# Speech synthesis using a maximally decimated pseudo QMF bank for embedded devices

*Nobuyuki Nishizawa and Tsuneo Kato*

KDDI R&D Laboratories Inc., Japan

{no-nishizawa, tkato}@kddilabs.jp

## Abstract

A fast speech waveform generation method using a maximally decimated pseudo quadrature mirror filter (QMF) bank is proposed. The method is based on subband coding with pseudo QMF banks, which is also used in MPEG Audio. In the method, subband code vectors for speech sounds are synthesized from magnitudes of spectral envelope and fundamental frequencies for periodic frames, and then waveforms are generated by decoding of the vectors. Since the synthesizing of the vectors is performed at the reduced sampling rate by the maximal decimation and the decoding is processed with fast discrete cosine transformation algorithms, faster speech waveform generation is achieved totally. Although pre-encoded vectors for noise components were used to reduce the computational costs in our former studies, in this study, all code vectors for noise components are made with a noise generator at run time for small footprint systems. In contrast, a subjective test for synthetic sounds by HMM-based speech synthesis using mel-cepstrum showed the proposed method was comparable to our former method and also the conventional method using a mel log spectrum approximation (MLSA) filter in quality of sounds.

**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 for embedded devices since it can generate high-quality sounds with small footprints such as several hundred kilobytes or megabytes. However, in the computational cost on the devices, the HMM-based speech synthesizers are often inferior to concatenative speech synthesis with a small waveform segment database optimized for embedded systems. This is because waveform generation in the HMM-based speech synthesis is performed by means of costly signal processing like the mel log spectrum approximation (MLSA) filters [3] in which several hundreds multiply-accumulate operations per output sample are required.

As a more efficient method for waveform generation, we studied speech synthesis using filter banks performed at the reduced sampling rate [4, 5]. In the studies, the pseudo quadrature mirror filter (QMF) banks [6, 7], which are also used in MPEG Audio [8], have been the main focus because faster processing is possible with fast Fourier tansformation (FFT) or similar fast discrete cosine transformation such as Lee's DCT [9]. Our previous method [5] was based on a filter bank-based speech synthesis. In the method, white source waveforms such as noise sequences or impulses were initially decomposed into several bands by a set of band-pass filters. Then, the amplitudes of the decomposed (i.e. band-limited) waveforms were scaled for each band to build spectral features. Finally the decomposed and scaled waveforms were composed into speech waveforms by simple summation. These operations were performed on the subband domain in the subband coding system. Thus, in the method, decomposed source waveforms were initially subband-coded. The scaling and the summation were performed against the code vectors at the reduced sampling rate, not waveforms. Since the code vectors from band limited waveform could be sparse, the computational cost for the waveform generation can be reduced even though the cost for decoding of the code vectors is taken into account. Moreover, sinusoidal synthesis [10] performed in the subband domain was also introduced for accurate reproduction of spectra, where the amplitude of each harmonic component is directly configurable independent of the filter bank-based speech synthesis. By these techniques, fast processing without degradation in quality of sounds was achieved.

However, for fast processing, use of predecomposed and pre-encoded source waveforms was expected in the method. Although it removed the filter banks for the source waveform decomposer and encoder from speech synthesizers, it instead required large storage for the predecomposed and pre-encoded code vectors of impulses and noises. Therefore, in this study, an improved method to generate code vectors without code vector storage is proposed. In the proposed method, aperiodic components are directly constructed by the pseudo QMF bank from

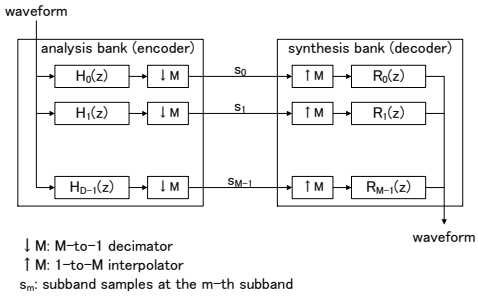


Figure 1: Block diagram of the subband coding system based on the maximally decimated filter banks.

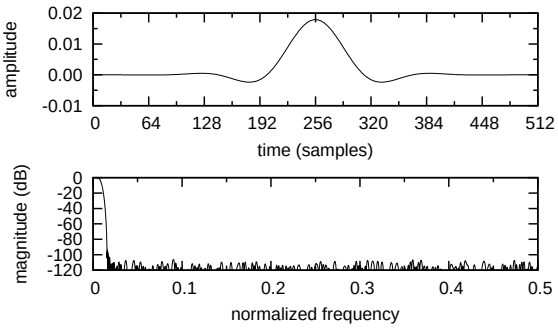


Figure 2: Impulse response and magnitude-frequency response of the prototype filter for a 32-band pseudo QMF bank.

