Mel-LSP Parameterization for HMM-based Speech Synthesis

Naotoshi Nakatani, Kazumasa Yamamoto and Hiroshi Matsumoto

Faculty of Engineering, Shinshu University, Nagano, Japan
{naonaka, kyama, matsu}@sp.shinshu-u.ac.jp

Abstract

In HMM-based speech synthesis using mel-cepstral parameters, it has been observed that formant peaks tend to be flattened in the synthetic speech. To alleviate this problem this paper investigates Mel-LSP (line Spectral Pairs) based speech synthesis. First, using vowel spectra synthesized by four formants, it is shown that the formant flattening for the centroid of mel-LSP frequencies is mach less than that for mel-cepstra. After overviewing the closed form of Mel-LPC analysis, a structure of the Mel-LSP synthesis filter is presented. On the basis of this mel-LSP parameterization, the mora HMMs are trained using the mel-LSP parameters and short sentences are synthesized using them. The speech quality of these synthetic speech are compared with that of speech synthesized by the mel-cepstrum based HMMs. In A-B preference tests, Mel-LSP-based synthetic speech were chosen 61% of time over Mel-cepstrum based one.

1. Introduction

In recent years, various speech synthesis approaches based on Hidden Markov Models (HMMs) have been proposed. These approaches can be broadly classified into those that is based on concatenation of natural speech segments [1], [2], [3] and those that generate speech parameter sequence from HMMs themselves [4], [5]. The former approaches achieve a high quality in naturalness and intelligibility, but require the huge amount of storage capacity for speech data. The latter approaches do not require such a storage, but have not yet achieved the speech quality comparable to the former approaches.

In HMM-based speech parameter generation, the spectral peaks tend to be wider and less significant in the synthetic speech spectra. This "formant flattening" which results in a degradation in speech quality is caused by averaging feature vectors on the cepstral domain over the wider variety of contextual variations. Therefore, to alleviate this phenomenon it is effective to reduce coarticulation effects by context dependent HMMs and also to increase number of mixture Gaussians [6].

Another approach to reduce this phenomenon is the usage of formant or formant-like parameters as a feature vector. Although formant frequencies as a feature vector might completely avoid "formant flattening", automatic formant tracking is very difficult problem. An alternative parameter is LSP (Line Spectral Pair) frequency [7]. LSP frequencies have been widely utilized in speech coding since LSP parameters exhibit good interpolation properties and low distortion in quantization [8],[9]. In trainable speech synthesis based on HMMs, several studies have attempted the LSP parameterization [10], [11].

This paper investigates a LSP-based HMM speech synthesis. In particular, to realize auditory like frequency resolution, we use the warped LSP frequencies based on the warped LPC analysis (hereafter the term "warped" is referred to as "mel"). Unlike prevous works [10],[11], the mel-LSP trajectories are directly generated from HMMs with the static and dynamic mel-LSP parameters [4], and then speech is synthesized using a mel-LSP digital filter.

This paper is organized as follows. Section 2 quantitatively compares the formant flattening between the centroids of the mel-LSP and the mel-cepstral vectors. Section 3 first overview the closed form of the mel-LPC analysis, and then presents the mel-LSP synthesis filter. Section 4 evaluates the speech quality comparing with the mel-cepstrum based HMM speech synthesis. The final section presents concluding remarks and future works.

2. Spectral Comparison of LSP and Cepstral Centroids

First, this section compares the degree of formant flattening due to averaging LSP and cepstral parameters. For this purpose, synthetic vowel spectra composed of the four formant frequencies $\boldsymbol{x}^F = \{F_1, F_2, F_3, F_4\}$ are generated such that each F_i has the independent Gaussian distribution $N(\mu_i, \sigma^2)$ in which σ is set to the same value for the four formants. The formant band width B_i is determined as a function of F_i,

