Speech Synthesis by Modeling Harmonics Structure with Multiple Function

Toru Nakashika[1], *Ryuki Tachibana*[2], *Masafumi Nishimura*[2], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1]

[1]Department of Computer Science and Systems Engineering, Kobe University, Japan
[2]IBM Research - Tokyo, Japan

nakashika@me.cs.scitec.kobe-u.ac.jp, ryuki@jp.ibm.com, nisimura@jp.ibm.com,
takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

In this paper, we present a new approach for the speech synthesis, in which speech utterances are synthesized using the parameters of spectro-modeling function (Multiple function). With this approach, only harmonic-parts are extracted from the phoneme spectrum, and the time-varying spectrum corresponding to the harmonics or sinusoidal components is modeled using the Multiple function. We introduce two types of the functions, and present the method to estimate the parameters of each function using the observed phoneme spectrum. In the synthesis stage, speech signals are generated from the parameters of the Multiple function. The advantage of this method is that it only requires a few speech synthesis parameters. We discuss the effectiveness of our proposed method through experimental results.

Index Terms: speech synthesis, text-to-speech, multiple function, harmonic-temporal structure, EM algorithm

1. Introduction

Text-to-speech (TTS) systems, which artificially produce human voices from a text, are used for many applications, including public address systems and speech devices for those who have difficulty speaking clearly. Various methods of speech synthesis technology have been proposed to date. Concatenative synthesis [1, 2], one of the most widely-used speech synthesis methods, is a method that synthesizes a speech signal by selectively concatenating speech fragments. Another method for synthesizing speech signals, Additive synthesis [3, 4], has also been proposed. In this method, a speech signal is synthesized by summing sinusoidal waves corresponding to each harmonic at a certain rate.

Concatenative synthesis is a method that generates a speech signal by means of appropriate concatenation of short-term-recorded speech pieces. This method produces relatively natural speech because it uses the uncompressed recorded speech data. It does have a problem, however, in that it might generate unnatural speech if the selected speech pieces are not selected properly. Furthermore, because the Concatenative synthesis method needs a large amount of speech fragment data included in a database, an enormous amount of computational resources, such as memory, disk space and CPU allocation, is required.

In the Additive synthesis approach, the synthesized speech is generated by adding sine waves of harmonic partials as formant information [3, 4]. Since this method represents speech using only a few parameters rather than all the speech signals, fewer computational resources are needed than for the Concatenative synthesis. The naturalness of the synthesized speech signals, however, tends to be inferior to that produced by the Concatenative approach.

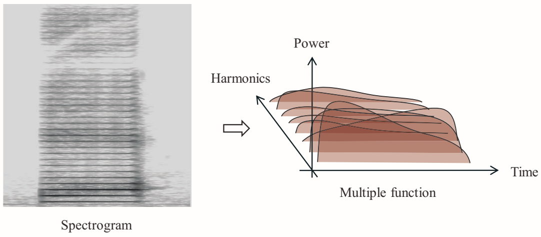


Figure 1: Modeling of an envelope shape in a phoneme spectrum. Only harmonic-parts are extracted, and replaced by a Multiple function.

Based on the Additive synthesis approach, we propose a speech synthesis technique in which the time-varying power (intensity) spectrum of each harmonic for each individual phoneme is approximated by a Multiple function (see section 2), and then the output speech signal is given from the parameters of the Multiple function (Fig. 1). Only harmonic-parts are extracted from the phoneme spectrum, and the harmonic-temporal structure of speech utterances is modeled using the Multiple function (non-harmonic partials in voiced speech are not modeled).

2. Multiple function

In this paper we propose a method of speech synthesis, in which the harmonic-temporal shape of the utterance spectrum is approximated to a Multiple function. We consider that the Multiple function best models an envelope shape given the following constraints:

- When the discrete-frequency and continuous-time domain are used.
- When the integral value over the whole time/frequency space equals 1.
- When the parameters are estimated by an EM algorithm or Maximum Likelihood.

