Histogram-based spectral equalization for HMM-based speech synthesis using mel-LSP

Yamato Ohtani, Masatsune Tamura, Masahiro Morita, Takehiko Kagoshima, Masami Akamine

Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation, Japan

yamato.ohtani@toshiba.co.jp, masatsune.tamura@toshiba.co.jp, masahiro.morita@toshiba.co.jp, takehiko.kagoshima@toshiba.co.jp, masa.akamine@toshiba.co.jp

Abstract

This paper describes a statistical spectral parameter emphasis technique for HMM-based speech synthesis using mel-scaled line spectral pair (mel-LSP). Spectral parameter emphasis is effective for compensating over-smoothed spectra in HMM-based speech synthesis. However, there is no conventional technique that satisfies such requirements as automatic tuning for different speakers and real-time synthesis for mel-LSP. In the proposed method, the cumulative distribution function (CDF) is calculated from the histogram of spectral parameters that are extracted from training speech data. In the same manner, CDF of spectral parameters that are generated from HMMs is constructed. Then an emphasis rule is trained so that the CDF of generated parameters equals to that of training data. After generating a spectral parameter sequence from HMMs, the spectral parameter sequence is emphasized by using the rule. Experimental results show that our proposed method improves speech quality.

Index Terms: speech synthesis, hidden Markov model, parameter emphasis, mel-LSP, histogram equalization

1. Introduction

Recently, statistical speech synthesis based on hidden Markov model (HMM) has been studied widely [1][2]. In this approach, spectral and source features in a source-filter model of speech production, and durations are simultaneously modeled by HMMs in training. In synthesis, these parameter sequences are generated from the HMMs based on the maximum likelihood (ML) criterion. One of the advantages of HMM-based speech synthesis over concatenative speech synthesis is flexibility in producing different voices, which is achieved by using speaker adaptation techniques [3]. However, compared to the concatenative speech synthesis, synthetic speech by this approach does not sound as good owing to over-smoothed spectra.

To address this problem, Yoshimura et al. proposed a postfilter for mel-cepstrum [4], and Ling et al. proposed a method to emphasize over-smoothed spectra for line spectral pair (LSP) [5]. These techniques enhance spectral features in post-processing using pre-determined enhancement parameters. They are simple and effective. However, it is not straightforward to determine proper enhancement parameters for different target speakers. Since there is no automatic way to determine the parameters, a time-consuming manual method must be used. As another approach, a stochastic emphasis method has been proposed. This is to generate acoustic parameters considering global variance (GV) [6]. This algorithm effectively alleviates over-smoothing effects in the case of using mel-cepstrum as the spectral parameter. For LSP parameters, parameter generation algorithms with GV were proposed in [7][8]. However, the parameter generation with GV cannot be carried out in a recursive way.

In this paper, we propose a spectral emphasis method based on histogram equalization (HE). HE is widely used in image processing to increase the contrast of images. Wu et al. applied HE to F0 trajectories in speech synthesis to reduce over-smoothing effects [9]. We apply HE to spectrum parameters. In the proposed method, cumulative distribution functions (CDFs) are calculated for LSP parameters in training data and for those generated from HMMs in each dimension in advance. Here, LSP parameters from HMMs are generated based on context features in training data so as to keep the contexts for the LSP generation and the training data consistent. Then, the emphasis rule is constructed so that the generated LSPs have the same CDFs as those of the LSPs in the training data. In the emphasis process, each dimension of the generated LSP parameter sequence is converted based on the emphasis rule. After this conversion, LSP parameters are rearranged to resolve disorder and reduce over-emphasizing effects.

The remainder of this paper is organized as follows. Section 2 describes the HMM-based speech synthesis system and the conventional parameter emphasis method for the HMM-based speech synthesis system. In section 3, the proposed method is described. In section 4, we present an experimental evaluation result. Finally, we conclude this paper in section 5.

2. HMM-based speech synthesis system

2.1. Overview of HMM-based speech synthesis

HMM-based speech synthesis includes two main stages, the training stage and the synthesis stage. In train-
2.2. Spectral parameter emphasis using postfiltering

In HMM-based speech synthesis systems, muffled speech is often synthesized from generated over-smoothed spectral parameter sequences. Figure 1 shows histograms both the mel-LSPs in the training data and generated mel-LSPs. Compared with the mel-LSPs in the training data, generated mel-LSPs have smaller variances due to over-smoothing. To alleviate the over-smoothing effects, many HMM-based speech synthesis systems employ an ML-based parameter generation considering GV [6]. This algorithm requires parameter generation for a whole utterance and is not applicable to parameter generation with any shorter unit with fixed length. Also, according to [12], the traditional GV-based parameter generation cannot improve the over-smoothing in the case of using mel-LSPs as spectral features.