$$B_i = 50\{1 + F_i^2/(6 \times 10^6)\}. \quad (1)$$

The 8-th order prediction coefficients are directly derived by computing the four conjugate poles of the inverse filter from the four formants $\{F_i, B_i\}$ under the sampling frequency of 8kHz. Then, the 35-dimensional cepstral vector $\boldsymbol{x}^C$ and the 8-dimensional LSP vector $\boldsymbol{x}^L$ are computed from the prediction coefficients. For a given standard deviation σ, the centroid vectors $\bar{\boldsymbol{x}}^C$ and $\bar{\boldsymbol{x}}^L$ are calculated by averaging $\boldsymbol{x}^C$ and $\boldsymbol{x}^L$ over 100 tokens, respectively. As an example, the mean formant frequencies $\bar{\boldsymbol{x}}^F = \{\mu_1, \mu_2, \mu_3, \mu_4\}$ were set to $\{780, 1240, 2720, 3350\}$ (Hz) corresponding to the vowel /a/. Figure 1 compares three spectra corresponding to $\bar{\boldsymbol{x}}^F$, $\bar{\boldsymbol{x}}^L$ and $\bar{\boldsymbol{x}}^C$ for $\sigma = 200Hz$.

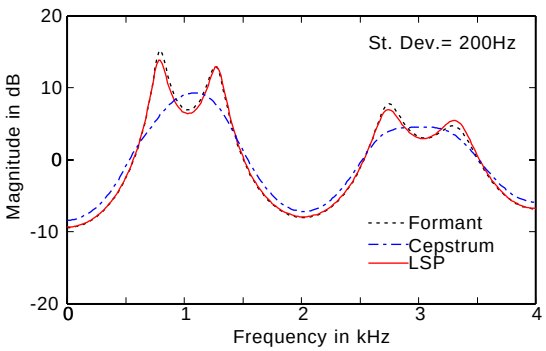


Figure 1: Comparison of the spectra corresponding to the centroids of formant frequencies, LSP and cepstral coefficients.

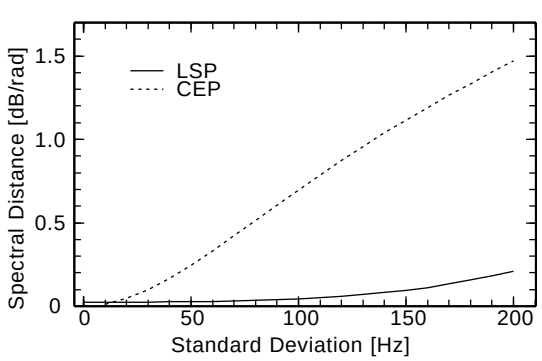


Figure 2: Spectral distances between centroid of formant vector and centroids of LSP and Cepstrum vectors as a function of standard deviation of formant frequencies.

It is clear that while the formant peaks on the spectrum of $\bar{\boldsymbol{x}}^C$ are sevearly flattened, the spectrum of the LSP centroid $\bar{\boldsymbol{x}}^L$ remains almost the same as that of the mean formant $\bar{\boldsymbol{x}}^F$. Figure 2 also shows the spectral distances (dB/radian) between the spectrum of $\bar{\boldsymbol{x}}^F$ and those of $\bar{\boldsymbol{x}}^C$ and $\bar{\boldsymbol{x}}^L$ as a function of the standard deviation σ. From this figure, the spectral distance between the spectra of $\boldsymbol{x}^F$ and $\boldsymbol{x}^L$ is much smaller than that between the spectra of $\boldsymbol{x}^F$ and $\boldsymbol{x}^C$. Although some of the LSPs in natural speech contribute to the glottal spectral shape, the formant flattening effect might be decreased by the use of LSP.

3. Mel-LSP based Speech Analysis/Synthesis

3.1. Mel-LPC analysis

This section overviews the Mel-LSP analysis. First, for frequency warped speech signal $\tilde{x}[n]$ ($n = 0, .., \infty$), which is bilinear transformed from a windowed input speech signal $x[n]$ ($n = 0, 1, ..N-1$) [12], the following all-pole model is defined [13].

$$\tilde{H}_\alpha(\tilde{z}) = \frac{\tilde{\sigma}_e}{\tilde{A}(\tilde{z})} = \frac{\tilde{\sigma}_e}{1 + \sum_{k=1}^{p} \tilde{a}_k \tilde{z}^{-k}} \tag{2}$$

$$\tilde{z}^{-1} = \frac{\tilde{z}^{-1} - \alpha}{1 - \alpha \tilde{z}^{-1}} \tag{3}$$

where $\tilde{a}_k$ is the k-th mel-prediction coefficient and $\tilde{\sigma}_e^2$ is the residual energy.

