ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speech Synthesis Using a Non-maximally Decimated Filter Bank for Embedded Systems

Nobuyuki Nishizawa, Tsuneo Kato

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

Abstract

A novel speech waveform generation method using a non-maximally decimated filter bank is proposed, where spectral features of synthetic sounds are created by amplitude modification of subband samples that are pre-decomposed from impulse or noise waveforms. The proposed method uses two synthesis banks of the maximally decimated pseudo quadrature mirror filter (QMF) bank structure which is similar to that in the MPEG audio decoder. Consequently, the computational complexity of the proposed method is $O(logN)$ per sample, while that of the conventional method based on the source-filter model with an auto-regressive (AR) filter or mel log spectrum approximation (MLSA) filter is $O(N)$ per sample. A MOS test for resynthesized speech sounds from the results of analyzing natural speech sounds showed the proposed method achieved scores similar to those of the conventional method using the MLSA filter for a female narrator.

Index Terms: HMM-based speech synthesis, pseudo quadrature mirror filter bank, non-maximally decimation, embedded systems

10.21437/Interspeech.2012-366

1. Introduction

High-quality sounds by speech synthesis have been required even for mobile or embedded systems, while a small footprint is simultaneously a requirement. HMM-based speech synthesis[1][2] is suitable for such systems since it can generate high-quality sounds with only several hundred kilobyte or several megabyte data.

However, the computational complexity of waveform generation by HMM-based speech synthesizers can be problematic, especially for such systems. In waveform generation, filtering to create spectral features constitutes a large part of the entire processing cost. Although the mel log spectrum approximation (MLSA) filter[3] is often adopted for speech synthesis systems using mel-cepstra for efficient processing, the computational complexity of the MLSA filtering is still high for embedded systems. Since rational expressions represented by fourth or fifth-order Padé approximations in z space are generally introduced for MLSA filters and filtering with mel frequency warping doubles the count of the product-sum operations, the count of product-sum operations per sample of the MLSA filters roughly exceeds eight or ten times the dimensions of the mel-cepstrum, while its computational complexity is $O(N)$ per sample where N corresponds to the dimensions of the mel-cepstrum.

Although computational cost for speech synthesis using linear spectral pairs (LSP)[4][5] can be smaller than that using an MLSA filter, the computational complexity is still $O(N)$ per sample. By contrast, an $O(\log N)$ speech generation method based on a non-maximally decimated filter bank is proposed in this study. In the proposed method, spectral features of synthetic speech sounds are created by amplitude modification of subband samples from source waveforms. The filter bank of the proposed method is based on a pseudo quadrature mirror filter (QMF)[6] structure, which is similar to the filter banks used in MPEG audio[7]. Since this non-maximally decimated bank can be constructed with two maximally decimated pseudo QMF banks, a fast algorithm based on discrete cosine transform (DCT) of $O(N \log N)$ per operation for the pseudo QMF banks of MPEG audio[8] is directly applicable to the filter banks used in the proposed method. When such algorithms are employed, the computational complexity per sample can be reduced to $O(\log N)$ by decimation. Therefore, the proposed method may be more efficient than systems using MLSA filters or LSP. Moreover, excitation source waveforms can be generated by combination of pre-composited impulse and noise waveforms in the subband space. Thus, only composition of sub-band waveforms is performed in the speech synthesizers.

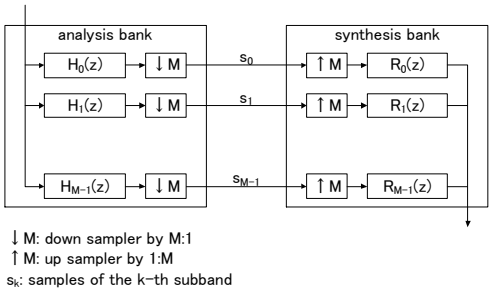


Figure 1: *Structure of maximally decimated filter bank.*

The goal of this study is to develop a new waveform generation method as a substitute for the MLSA filters in HMM-based speech synthesizers for embedded systems. Therefore, this study also explains that mel-cepstral coefficients as input can be converted into amplitude modification factors for subbands by $O(N \log N)$ algorithms. Since this conversion is performed for each decimated sample vector in the subband domain, $O(\log N)$ waveform generation can be achieved.

2. Pseudo QMF banks

Figure 1 shows a block diagram of the maximally decimated filter banks where the number of subbands is M. Input waveforms are equally decomposed into subbands in the analysis bank, then the decomposed waveforms are composed in the synthesis bank. In maximally decimated banks, down-sampling and up-sampling by M are performed after and before filtering, respectively.

