INTERSPEECH 2009 BRIGHTON

Pulse Density Representation of Spectrum for Statistical Speech Processing

Yoshinori Shiga

Spoken Language Communication Group, MASTAR Project
National Institute of Information and Communications Technology (NICT), Japan
yoshi.shiga@nict.go.jp

10.21437/Interspeech.2009-146

Abstract

This study investigates a new spectral representation that is suitable for statistical parametric speech synthesis. Statistical speech processing involves spectral averaging in the training process; however, averaging spectra in the domain of conventional speech parameters over-smooths the resulting means, which degrades the quality of the speech synthesised. In the proposed representation, high-energy parts of the spectrum, such as sections of dominant formants, are represented by a group of high-density pulses in the frequency domain. These pulses' locations (i.e., frequencies) are then parameterised. The representation is theoretically capable of averaging spectra with less over-smoothing effect. The experimental results provide the optimal values of factors necessary for the encoding and decoding of the proposed representation towards the future applications of speech synthesis.

Index Terms: spectral analysis, cepstral analysis, feature extraction, speech synthesis

1. Introduction

In the field of speech synthesis, statistical approaches that are based on the hidden Markov model (HMM) are currently receiving a considerable deal of attention [1][2][3][4]. These approaches are not only capable of acquiring the characteristics of speech automatically from a speech database, but also capable of producing speech with various voice characteristics and speaking styles through techniques such as speech morphing and speaker adaptation. Such flexibility provides a great advantage over the other leading speech synthesis techniques.

Recently, however, a limitation of speech synthesis using statistical techniques has been pointed out; the statistical process causes excessively-smoothed spectra, which degrades the quality of synthetic speech [5][3]. We consider that the major cause of this problem is inadequate spectral averaging for conventional speech representations. Being tractable with a statistical framework, the cepstrum is most commonly used in the current statistical speech processing. However, averaging several spectra whose formants are located in slightly different frequencies easily causes the formants to be 'dull-edged'. Figure 1 illustrates this effect, where two spectra, designed using all-pole modelling, are averaged in the cepstral domain. The resulting mean spectrum (thick solid line) has dulled formants and differs considerably from the one where the locations of the corresponding poles are properly averaged (dashed line).

There are some studies that adopt the line spectrum pair (LSP) as a speech representation for the HMM-based speech synthesis [2]. In the LSP representation, formants are averaged properly in the majority of cases. LSP is, however, based on the all-pole model, which is theoretically poor for representing spectral details other than poles (formants). Wouters and Macon [6] improve the quality of speech synthesised from the all-pole model by compensating the error between the sinusoidal spectrum and its all-pole fit. In addition, it is well-known that the inherent order of the parameters does not always correspond between the frames to be processed [7]. Furthermore, it is reported that substantial spectral manipulation (e.g., for speaker adaptation) occasionally produces LSPs that do not maintain the original ordering of the parameters [8][4], which leads to severe degradation in the quality of speech synthesised.

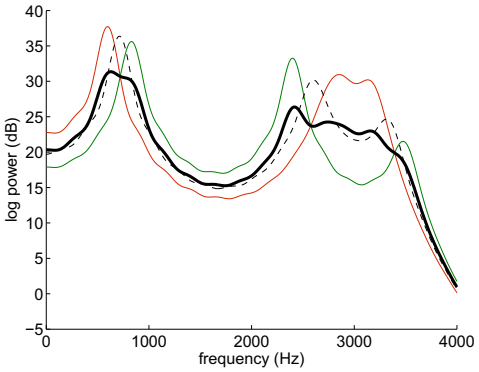


Figure 1: Spectral averaging in the cepstral domain. Log-power spectra (thin solid lines) to be averaged, the resulting mean (thick solid line), and an ideal interpolation (dashed line) that has poles at the mean positions on the z-plane.