Satisfying these conditions, we define the Multiple function as in Eq. (1). This function can express the harmonic-temporal spectral structure as a whole, by preparing individual amplitude functions corresponding to each harmonic.

$$q(t, n; \boldsymbol{\Theta}, \boldsymbol{\pi}) = \sum_{n} \pi_n \, p_n(t; \Theta_n) \qquad (1)$$

where t is a variable of time, and n is an index of harmonic. $p(t; \boldsymbol{\Theta})$ means Partial function, which is the time-varying amplitude function to a harmonic, and satisfies

$$\forall n, \int p_n(t)dt = 1. \tag{2}$$

Multiple function has two kinds of parameters: $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$. $\boldsymbol{\Theta}$ implies the parameter matrix of the Partial function, and $\boldsymbol{\pi}$ represents the multiply rate vector within the Partial functions. $\boldsymbol{\pi}$ satisfies

$$\sum_n \pi_n = 1, \quad \forall n, \pi_n > 0, \tag{3}$$

and its parameters can be estimated as follows:

$$\kappa_n = \frac{\int g_n(t)dt}{\int g_1(t)dt} \tag{4}$$

$$\pi_n = \frac{\kappa_n}{\sum_m \kappa_m} \tag{5}$$

where κ_n means the intensity ratio between the 1st harmonic and the n-th harmonic. $g_n(t)$ is the observed intensity value of the n-th harmonic.

In this paper, we introduce two types of the Multiple functions: Multi-Gaussian Mixture Model and Multi-Beta Mixture Model. Since both of these functions are based on mixture models, it is expected that the two can represent complicated shapes, such as the utterance spectrum, which have multiple peaks in the time domain. We will describe the models in more detail in following subsections.

2.1. Multi-Gaussian Mixture Model

We define here Multi-Gaussian Mixture Model (MGMM) as one of the Multiple functions, whose Partial function is as a Gaussian Mixture Model (GMM). The MGMM is formulated as in Eq. (1) and (6).

$$p_n(t; \nu_n, \mu_n, \sigma_n) = \sum_l \nu_{n,l} \frac{1}{\sqrt{2\pi}\sigma_{n,l}} \exp\left\{-\frac{(t-\mu_{n,l})^2}{2\sigma_{n,l}^2}\right\} \tag{6}$$

where

$$\forall n, \sum_l \nu_{n,l} = 1, \quad \forall n, l, \nu_{n,l} > 0. \tag{7}$$

$\nu_{n,l}$ is mixing rate, and l represents an index of mixture components.

Next, we derive the update rules for parameters ν_n, μ_n and σ_n in Eq. (6). To achieve this, we introduce Kullback-Leibler (KL) divergence for the evaluation function J. The KL divergence is a measure of the difference between two distributions. We define the evaluation function J as below:

$$J = \sum_n J_n = \sum_n \int_{-\infty}^{\infty} g_n(t) \log \frac{g_n(t)}{p_n(t)} dt. \tag{8}$$

We also define $u_{n,l}$ and $v_{n,l}$ as

$$u_{n,l} = \frac{\nu_{n,l}}{\sqrt{2\pi}\sigma_{n,l}} \exp\left\{-\frac{(t-\mu_{n,l})^2}{2\sigma_{n,l}^2}\right\} \tag{9}$$

$$v_{n,l} = \int_{-\infty}^{\infty} \frac{g_n(t)u_{n,l}}{p_n(t)} dt \tag{10}$$

respectively.

J in Eq. (8) is the KL divergence between MGMM and the observed spectrum shape. The smaller the J is, the closer the MGMM and the observed spectrum of a speech signal are.

Using Lagrange multipliers, we obtain update rules (11), (12), and (13) for MGMM parameters, which minimize the evaluation function J under condition (7).

$$\hat{\nu}_{n,l} = \frac{v_{n,l}}{\sum_m v_{n,m}} \tag{11}$$

$$\hat{\mu}_{n,l} = \frac{\int_{-\infty}^{\infty} \frac{t \cdot g_n(t)u_{n,l}}{p_n(t)} dt}{v_{n,l}} \tag{12}$$

$$\hat{\sigma}_{n,l} = \sqrt{\frac{\int_{-\infty}^{\infty} \frac{(t-\mu_{n,l})^2 g_n(t)u_{n,l}}{p_n(t)} dt}{v_{n,l}}} \tag{13}$$

Thus, updating parameters as in from (9) to (13) iteratively, the parameters of MGMM can be optimized gradually.

2.2. Multi-Beta Mixture Model

Next, we define Multi-Beta Mixture Model as one of the Multiple functions, where its Partial function is as a Beta Mixture Model (BMM). The Multiple function can be formulated as in Eq. (1) and (14).

$$p_n(t; \nu_n, \alpha_n, \beta_n) = \sum_l \nu_{n,l} \frac{1}{B(\alpha_{n,l}, \beta_{n,l})} t^{\alpha_{n,l}-1}(1-t)^{\beta_{n,l}-1} \tag{14}$$

where $B(\alpha, \beta)$ is a Beta function. Eq. (14) is definitional formulation for BMM, and its parameters can be estimated by EM algorithm [7]. The update rules of the parameters in M-step are as follows. (For brevity, we omit the description of the details of the derivation of the update rules.)