Two methods are now often used for subband coding with maximally decimation[9]. One is pseudo Quadrature Mirror Filter (QMF) bank[6], and the other is modified discrete cosine transformation (MDCT)[10]. Generally speaking, the former can achieve higher stopband attenuation in the analysis and synthesis filter banks, while strictly perfect reconstruction of input signals is difficult in practice. In this study, the pseudo QMF

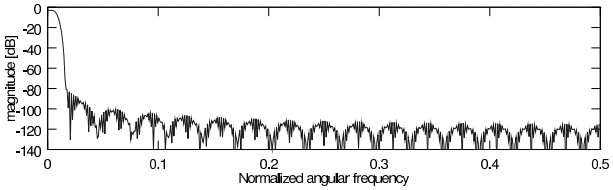


Figure 2: *Magnitude response of the prototype filter for the QMF bank* ($M = 32$).

bank is adopted since higher stopband attenuation of the filters is required to cut alias components caused by down-sampling or up-sampling.

In the QMF bank, the analysis and synthesis filters are made by cosine modulation of the impulse response of a prototype filter, where the magnitude response of each filter is shifted along the frequency axis. Where $h(n)$ is the impulse response of the prototype filter, the impulse response of the filters of the analysis and synthesis banks are given by:

$$h_k(i) = 2h(i)\cos[(k+\frac{1}{2})\frac{i}{M}\pi - \theta_k] \quad (1)$$

$$r_k(i) = 2h(i)\cos[(k+\frac{1}{2})\frac{i}{M}\pi + \theta_k] \quad (2)$$

where θ_k satisfies the following equation:

$$\theta_{k+1} = \theta_k \pm \frac{\pi}{2} \quad (3)$$

In this study, θ_k is defined as:

$$\theta_k = \frac{2k+1}{4}\pi \quad (4)$$

and $(16 \times M)$-tap finite impulse response (FIR) filters are used for the filtering. This design is similar to the pseudo QMF banks used for the MPEG audio standard.

The prototype filter should be designed such that the squared magnitude response becomes a M-th band filter[11]. In this study, the filter was designed by the algorithm described in [12]. The edge of the stopband was set to π/M. In this case, the stopband attenuation is over -70dB for all filters. Figure 2 shows the magnitude response of the 512-tap prototype filter where $M = 32$. In the pseudo QMF banks, although aliasing of neighboring subbands cancel each other out, other alias components are directly cut by the filters. Therefore, higher stopband attenuation is desired.

3. Non-maximally decimated filter bank for speech synthesis

When filter banks can create spectral features of speech sounds, input waveforms of the speech synthesizer using the filter banks can consist only of white source waveforms like impulse trains or white noise waveforms. Although such speech synthesis can be easily achieved using systems without decimation, they have no advantage in terms of computational cost. This is because the computational complexity of filter banks is $O(M \log M)$ per operation even when a fast algorithm like fast Fourier Transformation (FFT) is used, while those for the AR or MLSA-based speech synthesizers are $O(M)$ per sample. However, systems based on filter banks with decimation can reduce complexity to $O(\log M)$ per sample. Thus, decimated filter banks can make speech synthesis faster, provided that aliases caused by down-sampling and up-sampling are appropriately taken into account.

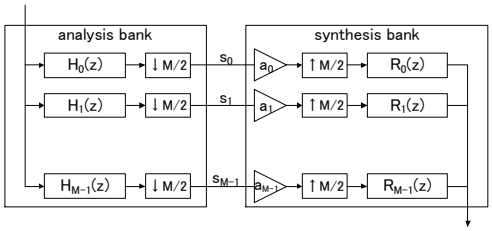


Figure 3: *Block diagram of the non-maximally decimated QMF bank.*

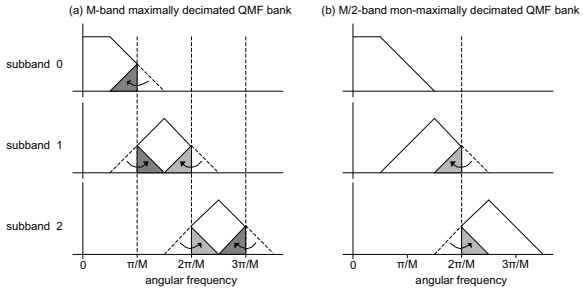


Figure 4: *Aliases in subband decomposition using the maximally and non-maximally decimated QMF banks.*

For audio coding, maximally decimated filter banks are generally used to reduce the bitrate as much as possible. However, since alias cancellation in the pseudo QMF banks is performed only where subband vectors are kept between decomposition and composition, the maximally decimated QMF banks cannot create spectral features without aliasing. Therefore, to relax this restriction, the decimation factor is reduced to $M/2$ in this study. Such filter banks are also called oversampled filter banks[13]. Figure 3 show the block diagram of the QMF bank used for the proposed speech synthesizer.