To alleviate over-smoothing of mel-LSP parameters, the speech synthesis system used in this paper employs a postfilter like [5]. The closeness between adjacent mel-LSP parameters represents the strength of spectral peak. The postfilter moves respective mel-LSP parameter pairs tend to be made closer, and then the spectral envelope is enhanced. However, we need to determine a proper enhancement parameter experimentally for each speaker. Moreover, there is no assurance that a distribution of emphasized parameters postfiltering moves closer to that of parameters extracted from actual speech data.

3. Spectral parameter emphasis based on histogram equalization

Conventional parameter emphasis methods using the postfilter require us to optimize a set of enhancement parameters manually, which requires great efforts. In this section, we introduce histogram equalization (HE) in order to achieve automatically spectral emphasis.

Figure 2 shows the process of the proposed method. In training, we construct a cumulative distribution function (CDF) of spectral parameters in the training database. In the same way, we also construct a CDF of spectral parameters generated from HMM using the context-features of training sentences. Then we train an emphasis rule by extracting the relationship between their CDFs. In emphasis, we apply the emphasis rule to a generated spectral parameter sequence and then obtain an emphasized parameter sequence. These details describes each in the following sections.

3.1. Training of the emphasis rule

First we find the maximum $y_{\text{max}}^{(d)}$ and minimum $y_{\text{min}}^{(d)}$ of the $d^{th}$ dimension of spectral parameters among training data $y_n = \left( y_n^{(1)}, y_n^{(2)}, \ldots, y_n^{(D)} \right)^T$, where $T$ denotes transposition of the vector. Then, we construct a histogram of the $d^{th}$ dimension of spectral parameters $h_y^{(d)}(i)$ of training data by dividing the range $[y_{\text{min}}^{(d)}, y_{\text{max}}^{(d)}]$ into $I$ equal width parts. We calculate a CDF of the $d^{th}$ dimension of
spectral parameters as follows:
\[
f_y^{(i)} = \sum_{j=1}^{N} h_y^{(j)} N, \tag{1}
\]
where \(N\) denotes the total number of frames in the training data. In the same manner, the maximum \(x_{\text{max}}^{(d)}\) and minimum \(x_{\text{min}}^{(d)}\) of \(d\)th dimension are extracted from spectral parameters generated from HMMs \(x_m = [x_m^{(1)}, x_m^{(2)}, \ldots, x_m^{(D)}]^T\) and a CDF of generated spectral parameters is calculated.

Second, a look-up table is constructed as an emphasis rule. We look for the \(i\)th bin of \(f_y^{(i)}(i)\) that satisfies \(f_y^{(i)}(i) \leq p_k < f_y^{(i)}(i+1)\), where \(p_k\) represents the \(k\)th cumulative probability \((0 < p_k < p_{k+1} \leq 1)\). Then the value \(y_{pk}^{(d)}\) corresponding to \(p_k\) is estimated as follows:
\[
y_{pk}^{(d)} = \frac{p_k (y_{i+1}^{(d)} - y_i^{(d)}) - f_y^{(i)}(i) y_{i+1}^{(d)} + f_y^{(i+1)}(i+1) y_i^{(d)}}{f_y^{(i+1)}(i+1) - f_y^{(i)}(i)}.
\tag{2}
\]
We also find the \(j\)th bin of \(f_x^{(j)}(j)\) that satisfies \(f_x^{(j)}(j) \leq p_k < f_x^{(j+1)}(j+1)\) and calculate the value \(\pi_{pk}^{(d)}\) corresponding to \(p_k\) as follows:
\[
\pi_{pk}^{(d)} = \frac{p_k (x_{j+1}^{(d)} - x_j^{(d)}) - f_x^{(j)}(j) x_{j+1}^{(d)} + f_x^{(j+1)}(j+1) x_j^{(d)}}{f_x^{(j+1)}(j+1) - f_x^{(j)}(j)}.
\tag{3}
\]
Finally, we construct a look-up table for the \(d\)th dimension of the LSP parameter
\[
T(d) = \left[\frac{\pi_{pk}^{(d)}}{y_{pk}^{(d)}}, \frac{\pi_{pk+1}^{(d)}}{y_{pk+1}^{(d)}}, \ldots, \frac{\pi_{pk}^{(d)}}{y_{pk}^{(d)}}, \ldots, \frac{\pi_{pk+1}^{(d)}}{y_{pk+1}^{(d)}}\right],
\tag{4}
\]
where,
\[
\frac{\pi_{pk}^{(d)}}{y_{pk}^{(d)}} = \frac{x_{\text{min}}^{(d)}}{y_{\text{min}}^{(d)}}, \frac{\pi_{pk}^{(d)}}{y_{pk}^{(d)}} = \frac{x_{\text{max}}^{(d)}}{y_{\text{max}}^{(d)}}.
\tag{5}
\]