We have investigated a spectral representation that is tractable within the statistical speech processing, as in the case of the cepstrum, and is capable of interpolating speech formants properly, as in the case of LSP. In the previous paper [9], we theoretically examined the proposed representation, reporting some results from preliminary experiments. In this study, preparatory to the applications of statistical speech processing, we investigate the accuracy of this representation against the various factors involved in the encoding and decoding of the representation. The remaining part of this paper is organised as follows: a concise explanation of the proposed representation is presented in Section 2, the results of experiments are presented and discussed in Section 3 and the conclusion is provided in Section 4.

2. Pulse density representation of spectrum

2.1. Frequency-domain pulse-density modulation

Let us consider representing high-power sections of the log-power spectrum, such as sections of dominant formants, with a group of high-density line spectra. As illustrated schemati-

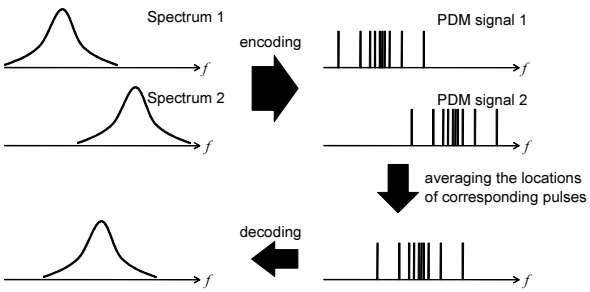


Figure 2: Spectrum interpolation based on the proposed line spectrum representation

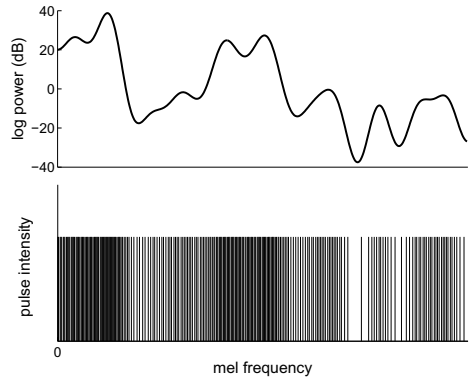


Figure 3: PDM of an actual log-power spectrum ($N = 1024$). The log power values of the spectrum are represented by the relative density of pulses.

cally in Fig. 2, such a line spectrum representation should (similar to the case of LSP) achieve the desired spectral averaging by interpolating formants on the basis of the frequency of each line spectrum. In contrast to LSP, however, this representation should (with a sufficient number of line spectra) be able to preserve the spectral detail that LSP fails to represent.

Such a representation can be realised by applying pulse-density modulation (PDM) in the frequency domain. It should be noted that instead of the term 'line spectrum', we hereafter use the term 'pulse' of digital modulation. PDM is a form of modulation that encodes the amplitude of a signal into the relative density of pulses and is often achieved by the delta-sigma modulator (DSM) [10]. The DSM used in this study is the standard one-bit DSM [9], which includes a one-bit quantizer generating either a '$+1$' or '-1' depending on the amplitude of the input signal. For theoretical details, see [9].

The input to DSM is a log-power spectrum envelope that contains no harmonic component. For efficient modulation, the DC component of the envelope is removed beforehand. Figure 3 shows an example of PDM of a log-power spectrum from actual speech, where the pulses were generated at frequencies when DSM outputs $+1$'s. The upper and lower graphs correspond to the input and output of DSM, respectively.

For the purpose of achieving the intended spectral interpolation, as shown in Fig. 2, our spectral representation retains the locations of output pulses (i.e., frequencies at which DSM generates '$+1$'). The representation consists of $N/2$ frequencies for a given N-point log-power spectrum since DSM outputs exactly equal number of positive and negative pulses, as long as the input has no DC component. The representation will hereafter be referred to as the *pulse-density-modulation spectrum* (PDM spectrum).