On the basis of minimum prediction error energy for $\tilde{x}[n]$ over the infinite time span, $\tilde{a}_k$ and $\tilde{\sigma}_e$ are given by Durbin's algorithm from the autocorrelation coefficients $\tilde{r}[m]$ of $\tilde{x}[n]$ defined by

$$\tilde{r}[m] = \sum_{i=0}^{\infty} \tilde{x}[n]\tilde{x}[n-m], \tag{4}$$

which is referred to as mel-autocorrelation function.

The mel-autocorrelation coefficients can be easily calculated from the input speech signal $x[n]$ via following two steps as shown in fig. 3 [14], [15]. First, the generalized autocorrelation coefficients are calculated as

$$\tilde{r}_\alpha[m] = \sum_{i=0}^{N-1} x[n] x_m[n] \tag{5}$$

where $x_m[n]$ is the output signal of an m-th order all-pass filter $\tilde{z}^{-m}$ excited by $x_0[n] = x[n]$. That is, $\tilde{r}_\alpha[m]$ is defined

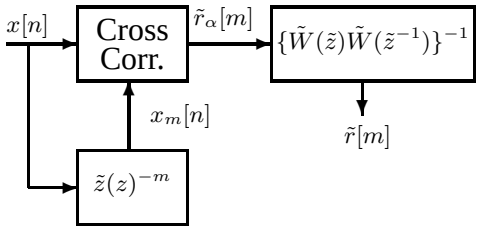


Figure 3: *Mel-autocorrelation function.*

by replacing the unit delay zi with the first order all-pass filter $\tilde{z}(z)^{-1}$ in the definition of conventional autocorrelation function. Due to the frequency warping, $\tilde{r}_\alpha[m]$ includes the frequency weighting $\tilde{W}(e^{j\tilde{\omega}})$ given by

$$\tilde{W}(\tilde{z}) = \frac{\sqrt{1-\alpha^2}}{1 + \alpha \tilde{z}^{-1}}. \tag{6}$$

Thus, the weighting is then removed by inverse filtering in the autocorrelation domain using $\{\tilde{W}(\tilde{z})\tilde{W}(\tilde{z}^{-1})\}^{-1}$. $\tilde{r}[m]$ (Mel-autocorrelation Coefficient) is obtained by

$$\tilde{r}[m] \quad = \quad \beta_0 \tilde{r}_\alpha[m] + \beta_1 \{\tilde{r}_\alpha[m-1] + \tilde{r}_\alpha[m+1]\} \tag{7}$$

where

$$\beta_0 \quad = \quad (1+\alpha^2)(1-\alpha^2)^{-\frac{1}{2}}, \tag{8}$$

$$\beta_1 \quad = \quad \alpha(1-\alpha^2)^{-\frac{1}{2}}. \tag{9}$$

As in the conventional LPC, we have a symmetric polynominal $\tilde{Q}(\tilde{z})$ and an anti-symmetric polynominal $\tilde{P}(\tilde{z})$ defined by

$$\tilde{P}(\tilde{z}) = \tilde{A}(\tilde{z}) + \tilde{z}^{-(p+q)} \tilde{A}(\tilde{z}^{-1}), \tag{10}$$

$$\tilde{Q}(\tilde{z}) = \tilde{A}(\tilde{z}) - \tilde{z}^{-(p+q)} \tilde{A}(\tilde{z}^{-1}). \tag{11}$$

Zeros of $\tilde{P}(\tilde{z})$ and $\tilde{Q}(\tilde{z})$ are on the unit circle, and then their frequencies $\tilde{\omega}$s except 0 and π are called the Mel-LSP frequencies. The even and odd numberd Mel-LSP frequencies corresponds to zeros of $\tilde{P}(\tilde{z})$ and $\tilde{Q}(\tilde{z})$, respectively, and these frequencies are interlaced with each other.