### 3.2. Spectral parameter emphasis

In the emphasis process, we convert a generated spectral parameter to an emphasized parameter dimension by dimension. An index \(k\) of \(T(d)\) corresponding to an arbitrarily generated spectral parameter \(x_t^{(d)}\) is selected by satisfying \(\pi_{pk}^{(d)} \leq x_t < \pi_{pk+1}^{(d)}\). Then the emphasized spectral parameter \(y_t^{(d)}\) is determined with \(\pi_{pk}^{(d)}, \pi_{pk+1}^{(d)}, y_{pk}^{(d)}\) and \(y_{pk+1}^{(d)}\) as follows:
\[
y_t^{(d)} = \frac{y_{pk+1}^{(d)} - y_{pk}^{(d)}}{\pi_{pk+1}^{(d)} - \pi_{pk}^{(d)}} (x_t - \pi_{pk}^{(d)}) + y_{pk}^{(d)}.
\tag{6}
\]

![Figure 3: Cumulative frequency functions of the 20th mel-LSP coefficient.](image)

Note that (6) means a linear interpolation because \(\pi_{pk}^{(d)}\) is always smaller than \(x_t^{(d)}\). This paper uses mel-LSP parameters as spectral parameters. In this case, disorder and over-emphasis are sometimes caused by the above emphasis because each dimension is converted independently. We adjust intervals between adjacent mel-LSP coefficients based on frequency delta information and rearrange them so as to be ordered.

Figure 3 shows CDFs of the 20th mel-LSP coefficient. The CDFs of the mel-LSP generated from the training context features (Figure 3 (a)) and from the open data (Figure 3 (b)) are different from that of the training mel-LSP. We can see this figure that the CDF of the emphasized mel-LSP gets closer to that of the training mel-LSP. Figure 4 shows a spectrum without emphasis and an emphasized spectrum using the proposed method. The emphasized spectral envelope includes larger peaks and valleys than the non-emphasized spectral envelope.

### 4. Experimental evaluation

#### 4.1. Experimental conditions

We conducted a subjective evaluation of the proposed method and compared to those of other emphasis methods. In this evaluation, we trained a HMM for a female voice using its speech corpus consisted of 2237 utterances. The sampling rate of the speech data was 22.05 [kHz]. We used 40-dimensional mel-LSPs including the
power and their deltas. BAP consisted of 5-dimensional components (0–1.4, 1.4–3.1, 3.1–5.5, 5.5–8.3 and 8.3–11 [kHz]) and their deltas. F0 parameter vectors were composed of logarithmic F0 and its delta and delta-delta. The number of bins of the look-up table is 10000.

In the evaluation, we employed four types of synthesized speech: i) non-emphasized speech (BASELINE), ii) emphasized speech using the postfilter (CONVENTIONAL), iii) emphasized speech using the proposed method (PROPOSED) and iv) speech using a GV-based generation algorithm (GV) [6]. In addition, the enhancement parameter of the postfilter used in “CONVENTIONAL” was optimized in the preliminary experiment.

We conducted an opinion test on speech quality using our crowd-sourcing system. In the opinion test, each listener evaluated synthesized speech using a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor and 1: bad). The number of listeners was 18, and each listener evaluated 80 samples that were not included in the training database.

4.2. Result of subjective evaluation

Figure 5 shows the result of the subjective evaluation of speech quality. We can see that speech quality of “PROPOSED” outperforms that of “BASELINE.” “PROPOSED” also has a better improvement compared with “GV.” Therefore, the proposed method can emphasize spectral parameters efficiently. Speech quality of “PROPOSED” is almost the same as that of “CONVENTIONAL.” However, “PROPOSED” has an advantage in terms of parameter adjustment over “CONVENTIONAL” because “CONVENTIONAL” needs manual adjustment of the enhancement parameter and it is difficult to determine a proper parameter.

This result suggests that the proposed method can emphasize generated spectral parameters efficiently and improve speech quality.

5. Conclusions

In this paper, we have proposed an equalization method for spectral parameters generated from HMMs. In the proposed method, the emphasis rule is trained by relating the cumulative distribution function of spectral parameters generated from HMMs to that of those in the training data. Then arbitrary generated spectral parameter sequences are emphasized based on the rule. Subjective evaluation results show that our proposed method is effective emphasize spectrum and improve speech quality.

6. References