Frequency-domain global variations (e.g., the spectral tilt), which have an adverse effect on the estimation of the PDM spectrum [9], are suppressed by 'high-pass liftering' prior to the input of the spectrum to DSM. The input, $\hat{X}_{\rm in}(\Omega)$, is given as

$$\hat{X}_{\rm in}(\Omega) \approx 2 \sum_{n=1}^{p} L(n)\, c(n) \cos(n\Omega) \tag{1}$$

where $c(n)$ and $L(n)$ denote the $n^{\rm th}$ cepstral coefficient and the lifter, respectively, and p is the order of cepstrum. For the normalised frequency Ω, we may use the linear or mel frequency scale. The following lifter is adopted for the experiments in the next section:

$$L(n) = \begin{cases} n/\gamma, & 1 \le n < \gamma \\ 1, & n \ge \gamma \end{cases} \tag{2}$$

where γ satisfies $1 \le \gamma \le p$. This lifter suppresses the low-quefrency energy as γ increases, thereby reducing the spectral global variation.

2.2. PDM cepstrum

The delta-sigma modulation relies on the technique of oversampling to reduce the quantization noise. Accordingly, the input log-power spectrum should be represented with a sufficiently large number of data points. This means that the PDM spectrum itself consists of a large number of dimensions; it is therefore necessary to reduce the dimensionality for the use of the proposed representation in typical applications of statistical speech processing.

We achieve this reduction on the basis of a sine series expansion. The PDM spectrum, $\Omega_{\rm PF}(m)$ ($m = 0, 1, 2, \cdots, N/2$), can be expanded as

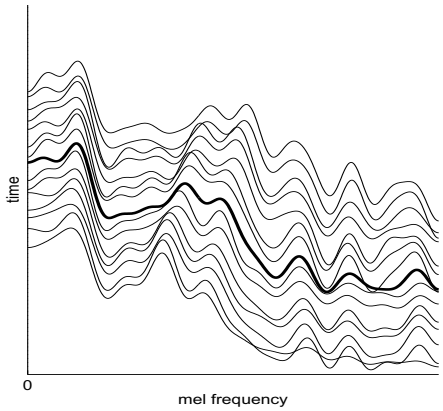
$$\Omega_{\rm PF}(m) = \frac{2\pi m}{N} + 2 \sum_{n=1}^{q} c_{\rm PF}(n) \sin \frac{2\pi m n}{N}. \tag{3}$$

It is to be noted that $c_{\rm PF}(n)$ can be efficiently determined by using the inverse-FFT algorithm. Most of the information on $\Omega_{\rm PF}(m)$ is considered to be preserved by the terms of lower orders. Truncating the expansion at an appropriate order can thus provide a good approximation.

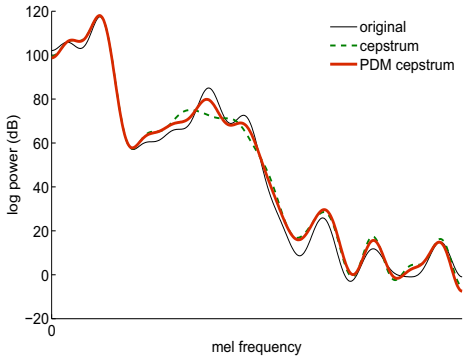
In order to reconstruct the log-power spectrum during decoding, the $0^{\rm th}$ cepstral coefficient of the original log-power spectrum is practically stored in $c_{\rm PF}(0)$. The set of coefficient $c_{\rm PF}(n)$ will hereafter be referred to as *PDM cepstrum*.