3.2. Mel-LSP synthesis filter

For a given set of Mel-LSP frequencies, $\{\tilde{\omega}_1, \tilde{x}_2, \cdots, \tilde{\omega}_p\}$ in increasing order, the synthesis filter $\tilde{A}^{-1}(\tilde{z}(z))$ is given by

$$\tilde{A}(\tilde{z}(z)) = \frac{1}{2}\{\tilde{P}(\tilde{z}) + \tilde{Q}(\tilde{z})\}. \tag{12}$$

A direct form filter is realized as in [13] by converting Mel-LSP frequencies into mel-prediction coefficients $\{\tilde{a}_i\}$. This paper presents another form of the synthesis filter by directly using Mel-LSP frequencies as in the conventional LSP-based filter with some modifications.

For the even order p, the polynominals are factored as

$$\tilde{P}(\tilde{z}) = (1 - \tilde{z}^{-1}) \prod_{i=2,4,\cdots,p} g_i \cdot \tilde{F}_i(\tilde{z}), \tag{13}$$

$$\tilde{Q}(\tilde{z}) = (1 + \tilde{z}^{-1}) \prod_{i=1,3,\cdots,(p-1)} g_i \cdot \tilde{F}_i(\tilde{z}) \tag{14}$$

where

$$g_i \cdot \tilde{F}_i(\tilde{z}) = 1 - 2\tilde{z}^{-1} cos\, \tilde{\omega}_i + \tilde{z}^{-2}. \tag{15}$$

In order to avoid lag-free loops in the filter, $\tilde{F}_i(\tilde{z})$ is transformed into equation 17.

$$\tilde{F}_i(z) = 1 + R(z)\, T_i(z), \tag{16}$$

$$R(z) = \alpha + \tilde{z}^{-1} = \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}, \tag{17}$$

$$T_i(z) = \frac{C_i + z^{-1}}{1-\alpha z^{-1}}. \tag{18}$$

In the above equations, g_i and C_i are defined by

$$g_i = 1 + 2\alpha\, cos\, \tilde{\omega}_i + \alpha^2, \tag{19}$$

$$C_i = -2(\alpha + cos\, \tilde{\omega}_i)/g_i. \tag{20}$$

By factoring out $R(z)$ in $F_i(z)$ and $1 \pm \tilde{z}^{-1}$, we have

$$\tilde{A}(z) = \frac{1}{G}\Big[1 + R(z)\Big\{ G_e \sum_{\substack{i=2\\ even}}^{p} T_i(z) \prod_{\substack{j=0\\ even}}^{i-2} \tilde{H}_i(z) - \frac{G_e}{1+\alpha} \prod_{\substack{i=2\\ even}}^{p} \tilde{H}_i(z) + G_o \sum_{\substack{i=1\\ odd}}^{p} T_i(z) \prod_{\substack{j=1\\ odd}}^{i-2} \tilde{H}_i(z) + \frac{G_o}{1-\alpha} \prod_{\substack{i=1\\ odd}}^{p} \tilde{H}_i(z) \Big\}\Big] \tag{21}$$

where

$$G = \frac{2}{g_e + g_o}, \tag{22}$$

$$G_e = \frac{g_e}{g_e + g_o}, \tag{23}$$

$$G_o = \frac{g_o}{g_e + g_o}, \tag{24}$$

$$g_e = (1+\alpha) \prod_{\substack{i=2\\ even}}^{p} g_i \quad and \quad g_o = (1-\alpha) \prod_{\substack{i=1\\ odd}}^{p} g_i. \tag{25}$$

For the odd order p, the polynominals are factored as

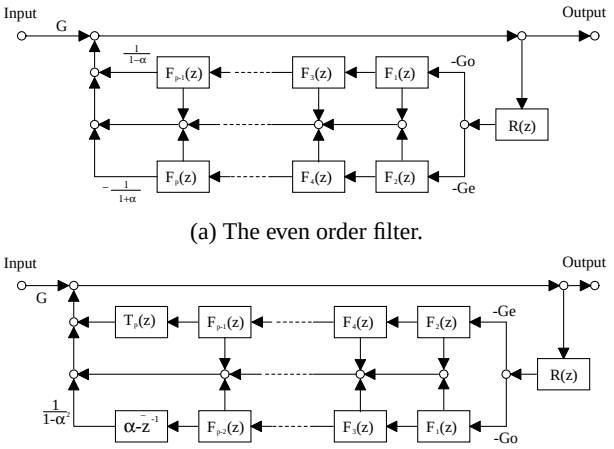
$$\tilde{P}(\tilde{z}) = (1-\tilde{z}^{-2}) \prod_{i=2,4,\cdots,p-1} g_i \cdot \tilde{F}_i(\tilde{z}), \tag{26}$$

$$\tilde{Q}(\tilde{z}) = \prod_{i=1,3,\cdots,p} g_i \cdot \tilde{F}_i(\tilde{z}). \tag{27}$$