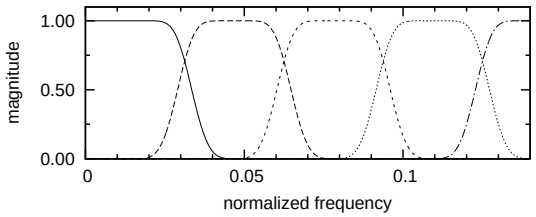


Figure 3: Magnitude-frequency response of the filters for 0th (lowest) to 4th bands in the analysis bank. (The filters correspond to $H_0(z)$ to $H_4(z)$ in Figure 1.)

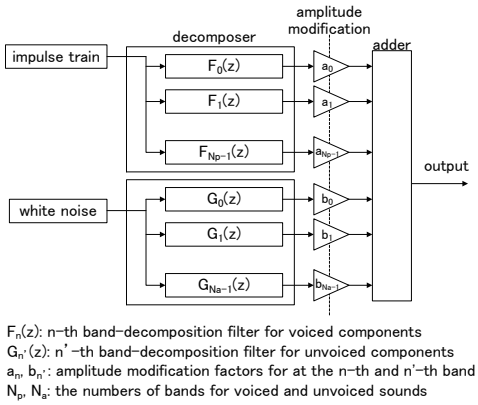


Figure 4: Block diagram of the filter bank based speech synthesizer.

white noise at run time. In contrast, all periodic components are synthesized by sinusoidal synthesis, which is similar to our former method. The result of a subjective test indicates that the proposed method can generate speech waveforms with quality comparable to those of our former method and the conventional MLSA filter-based method.

The rest of the paper is organized as follows. First, in Section 2, our former waveform generation method based on filter banks is introduced. Section 3 explains the proposed method. Section 4 gives a subjective test to evaluate degradation in speech sound quality by the proposed method. Finally, section 5 concludes the paper.

## 2. Waveform generation performed on subband coding

### 2.1. Subband coding based on maximally decimated pseudo QMF bank

Figure 1 shows a block diagram of a subband coding system with maximally decimated filter 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.

Commonly, analysis and synthesis filters of pseudo QMF banks are made by cosine modulation of the impulse response of a prototype filter. The cosine modulation corresponds to shifting of the magnitude response along the frequency axis. In this study, the impulse responses of the filters of the analysis and synthesis banks are given by:

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

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

where $h(i)$ is the impulse response of the prototype filter. These equations are from the MPEG Audio specification [8]. 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. This configuration is the same as that in our former study [5]. Thus, filter banks that are components of highly optimized MPEG Audio decoders like [11] are directly applicable to systems for the proposed method. Figure 2 shows plots of the coefficients and magnitude-frequency response of the prototype filter, respectively. Figure 3 shows the magnitude-frequency responses of the cosine modulated filters for the analysis bank.

### 2.2. Filter bank-based speech synthesis on the subband coding system

Figure 4 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 [12] 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.

In contrast, the cosine modulation of the filter coefficients corresponds to shifting of the magnitude response along the frequency axis. The cosine modulation for the n-th filter $f_n(i)$ for band-decomposition is performed by the following equation:

$$f_n(i) = 2f(i)\cos\Big(\frac{\pi}{2N}(2n+1)i\Big)\ (0 \le n \le N-1) \quad (3)$$

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

Thus, a filter bank for white source waveform decomposition can be built where the summation of all outputs of bands approximately performed the all pass characteristic:

$$|F_n(\omega) + F_{n+1}(\omega)| \approx 1 \quad (4)$$

$$\text{for } \frac{\pi}{N}(n+\frac{1}{2}) \le \omega \le \frac{\pi}{N}(n+\frac{3}{2}),\ 0 \le n < N-1$$

Consequently, the magnitude at $\omega$ is given as:

$$\begin{aligned}&|a_n F_n(\omega)S(\omega) + a_{n+1}F_{n+1}(\omega)S(\omega)| \\ &\quad = a_n|F_n(\omega)| + a_{n+1}|F_{n+1}(\omega)| \end{aligned} \quad (5)$$

where $a_n$ and $S(\omega)$ denote the amplitude modification factor for n-th band and the magnitude response of a white noise, respectively, and $|S(\omega)| = 1$.