Figure 4 schematically shows aliases in the maximally and non-maximally decimated QMF banks, respectively. Referring the left part of the figure, in the maximally decimated banks, after filtering by $H_k(z)$ of the analysis bank, M:1 down-sampling causes aliasing at $k\pi/M$ $(1 \leq k \leq M-1)$ in the angular frequency range. In this case, the aliases are canceled by filtering of $R_k(z)(= \sum_{i=0}^{L-1} r_k(i)z^{-i})$ after up-sampling in the synthesis bank. Since the alias cancellation of the QMF banks is performed between neighboring subbands, $a_k = a_{k+1}$ $(0 \leq k \leq M-2)$ where a_k denotes amplitude modification factor for subband k are required for the alias cancellation. In other words, all a_k should be equal. On the other hand, in the $M/2$ decimated filter banks, aliasing is caused only at $2l\pi/M$ $(1 \leq l \leq M/2-1)$ after down-sampling. Therefore, although $a_{2l+1} = a_{2l+2}$ $(0 \leq l \leq M/2-2)$ are still required, there is no restriction between a_{2l} and a_{2l+1}, i.e., the magnitude of the spectrum is independently configurable at $(M/2+1)$ points. In this study, amplitude modification factors a_k are extracted from mel-cepstral coefficients, and each pair of a_{2l-1} and a_{2l} is controlled by the magnitude at $2l\pi/M$.

Obviously, this non-maximal decimation doubles the computational cost. Nevertheless, since it is still $O(logM)$ per sample, it is still acceptable unless the number of subbands is far more than that of the dimensions of the conventional spectral feature vector. In practice, a filter bank with $M/2$:1 decimation can be configured with two sets of the maximally decimated filter banks. It is schematically explained in Figure 5, i.e., the sum of outputs of two maximally decimated QMF banks to which subband vectors are alternatively issued equals the output of the $M/2$:1-decimated QMF bank since $R_k(z)$ are FIR systems. In our design, the structure of the QMF banks is similar to that

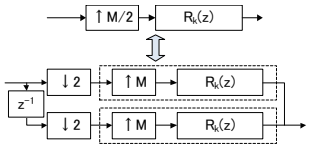


Figure 5: *Conversion of a component for a subband of a non-maximally decimated filter bank.*

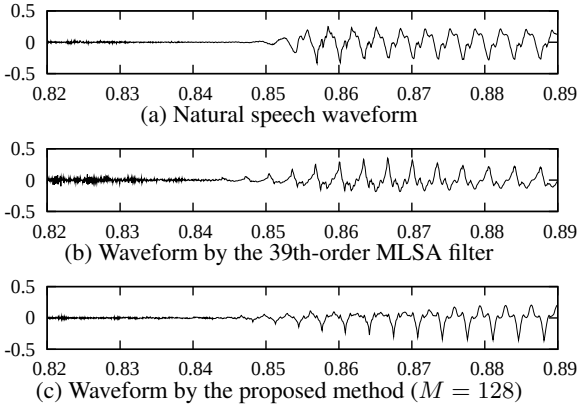


Figure 6: *Waveforms of* /chiisana/ *by a female narrator.*

of the MPEG audio decoder. As mentioned previously, several fast DCT-based algorithms of $O(logN)$ per sample for the pseudo QMF banks have already been proposed[8]. These techniques, with the exception of excessively optimized techniques where M equals 32 (which is the number of subbands used in the MPEG audio standard), are directly applicable to the proposed method.

Although the AR or MLSA filters are minimum phase systems, the proposed synthesizer is a linear phase system. Figure 6 shows examples of three types of waveforms. The figure shows that the shapes of the waveforms are quite different. Since the quality of the sounds might be also quite different from conventional speech synthesis methods, they should be subjectively examined for practical use. This will be presented in Section 5.

4. Design of the proposed speech synthesizer to reduce computational cost

4.1. Predecomposition of source waveforms

When all spectral features are generated by a filter bank, input waveforms of the filter bank can be limited to white source waveforms like white Gaussian noise or impulse train. Moreover, since the filter bank in the proposed method is an FIR system, combination of source components can also be easily performed on the subband domain. Therefore, only predecomposed waveforms for such sources are used in the proposed method. Consequently, the proposed speech synthesizer requires no analysis bank.