2.3. PDM cepstrum to log-power spectrum: decoding

For applications such as parametric speech synthesis, it is vital to reconstruct the log-power spectrum. The decoding process of widely used time-domain PDM is very simple; a PDM pulse train is converted back through a low-pass filter. For our frequency-domain PDM, a *low-pass lifter* is accordingly used to decode the PDM spectrum. The procedure is as follows: (i) convert the PDM cepstrum $c_{\rm PF}(n)$ into the PDM spectrum $\Omega_{\rm PF}(m)$ using (3); (ii) reconstruct an N-point PDM pulse train by producing '$+1$' at discrete frequencies given by $\Omega_{\rm PF}(m)$, and '-1' at all the other frequencies; (iii) compute the inverse discrete Fourier transform of the reconstructed PDM pulse



(a) Spectra to be averaged (5-ms frame-shift)



(b) Mean log-power spectra

Figure 4: Comparison of spectral averaging in the cepstrum domain and PDM-cepstrum domain. For specifics of the experimental conditions, see [9].

train; (iv) apply the inverse lifter $1/L(n)$ to the result obtained from step iii in order to recover the global variations suppressed above; and (v) perform a discrete Fourier transformation on the above liftered signal for reconstructing a log-power spectrum. According to the computation given in steps iv and v above, the reconstructed log-power spectrum, $\hat{X}_{\rm rec}(\Omega)$, is now given as

$$\hat{X}_{\rm rec}(\Omega) = c_{\rm PF}(0) + 2\,v_c \sum_{n=1}^{p} \frac{c_{\rm PD}(n)}{L(n)} \cos(n\Omega) \qquad (4)$$

where v_c and $c_{\rm PD}(n)$ are, respectively, the feedback gain of DSM and the $n^{\rm th}$ coefficient of the cepstrum that was computed in step iii. The truncation of cepstrum $c_{\rm PD}(n)$ at order p in (4) acts as the low-pass lifter of the PDM pulse train.

Figure 4 shows an example comparing spectral averaging in the standard-cepstrum domain and PDM-cepstrum domain from our preliminary experiments. Log-power spectra of consecutive frames extracted from speech with /r/-to-/I/ transition of English (Fig. 4a) are averaged in each domain. Such a type of averaging occurs during the training of HMMs, where spectra are averaged for frames that belong to the same state. As shown in Fig. 4b, the second and third formants are clearly shown in the case of the mean resulting from averaging in the PDM-cepstrum domain, whereas the standard cepstral averaging flattened out those formants.

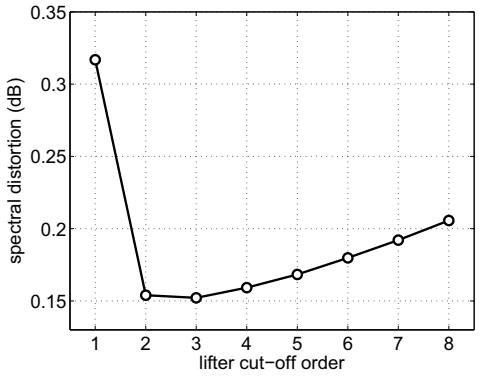


Figure 5: Spectral distortion versus cut-off order of the high-pass lifter

3. Experiments

The encoding and decoding of the PDM cepstrum distort the original spectrum. The quantization noise out of DSM is one of the causes. Here, we evaluate the accuracy of the representation with respect to values of factors involved.

3.1. Method and data

The accuracy was measured by the RMS value of spectral distortion across all the frames included in the data set. The spectral distortion was measured by the cepstrum distance:

$$d = \frac{10}{\ln 10}\sqrt{2\sum_{n=1}^{p}\left(c(n) - v_c\frac{c_{\rm PD}(n)}{L(n)}\right)^2} \quad ({\rm dB}) \qquad (5)$$

where $c(n)$ and $c_{\rm PD}(n)$ denote the $n^{\rm th}$ coefficients of the original cepstrum and the cepstrum computed during the procedure in Section 2.3, respectively. We investigated the distortion against the factors used in the conversion between the spectrum and the PDM cepstrum, which are as follows: (a) the cut-off order γ of the high-pass lifter $L(n)$ in (2), (b) the number of points of spectrum input to DSM, (c) the DSM feedback gain v_c and (d) the order of PDM cepstrum.