Of course, the filter bank-based method increases the computational cost due to processing for multiple bands. As a method to reduce the computational cost, sampling rate reduction with the bandwidth limitation was studied. Consequently, the filter bank-based speech synthesizer was implemented on the subband coding system based on the pseudo QMF bank in our former study [5]. In the method, encoding and decoding of the subband coding were performed after the band decomposition and after the composition of the bands. Since predecomposition and pre-encoding of the white source waveforms are adopted, there is neither filter bank for the source decomposition nor encoder for the subband coding.

Subband-coded vectors for band limited waveforms can be sparse. Elements of the coded vectors corresponding to the overlapped region of the stop bands of bandpass filters for source waveform decomposition and encoding of the subband coding can be regarded as zero since all components are cut in the region. For such sparse vectors, the computational cost for the scaling to construct spectral features is not high. Although multiple elements in the coded vector were non-zero values due to the overlapped structure of the filter bank in the encoder, affection of sampling rate reduction by the subband coding is more significant. Consequently, the computational cost can be reduced by adoption of the subband coding.

### 2.3. 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 the 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.

For this conversion, initially, the mel-cepstrum for each frame is converted into a log-power in the mel-scale spectrum by a DFT or DCT operation. Then, at the center of each subband in the mel frequency scale, the log-power coefficient is extracted from a log-spectral envelope built by linear interpolation. Next, the log power coefficients, which equal the amplitude modification factors where the power of source is normalized, are converted into linear power coefficients. Power operations in this conversion can be effectively calculated with table lookup, shift operation and linear interpolation.

In the above algorithms, the worst operation in terms of computational complexity is DFT (or DCT), the complexity of which can be $O(N \log N)$ per frame, i.e. $O(\log N)$ per sample.

### 2.4. Sinusoidal synthesis in the subband domain

In our first study [4], spectral features of voiced sounds were constructed by the scaling of each element of coded vectors. However, insufficiency of the resolution along the frequency axis caused degradation of the quality of the synthetic sounds. Therefore, speech synthesis by si-

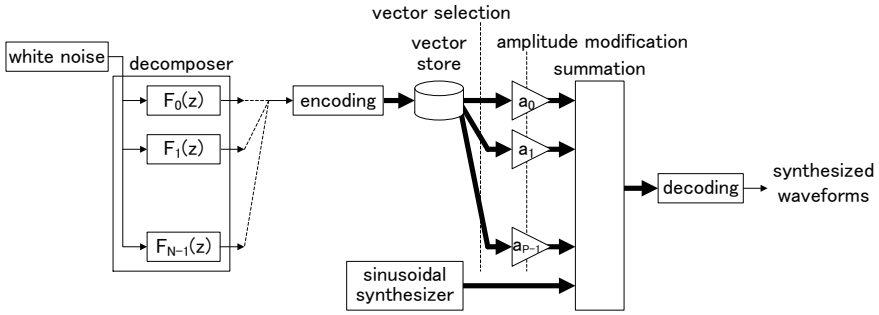


Figure 5: Block diagram of the speech synthesizer proposed in our former study [5]. Bold lines correspond to coded vectors.

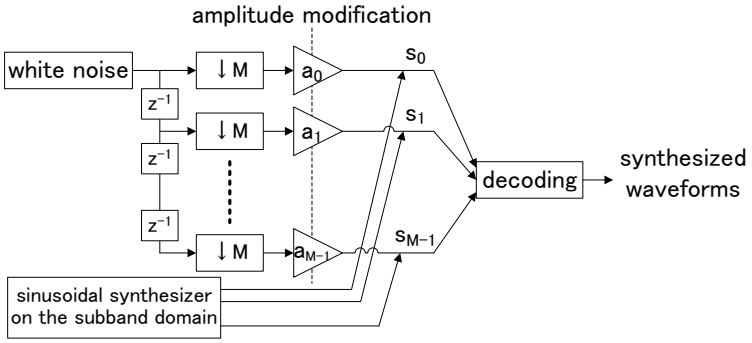


Figure 6: Block diagram of the speech synthesizer based on the proposed method.

nusoids [10] was introduced in our former study [5] 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 are reduced.

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

$$x(t) = \sum_k A_p(\omega_0) \cos\bigl(\omega_0 k(t - t_0)\bigr) \qquad (6)$$

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

Since magnitude-frequency and phase-frequency responses of the filters of the analysis bank are given, encoded results of sinusoids are easily obtained by calculations in the 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, the m-th element of the subband vector of encoded cosine waveform with angular frequency $\omega$ is given as follows:

$$x_{\omega,m}(t) = |H_m(\omega)| A_p(\omega) \cos\bigl(\omega(t - t_0) + \arg H_m(\omega)\bigr) \qquad (7)$$

where $A_p(\omega)$ denotes the magnitude of the periodic component at angular frequency $\omega$.