Then, the following form of inverse filter is obtained.

$$\tilde{A}(z) = \frac{1}{G}\Big[1 + R(z)\Big\{ G_o \sum_{\substack{i=1\\ odd}}^{p} T_i(z) \prod_{\substack{j=1\\ odd}}^{i-2} \tilde{H}_i(z) + G_e \sum_{\substack{i=2\\ even}}^{p-1} T_i(z) \prod_{\substack{j=0\\ even}}^{i-2} \tilde{H}_i(z) - \frac{G_e}{1-\alpha^2}(\alpha - \tilde{z}^{-1}) \prod_{\substack{i=2\\ even}}^{p-1} \tilde{H}_i(z) \Big\}\Big]. \tag{28}$$

From equations 21 and 28 the filter configurations for the even and odd order are shown in figures 4 and 5.



(a) The even order filter.

(a) The odd order filter.

Figure 4: The Mel-LSP speech synthesis filter.

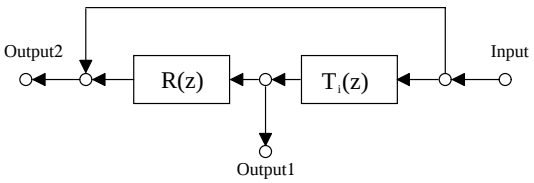


Figure 5: The i-th stage of the Mel-LSP synthesis filter.

4. EVALUATION

4.1. Experimental conditions

In training HMMs, we used 150 sentences uttered by each of 103 male speakers, which are from database of 50 phonetically balanced sentences (ASJ-PB) and 100 newspaper article texts (ASJ-JNAS). The sub-word models are 124 gender-dependent monosyllable HMMs. The structure of HMMs is a left-to-right model with 3 emitting states for vowels, double consonant(/q/), syllabic nasal(/N/) and silences, and with 5 emitting states for the other syllables. A state consists of a single Gaussian with a diagonal covariance.

The speech data was sampled at 16kHz. A speech segment of 25ms was weighted by Hamming window without preemphasis. A 12th order Mel-LPC analysis was conducted with a frame shift of 10ms. The frequecy warping factor α was set to 0.45. In Mel-LSP based synthesis, a feature vector composed of log-residual energy, 12 Mel-LSP frequencies, delta-log-residual energy, and 12 delta-Mel-LSP frequencies. In Mel-cepstrum based synthesis, the cepstral parameters are derived from the Mel-prediction coefficients instead of the mel-cepstral analysis in [**?**], and a feature vector composed of 15 mel-cepstral and 15 delta-mel-cepstral coefficients including the their 0th terms.

In this study, to compare the effect of synthesized spectral sequences on the speech quality, the state duration and excitation signal were used from natural speech. The state duration of each HMM is determined by Viterbi alignment between the feature vector sequence of the template speech and the connected HMMs corresponding to a given text. The excitation signal was obtained by inverse filtering the speech using the 12-th order Mel-prediction coefficients. In Mel-cepstrum based speech synthesis, speech was synthesized by means of the MLSA (Mel Log Spectrum Approximation) filter [16].

4.2. Subjective and spectral evaluation

Three phrases /arayuru geNjituwo/, /subete jibuNno hoohe/, and /negimageta noda/ were synthesized by both Mel-LSP and Mel-