In this study, a short-period cyclic pseudo noise sequence is used as the noise source. Therefore, pre-decomposition for a noise source waveform is quite simple. In this case, a sequence with $T_n/(M/2)$ subband sample vectors is required for the pre-decomposed waveform table, where T_n denotes the period of the pseudo noise waveform. In other words, the required table size for the noise equals $2T_n$. Of course, T_n should be configured to be long enough so that periodicity can be ignored. In this study, T_n is set to 4096 (0.256 s at 16 kHz sampling).

On the other hand, in this study, impulse trains are also generated from pre-decomposed waveforms. However, to reduce the size of the source waveform table, impulse trains are built by adding pre-decomposed impulses. This makes operations slightly complex because the impulse response of the filter of the QMF bank should be taken into account. Consequently, $(2L/M + 1)$ vectors are required to express an impulse waveform where L denotes the length of the filter. In other words, an impulse waveform is modeled with a subband vector sequence that consists of $(2L/M + 1)$ subband vectors. Moreover, for phase shift, $M/2$ kinds of impulses are necessary when the temporal precision of each impulse equals the sampling rate. In such a case, the required size of the table for impulses is $(ML + M^2/2)$.

In addition, power normalization is also required for impulse trains. Components of pre-decomposed impulse waveforms should be multiplied by $(T_0/T_s)^{1/2}$ where T_0 and T_s are the fundamental period and sampling period, respectively.

4.2. Calculation algorithm for amplitude modification factors from the mel-cepstrum

Since the feature vector of the target system is the mel-cepstrum, the amplitude modification factors for subbands are extracted from mel-cepstra. To reduce error in this conversion, more dimensions in the intermediate vectors of the conversion than the order of mel-cepstrum should be used. Consequently, this dimension can become the main contributor to the computational complexity of speech synthesis. Therefore, this extraction should be also performed by fast algorithms.

In the following, it is supposed the frame period of speech analysis is the same as the period of the subband vector sequence. Initially, the mel-cepstrum for each frame is converted into a log-power in the mel-scale spectrum by an FFT or DCT operation, the computational complexities of which are $O(N \log N)$. Then, at the center of each subband pair in the mel frequency scale, the log-power coefficient is extracted from a log-spectral envelop built by linear interpolation. Next, the log power coefficients at the centers of the subband pairs are converted into linear power coefficients. Power operations in this conversion can be effectively calculated with table lookup, shift operation and linear interpolation. Finally, when the target is voiced, the linear power coefficients are modified for amplitude modification factors because power normalization for the fundamental period is required when an impulse train is built from impulses. On the other hand, when the target is unvoiced, the linear power coefficients equal the amplitude modification factors of the subbands.

In the above algorithms, the worst operation in terms of computational complexity is FFT (or fast DCT), the complexity of which is $O(N \log N)$ per frame, i.e. $O(\log N)$ per sample. Thus, the computational complexity of the whole algorithm is still $O(\log N)$ per sample.

5. Subjective evaluation

To examine the practicality of the proposed method, a mean opinion score (MOS) test of the synthetic sounds was conducted. In the test, 10 subjects listened to synthetic speech sounds 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.

Stimuli consisted of resynthesized speech sounds from the analysis results of natural speech sounds spoken by one male and one female narrator. For speech analyses, Speech Processing Toolkit (SPTK) version 3.5[14] was used. In this test, the sampling rate was 16 kHz. In speech analysis, Blackman window was applied first, then 39th-order mel-cepstral coefficients where frequency warping coefficient $\alpha = 0.42$ were extracted with the *mcep* tool in the SPTK. Note that a 39th-order mel-cepstrum was also used in an HMM-based speech synthesis

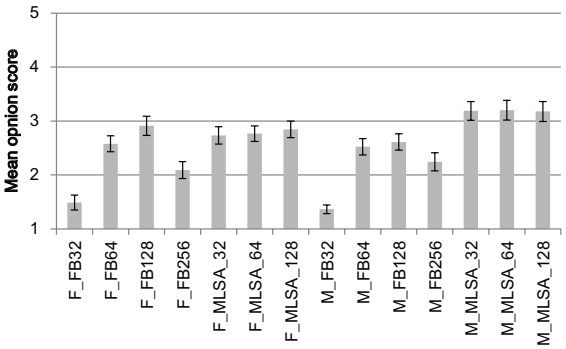


Figure 7: *Result of the subjective evaluation.* In conditions, prefix F and M correspond to the female and male narrator, and FBxx and MLSA_yy correspond to the synthetic sounds generated by the filter bank where the number of subbands is xx and those by the MLSA filter where the frame period is yy samples, respectively. The horizontal bars indicate the 95% confidence intervals.