Referring to Eq. (1), $\arg H_m(\omega) = -\pi(2m + 1)/4$. In contrast, $|H_m(\omega)|$ is easily obtained with a precalculated table. For example, a 512-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 overlapping structure of the analysis bank (refer to Fig. 3). By these operations, the sinusoidal synthesis is performed in the subband domain with the reduced sampling rate in the maximally decimated filter bank.

Consequently, the system shown in Fig. 5 was introduced in our former study.

## 3. Run-time generation of code vectors for aperiodic components

In the proposed method, elements of vectors for aperiodic components are generated on demand. In other words, aperiodic components are built by the combination of subband code vectors each of which has only one nonzero element. Referring to Fig. 1, in the first step of decoding of subband code vectors, bandwidth expansion is performed, then the expanded components are filtered by the band-pass filters in the decoder. However, different from the decoding of common subband coding based on the maximally decimated QMF banks, no alias cancel-

ing is taken into account for aperiodic components in the proposed method.

Now, two neighboring elements in the code vector are focused on in the following discussion. Commonly, prototype filters for pseudo QMF banks are designed with the condition where squared magnitude responses of analysis and synthesis banks approximately equal the magnitude responses of the N-th band filter to achieve approximately perfect reconstruction of the input. The prototype filter for the MPEG Audio also has similar properties. Therefore, in the subband coding system based on pseudo QMF banks, summation of the squared responses of neighboring subbands that overlap each other has an approximately all-pass characteristic:

$$\begin{aligned} &|H_m(\omega)R_m(\omega) + H_{m+1}(\omega)R_{m+1}(\omega)| \\ &\quad = |R_m(\omega)|^2 + |R_{m+1}(\omega)|^2 \approx 1 \end{aligned} \tag{8}$$

$$\text{for } \frac{\pi}{M}(m+\frac{1}{2}) \leq \omega \leq \frac{\pi}{M}(m+\frac{3}{2}),\ 0 \leq m < M-1$$

where $H_m(\omega)$ and $R_m(\omega)$ are the magnitude responses of the filters for the m-th subband in the analysis and synthesis bank, respectively. In general, when two elements are set from independent white noises, the variance of summation of the two elements equals the summation of the variances of the two elements. Thus, the magnitude at $\omega$ is given with independent white noises as:

$$\begin{aligned} &|a_m R_m(\omega)S_m(\omega) + a_{m+1}R_{m+1}(\omega)S_{m+1}(\omega)| \\ &\quad = (a_m^2|R_m(\omega)|^2 + a_{m+1}^2|R_{m+1}(\omega)|^2)^{1/2} \end{aligned} \tag{9}$$

where $S_m(\omega)$ denotes an independent white noise and $|S_m(\omega)| = 1$. Where $a_m = a_{m+1}$, spectral features become flat similar to our former method.

Thus, vectors for aperiodic components can be built with an independent white noise. Figure 6 shows a block diagram of the speech synthesizer based on the proposed method. Structure with delays and down-samplers in Fig. 6 corresponds to sequentially value assignment into elements of the coded vectors.

For example, in our former system [5], a cyclic quasi noise series for which the period was 4096 samples (128 frames) was used for aperiodic components. Since only neighboring 3 elements rather than $M$ $(= 32)$ elements for each vector of predecomposed bands were required to be stored for aperiodic components, the required size of the table is 12288 $(= 32 \times 3 \times 128)$ for the 32-band predecomposition. In contrast, the table is unnecessary for the proposed method. Consequently, in the proposed method, tables for cosine function for the sinusoidal synthesis, the filter coefficients and magnitude response of the 512-tap prototype filter for the pseudo QMF bank are necessary.

However, although the filter banks for the source decomposition and the subband coding became separated in our former system for flexibility of the source decomposition, the decomposition of noise in the proposed method depends on the design of the pseudo QMF bank again. Nevertheless, the result of the experiment in our former study where the sinusoidal synthesis was adopted implied that error in the reproduction of the spectral feature especially for low band was subjectively problematic only for periodic components since the frequency resolution in the aperiodic source decomposition was comparable to that in the subband coding.