$$\hat{\nu}_{n,l} = \frac{\sum_{i=1}^{K_n} z_{n,l,i}^*}{K_n} \tag{15}$$

$$\hat{\alpha}_{n,l} = \Psi^{-1}\left(\frac{1}{K_n}\sum_{i=1}^{K_n} \log\left(\frac{X_{n,i}}{1-X_{n,i}}\right) + \Psi(\beta_{n,l})\right) \tag{16}$$

$$\hat{\beta}_{n,l} = \Psi^{-1}\left(\frac{1}{K_n}\sum_{i=1}^{K_n} \log\left(\frac{1-X_{n,i}}{X_{n,i}}\right) + \Psi(\alpha_{n,l})\right) \tag{17}$$

where $\Psi(x)$ is the digamma function, and $\Psi^{-1}(x)$ is the inverse-digamma function. All $X_{n,i}$ are samples, obtained from the random generation along the observed amplitude spectrum of harmonic n. K_n is the number of samples $X_{n,i}$.

$z_{n,l,i}^*$ in Eq. (15) is the latent indicator variable [7]. This variable means the probability of the occurrence of a sample $X_{n,i}$ from the l-th component for the n-th BMM. $z_{n,l,i}^*$ can be updated in E-step as follows;

$$z_{n,l,i}^* = \frac{\hat{\nu}_{n,l} f_{n,l}(X_{n,i}|\hat{\alpha}_{n,l}, \hat{\beta}_{n,l})}{\sum_j \hat{\nu}_{n,j} f_{n,j}(X_{n,i}|\hat{\alpha}_{n,j}, \hat{\beta}_{n,j})} \tag{18}$$

$$f_{n,l}(X_{n,i}|\hat{\alpha}_{n,l}, \hat{\beta}_{n,l}) = \frac{X_{n,i}^{\hat{\alpha}_{n,l}-1}(1-X_{n,i})^{\hat{\beta}_{n,l}-1}}{B(\hat{\alpha}_{n,l}, \hat{\beta}_{n,l})}. \tag{19}$$

By adequately repeating calculations for the E-step and M-step updates as mentioned above, the parameters of MBMM $\Theta = \{\nu_n, \alpha_n, \beta_n\}$ can be estimated.

3. Speech synthesis from the parameters

In this section, we discuss the technique for synthesizing speech sounds from Multiple function parameters. The phoneme signals can be synthesized using the Additive synthesis approach [5].

In the Additive synthesis approach, the synthesized signal $s(t)$ can be formulated as

$$s(t) = \sum_n a_n(t) \sin\left(\frac{2\pi f_n t}{T}\right) \tag{20}$$

where f_n is the frequency of the n-th harmonic, and T is the voice activity term. Each harmonic of Multiple function has its own frequency f_n

$$f_n = n \cdot f_{pitch} \tag{21}$$

where f_{pitch} denotes the fundamental pitch of the speech sound.

Setting $a_n(t)$ as in Eq. (22), it is possible to synthesize a speech signal from the pre-learned model (Multiple function's) parameters.

$$a_n(t) = \pi_n \cdot p_n(\frac{t}{T}; \Theta_n) \tag{22}$$

where $p_n(t)$ is Partial function.

4. Experiments

4.1. Experimental setup

To evaluate the performance of our approach, we carried out an experiment in which we attempted to synthesize 5 long vowels, /a:/, /e:/, /i:/, /o:/ and /u:/, from their own Multiple functions (MGMM and MBMM). The training set for the experiment was recorded by a woman reader at a sampling rate of 22.05 kHz. First, we clipped the 5 phonemes from the recordings using Voice activity detection (VAD) [8]. Employing the PSOLA method [9], we forced the pitch of the signals to be set to 261 Hz over a given period. Using the normalized signals, we calculated each parameter of the Multiple functions for each phoneme. Since our emphasis was on the synthesis technique itself (matching the observed harmonic-temporal spectrum to Multiple function), we evaluated the efficiency of our method of synthesizing phoneme signals as the basic experiment, without synthesizing speech signals from a text using a text analysis technique.

The experimental conditions for both models (MGMM and MBMM) are shown in Table 1. B1 and B2 in Table 1 both refer to MBMM conditions, and G1, G2 and G3 are MGMM conditions. These conditions are different from the others in terms of the number of iterations or the number of mixtures. We set the number of harmonics for spectrum fitting to 20 in each model. In conditions B1 and B2, the number of K_n was set to 2,000 for all n.

As a point of reference, we also compared the result of Multi-Beta Distribution (MBD) condition (A1), whose Partial function is Beta distribution (same as in the case $l = 1$ in MBMM).