system in Blizzard Challenge 2005[15]. The fundamental frequency (F_0) was also extracted with the *pitch* tool in the SPTK. For the proposed method, the number of subbands M is 32, 64, 128 or 256 where the frame period is $M/2$ samples. Subband amplitude modification factors were determined from the extracted mel-cepstra 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 an MLSA filter were also prepared. As the MLSA filter and excitation source for speech synthesis, the *mlsadf* and *excite* tools in the SPTK were used. The frame period was set to 32, 64 or 128. Both mel-cepstra for the proposed method and the MLSA-filter-based method were the same where the frame periods were equal. Consequently, there were 7 conditions for one narrator.

For each condition, stimuli for 10 sentences that were similar to those from J01 to J10 of the ATR503 corpus[16] were prepared. Thus, 140 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 around 3 even for good conditions might be caused by the down-sampling. In the test, no information that high-frequency components of the sounds were cut off by the down-sampling were presented beforehand. For the female narrator, preference for the proposed method where $M = 128$ was comparable to that for the MLSA-based system, while it was poor where the number of subbands was small. In the proposed method, since the frame period is fixed to $M/2$, larger M also resulted in a lower preference score by insufficiency of temporal resolution. On the other hand, for the male narrator, the preference scores under any condition were inferior to that of the conventional method. Male speech sounds with lower F_0 may require higher resolution at low frequencies. If the number of subbands is increased while maintaining the frame period, better quality sounds may be generated. Although such a configuration proportionally increases computational costs, the computational cost can be still lower than that of the conventional method using the MLSA filter. For example, the total cost per sample of two FFT-like operations for calculation of the amplitude modification factors and composition of subbands even where M and the frame period are 256 and 64 samples, respectively, is significantly lower than that of the MLSA filtering where the order is 39 because the computational cost for the MLSA filtering is high as mentioned previously. However, to investigate the practicality of the proposed method in more detail, an examination under practical conditions should be conducted using practically optimized speech synthesizers because precise estimation of the computational cost, especially for the MLSA filtering, is also difficult.

6. Conclusion

A speech synthesis method using a non-maximally decimated filter bank to generate spectral features of synthetic speech sounds was proposed. To control the amplitude of each neighboring pair of subbands with alias cancelations, non-maximally decimated filter banks were adopted. In the proposed method, speech sounds were composed from pre-decomposed noise and impulse waveforms. Moreover, with a fast calculation method for the amplitude modification factors, $O(logN)$ speech waveform generation could be achieved. For evaluation, a MOS test using resynthesized speech sounds from the analysis results of natural speech sounds was conducted. The results of the MOS test showed that the proposed method can achieve scores similar to those of the conventional method using the MLSA filter for a female narrator.

In future work, the performance of the proposed method in terms of the computational cost will be examined using practical speech synthesizers.

7. 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, Sept. 1999.

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

[4] Nakatani, N., Yamamoto, K., Matsumoto, H. "Mel-LSP parameterization for HMM-based speech synthesis," in Proc. SPECOM. pp. 261–264, June 2006.

[5] Ling, Z. H., Wu, Y. J., Wang, Y. P., Qin, L. and Wang, R. H., "USTC system for Blizzard Challenge 2006 – An improved HMM-based speech synthesis method," in Proc. of Blizzard Challenge 2006 workshop, Sept. 2006.

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

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

[9] Painter, T., Spanias, A., "Perceptual coding of digital audio," Proc. of the IEEE, vol. 88, No. 4, pp. 451–513, April, 2000.

[10] 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.

[11] 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.

[12] Rossi, M., Zhang, J. and Steenaart, W., "Iterative Constrained Least Squares Design of Near Perfect Reconstruction QMF Banks," in Proc. CCECE '96, Calgary, Alta., Canada, vol. 2, pp. 766–769, May 1996.

[13] Bolcskei, H. and Hlawatsch, F., "Oversampled cosine modulated filter banks with perfect reconstruction," IEEE Trans. on Circ. and Syst. II, vol. 45(8), pp. 1057–1071. Aug. 1998.

[14] 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://sp-tk.sourceforge.net/.

[15] Zen, H. and Toda, T., "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in Proc. of Interspeech 2005–Eurospeech, Lisboa, Spain, pp. 93–96, Sept. 2005.

[16] 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).