## 4. Subjective evaluation

In practice, quality of speech sounds should be at least comparable to those of the conventional methods even when faster and smaller waveform generation is achieved. Although the difference between our proposed and former methods is limited only in aperiodic components, to evaluate the quality of sounds subjectively,a mean opinion score (MOS) test of the synthetic sounds was conducted.

In the test, the synthesis targets were generated by our HMM-based speech synthesizer using melcepstrum for a male and female voice. HMMs were trained from 6.1-hour and 5.8-hour sounds for the male and female voice, respectively. For spectral features, 39-order melcepstrum was used where the warping factor $\alpha = 0.4375$ $(= 7/16)$. All voiced sounds were synthesized only from periodic components. This corresponds to excitation only by impulse trains in the conventional methods. 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. In this test, the sampling rate was 16 kHz. Although the frame period of the HMMs was 5 ms, that in the waveform generation by the proposed method was 2 ms (32 samples) because the number of subbands was fixed at 32. This conversion was performed with linear interpolation. Subband amplitude modification factors were determined at the center frequencies of the bands, and extracted from the extracted melcepstrum through power spectra in the mel-scale with linear interpolation, where the number of dimensions for mel-spectrum was 64 for all conditions. For comparison, we also prepared speech sounds synthesized using a 39-order MLSA filter, and our former filter bank-based method using pre-encoded vectors where the number of subbands was 32. Although impulse trains were also used to generate periodic components in our former study [5], all periodic components for this test were synthesized from sinusoids.

For each condition, stimuli for 10 sentences that were similar to those from J01 to J10 of the ATR503 corpus [13] were prepared. The stimuli were randomly ordered for each subject and presented to both ears through closed-ear headphones in a silent room.

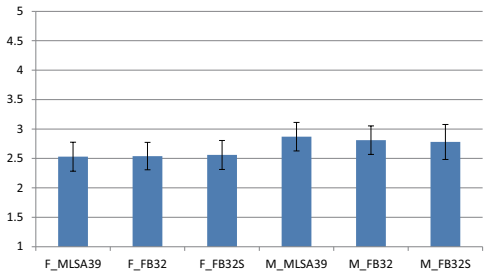


Figure 7: Result of the subjective evaluation. In the conditions, prefix F and M correspond to female and male sounds, and suffix MLSA39, FB32, and FB32S correspond to the conventional method using a 39-order MLSA filter, the conventional 32-band filter bank-based method with pre-encoded vectors and the proposed method, respectively. The error bars indicate the 95% confidence intervals.

Figure 7 shows the results of the test. The scores were comparable among all conditions for each of the male and female voices, i.e., the proposed method can be a substitute for the waveform generation method. Although approximatey 0.3 point differences were observed between the male and female voices, they would depend only on the HMMs because of the comparable results for all conditions in each voice.

It should be noted that the method based on filter banks for resynthesized speech with parameters extracted from natural speech sounds was subjectively superior to the method using the MLSA filter in the former study [5]. The result could be caused by fluctuation of the filter parameters estimated from natural speech sounds; MLSA filters with insufficient margins in Padé approximation can be unstable temporarily, especially when the parameters vary quickly. By contrast, the proposed method, which is a finite impulse response system, is stable anytime. Nevertheless, in the HMM-based speech synthesis, smoothed trajectories of the parameters with consideration of delta and delta-delta features are commonly generated [1]. This would be the reason why no difference was observed in this test.

## 5. Conclusion

This paper presented a waveform generation method using a pseudo QMF bank for embedded devices. In the method, all coded vectors were synthesized at run time with cosine functions for periodic components and a white noise generator for aperiodic components. The results of a subjective test using synthetic speech sounds indicated that the method was comparable to both the conventional methods with an MLSA filter and our former method by a filter bank with pre-encoded vectors in terms of the quality of sounds. Therefore, using the proposed method, speech synthesizers with smaller footprints than those of the conventional systems can be built without degradation in sound quality.

## 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] Imai, S., “Cepstral analysis synthesis on the mel frequency scale,” in Proc. ICASSP ’83, vol. 8, pp. 93–96, Apr. 1983.

[4] 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.

[5] Nishizawa, N. and Kato, T., “Speech synthesis using subband-coded multiband source components and sinusoids,” in Proc. ICASSP 2013, Vancouver, Canada, pp. 8002–8006, May. 2013.

[6] 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.

[7] 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.

[8] 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.

[9] Lee, B. G., “A new algorithm to compute the discrete cosine transform,” IEEE Trans. on ASSP, vol. 32(6), pp. 1243–1245, Dec. 1984.

[10] 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.

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

[12] 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.

[13] 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).