4.2. Results and discussion

Fig. 2 shows the results of fitting the harmonic-temporal spectrum of phoneme /e:/ to Multiple functions. Middle and bottom in the figure are the results of MGMM (G3) and MBMM (B2), respectively. We found that the harmonic-temporal structure of input phoneme signals is well-represented by both MGMM and MBMM: intensity-ratios between fundamental and each harmonic, rise and decay of the sound spectrum, or attack time and duration time of each spectral peak.

For clarity, we show the same results in Fig. 3. This figure indicates the comparison of the 2nd harmonic of phoneme /e:/. The vertical and horizontal axis in the figure indicate intensity and time, respectively. The shape of the time-varying spectrum of each model is compared to 'Base' (a), observed spectrum. We can see that the MBD (b), not being a mixture model, does not have an ability to represent the structure with more than 2 peaks, so MBD is not suitable for speech utterances. Meanwhile, MGMM or MBMM, especially G3 (d) or B2 (f), represents the 'Base' spectrum well. Comparing within the same type of Multiple functions, the more mixtures the model has, the more well-fitting to 'Base' the model is, and G3 or B2 is better than G2 (c) or B1 (e), respectively. When we compare different types of mixture models with the same number of mixtures, MBMM is more suitable to 'Base' than MGMM (see G2 and B2). We attribute this to the following: MBMM and MGMM are derived from Beta distribution and Gaussian distribution, respectively. Curvature of the Beta distribution tends to be larger than those of the Gaussian distribution in many cases [6]. Therefore, MBMM can approximate the spectrum shape, shown in time 0.3 to 0.7 in the figure. The above arguments also apply to the other phonemes.

Table 1: Experimental conditions.

	Model	Properties
A1	MBD	-
G1	MGMM	2 mixtures, 20 iterations
G2	MGMM	4 mixtures, 100 iterations
G3	MGMM	8 mixtures, 100 iterations
B1	MBMM	2 mixtures, 200 iterations
B2	MBMM	4 mixtures, 200 iterations

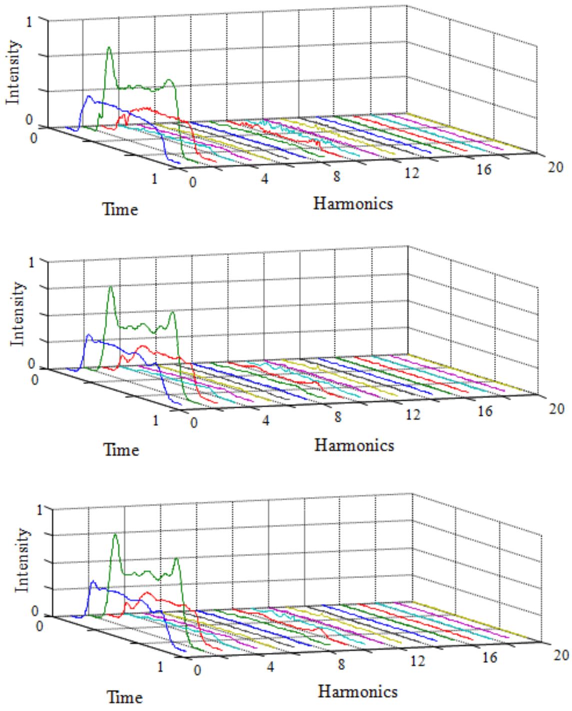
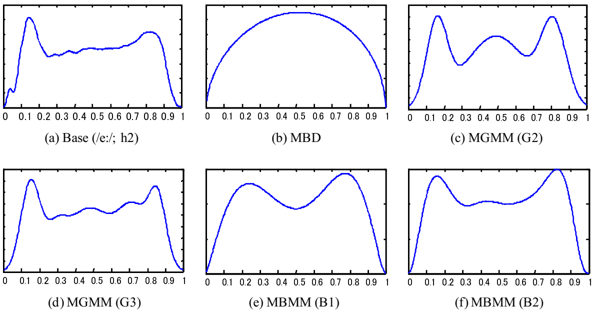


Figure 2: Experimental results. Observed spectrum envelopes of the phoneme /e:/ (top), modeling result of MGMM, G3 (middle) and result using MBMM, B2 (bottom).



(a) Base (/e:/; h2)

(b) MBD

(c) MGMM (G2)

(d) MGMM (G3)

(e) MBMM (B1)

(f) MBMM (B2)

Figure 3: Comparison of spectrum-modeling function shapes (two-dimensional view).

