A. W. and Bradley, A. B., "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in Proc. ICASSP '87, Dallas, TX, U.S.A., vol. 4, pp. 2161–2164, Apr. 1987.

[5] Nishizawa, N. and Kato, T., "Speech synthesis using a non-maximally decimated filter bank for embedded systems," in Proc. INTERSPEECH 2012, Portland, OR, U.S.A., Wed.O6d.04, Sep. 2012.

[6] Chen, W. H., Smith, C. H, and Fralick, S. C, "A fast computational algorithm for the discrete cosine transform," IEEE Trans. on Comm., vol. 25(9), pp. 1004–1009, Sep. 1977.

[7] ISO/IEC, JTC1/SC29/WG11 MPEG, "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s Part 3: Audio," IS11172-3, 1992.

[8] Quatieri, T. F. and McAulay, R. J., "Speech transformations based on a sinusoidal representation," IEEE Trans. on ASSP, vol. 34(6), pp. 1449–1464, Dec. 1986.

[9] Mintzer, F, "On half-band, third-band, and Nth-band FIR filters and their design," IEEE Trans. on ASSP. vol. 30(5), pp. 734–738, Oct. 1982.

[10] Hans, M. C. and Bhaskaran, V., "A fast integer implementation of MPEG-I audio decoder," HP Labs Technical Reports, HPL-96-03, Jan. 1996.

[11] Tokuda, K., Oura, K., Tamamori, A., Sako, S., Zen, H., Nose, T., Takahashi, T., Yamagishi, J. and Nankaku Y., "Speech Signal Processing Toolkit (SPTK)," http://sptk.sourceforge.net/.

[12] Camacho, A., "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music," Ph.D. Thesis, University of Florida, 116p., 2007.

[13] Imai, S., "Cepstral analysis synthesis on the mel frequency scale," in Proc. ICASSP '83, vol. 8, pp. 93–96, Apr. 1983.

[14] Abe, M., Sagisaka, Y., Umeda, T. and Kuwabara, H., "Speech Database User's Manual," ATR Interpreting Telephony Research Laboratories Technical Report, TR-I-0166, Japan, Aug. 1990 (in Japanese).