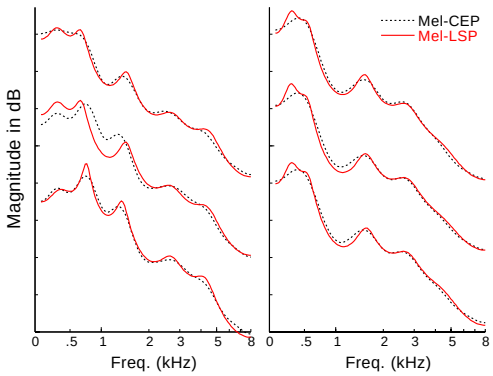


Figure 6: Comparison of Spectra synthesized by Mel-LSP and Mel-LPC cepstral based HMMs

cepsrum based HMM. Figure 6 shows an example of two spectral sequences generated by both the Mel-LSP and the Mel-cepstrum based HMMs. It is noted that the formant peaks of Mel-LSP based spectra are clear as compared to those of the Mel-cepstrum based spectra. The speech quality of these sentences were evaluated in A-B preference tests. Ten subjects participated in the listening test to compare preference for Mel-LSP based synthetic speech versus Mel-cepstrum based one. The percentage of trials for three sentences was 61% in which the Mel-LSP based synthetic speech is preferred over each of Mel-cepstrum based one.

5. Conclusions

In this paper we have presented the Mel-LSP based speech analysi/synthesis for HMM-based speech synthesis. The expeimental results have shown that the Mel-LSP based HMMs slightly improve the synthetic speech quality over the Mel-cepstrum based HMMs. However, the formant bandwidth of the Mel-LSP based HMMs is still wider than that of the natural speech. This seems to be caused by too simple HMM structure to deal with the contxtual as well as the speaker variation. Therefore, further investigation will be conducted using context dependent HMMs with many mixtures.

6. References

[1] Donovan, R.E., Woodland, P.C., "Automatic speech synthesiser parameter estimation using HMMs," Proc. of ICASSP-95, pp.640-643, 1995.

[2] Huang, X., Acero, A., et al., "Recent improvements on Microsoft's trainable text-to-speech synthesis -Whistler," Proc. of ICASSP-97, pp.959-962, 1997.

[3] Donovan, R.E. and Eide, E.M., "The IBM trainable speech synthesis system," Proc. of ICSLP-98, pp.1703-1706, 1998.

[4] Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., and Imai, S., "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," Proc. of EUROSPEECH, pp.757-760, 1995.

[5] Masuko, T., Tokuda, K., T., Kobayashi, T., and Imai, S., "Speech Synthesis using HMMs with dynamic features," Proc. of ICASSP96, Vol.1, pp.389-392, 1996.

[6] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., "Speech parameter generation algorithms for hmm-based speech synthesis," Proc. of ICASSP-00, vol.3, pp.1315-1318, 2000.

[7] Itakura, F., "Line spectrum representation of linear predictive coefficients of speech signals," J. Acoust. Soc. Am., 57, 535(A), 1975.

[8] Sugamura, N., and Itakura, F., "Speech data compression by LSP speech analysis-synthesis technique," Trans. IECE, Vol.J 64-A, No.8, pp.599-606, 1981.

[9] Soong F.K., and Juang B.-H., "Line spectrum pair (LSP) and speech data compression," Proc. of ICASSP-84, Vol.1, pp.1.10.1-4, 1984.

[10] Pellom, B.L., and Hansen, J.H.L., "Trainable speech synthesis based on trajectory modeling of line spectrum pair frequencies," IEEE Nordic Signal Processing Symposium, pp.125-128, Vigso, Denmark, June 1998.

[11] Dines J., and Sridharan, S., "Trainable speech synthesis with trended hidden Markov models," Proc. of ICASSP-01, Speech, p8.6, 2001.

[12] Oppenheim,A.V., and Johnson,D.H., "Discrete representation of signals," IEEE Proceedings, vol.60, no.6, pp.681-691, 1972.

[13] Strube,H.W., "Linear prediction on a warped frequency scale," J. Acoust. Soc. Amer., vol.68, no.4, pp.1071-1076, 1980.

[14] Matsumoto,H., Nakatoh, Y., and Furuhata, Y., "An efficient Mel-LPC analysis method for speech recognition," Proc. of ICSLP'98, pp.1051-1054, 1998.

[15] Matsumoto,H., and Moroto,M., "Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition," Proc. of ICASSP-01, pp.117-120, 2001.

[16] Imai, S., Sumita, K., and Furuichi, C., "Mel log spectrum approximation (MLSA) filter for speech synthesis," Trans. IEICE, vol.J66-A, 122-129, 1983.