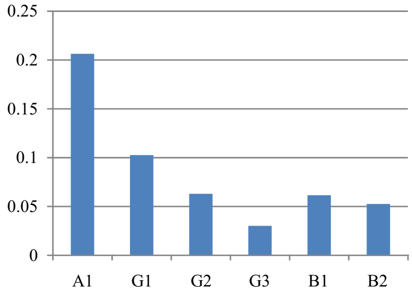


Figure 4: Comparison of DP distances.

Fig. 4 shows the comparison results among tests based on Dynamic Programming (DP) distance to original spectrum, the sum of the DP distances between temporal shape of each harmonic and the observed harmonic spectrum. The small value implies that the model shape is close to the original one. We can see that the more mixtures Multiple function has, the closer to the original spectral structure (G3 to G2, or B2 to B1, etc) it is. In the case of the same number of mixtures, MBMM is more suitable for representing a harmonic-temporal spectrum than MGMM is.

We chose a mean opinion score (MOS) evaluation as a preference test. The listening test was performed using 3 systems: Additive synthesis with observed harmonic-temporal spectrum (consisting of only harmonic partials with no phase; we call it the baseline), MGMM (G3), and MBMM (B2). Fifteen listeners participated in the test. Each listener was asked to rate the naturalness of each utterance on a scale of 1 (the worst) to 5 (the best; that is, raw data). The results of the MOS test are shown in Fig. 5. We see that MGMM or MBMM can produce acoustically more natural utterances than the baseline of Additive synthesis. We believe that this is because some noise or very small fluctuations occur in the recordings, and the approximation with MGMM or MBMM can ignore them.

Finally, we illustrate the parameters of each model in Table 2. The number of parameters $\gamma_{mgmm}, \gamma_{mbmm}$ for MGMM and MBMM, respectively, is given by

$$\gamma_{mgmm} = \gamma_{mbmm} = \rho \cdot (3 \cdot \lambda + 1) \tag{23}$$

where ρ is the number of harmonics (here, 20), and λ is the number of mixtures. We see that the number of required parameters for the Multiple function is much less than those of conventional Additive synthesis.

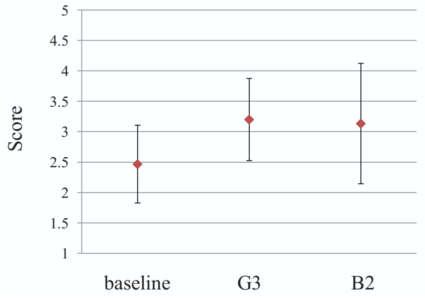


Figure 5: MOS listening test result.

Table 2: Number of parameters for each model.

	Base	G1	G2	G3	B1	B2
No. of parameters	21,240	140	260	500	140	260

5. Conclusion

In this paper, we proposed a method to synthesize speech utterances using the Multiple function parameters. We introduced two types of Multiple functions, MGMM and MBMM, and presented the update rules of their parameters. We conducted evaluation experiments under changing experimental conditions, such as the number of iterations or mixtures. We evaluated our approach by use of DP-based metric and MOS test from subjective and objective points of view, respectively. Through these experiments, we concluded that speech utterance is well-represented by the Multiple function with only a few parameters. When we consider the balance of the approximate precision and the number of parameters, MBMM is a more suitable model for the representation of speech utterances.

In the future, we plan to improve the iteration speed in estimating MBMM parameters, and devise more well-representing models. We are also planning to fit pitch contour or other phonemes to Multiple function and synthesize more general speech sounds, taking these points into consideration.

6. References

[1] Andrew J. Hunt and Alan W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," IEEE ICASSP, pp. 373-376, 1996.

[2] Gregory Beller, et al., "A hybrid concatenative synthesis system on the intersection of music and speech, " JIM, 2005.

[3] Remez, R. E., et al., "Talker identification based on phonetic information," Journal of Experimental Psychology: Human Perception and Performance, vol. 23, pp. 651-666, 1997.

[4] Remez, R. E., et al., "Speech perception without traditional speech cues," Science, pp.947–950, 1981.

[5] Xavier Rodet, "Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models," TFTS'97, 1998.

[6] T. Nakashika, et al., "Mathematical Modeling of Harmonic-Timbre Structure with Multi-Beta-Distribution," IEEE Workshop on Statistical Signal Processing, pp. 769-772, 2009.

[7] Yuan Ji, et al., "Applications of Beta-Mixture Models in Bioinformatics," Bioinformatics, vol.21, no.9, pp.2118-2122, 2005.

[8] J. Ramirez, et al., "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," Robust Speech Recognition and Understanding, pp. 1-22, 2007.

[9] X. Huang, et al., "Spoken Language Processing: A Guide to Theory, Algorithm and System Development," 2001.