The speech data used here is of 100 English utterances by a female speaker (SLT) from the CMU ARCTIC database [11]. Spectral envelopes were estimated from their waveforms (16-kHz sampling) by the STRAIGHT analysis [12] using 5-ms frame-shifts. Each of the envelopes was then converted into the $39^{\rm th}$-order mel-cepstrum [4] (which corresponds to $c(n)$ in (5)). The PDM cepstrum was finally computed from the spectrum obtained from the mel-cepstrum on a frame-by-frame basis.

3.2. Results and discussions

Figure 5 shows the relationship between the distortion and the cut-off order γ of the high-pass lifter $L(n)$ under the following conditions: $N = 2^{12}$, $v_c = 20.0$ dB and $q = 63$. The distortion became minimal (0.1521 dB) when $\gamma = 3.0$. It should be noted, however, that the distortion values are sufficiently small across all the orders. We therefore believe that γ may/should be determined so as to obtain the best result specifically for the applications.

Figure 6 shows the distortion against N, the number of points in spectrum input to DSM, when $\gamma = 3.0$, $v_c = 20.0$ dB and $q = 63$. The distortion is closely related to the quantization

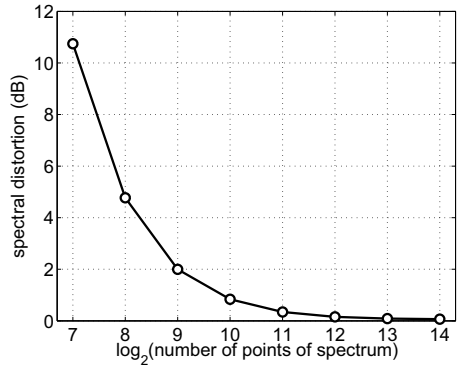


Figure 6: Spectral distortion against different number of points of spectrum input to DSM

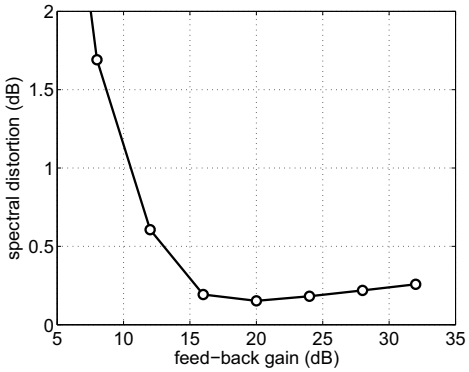


Figure 7: Spectral distortion as a function of DSM feedback gain

error of PDM. Benefiting from both Parseval's theorem and DSM's noise-shaping [10], the error decreases with increased oversampling. For example, in order to obtain the distortion of less than 1 dB, N should be at least 2^{10}.

Figure 7 shows the distortion as a function of the DSM feedback gain v_c, when $N = 2^{12}$, $\gamma = 3.0$ and $q = 63$. The distortion became minimal (0.1521 dB) when $v_c = 20.0$ dB. Theoretically, DSM cannot correctly encode a signal with amplitude greater than the feedback gain v_c and less than $-v_c$. Since the maximum of the absolute value of amplitude was 12.7 for the data set, the graph shows large distortion when $v_c < 12.0$ dB. When v_c becomes larger than 20 dB, the curve increases gradually. This is probably because of increasing quantization noise. It is noteworthy that the best accuracy is obtained when v_c is set slightly larger than the maximum of the absolute value of amplitude in the input to DSM.

Shown in Fig. 8 is the distortion with respect to the order of PDM cepstrum under $\gamma = 3.0$, $v_c = 20.0$ dB and $N = 2^{12}$. The distortion decreases as the PDM-cepstral order q increases. For example, in order to obtain the distortion of less than 1 dB, q should be at least 35–40.

4. Conclusions

We have investigated a new spectral representation suitable for statistical speech processing. We now focus our research direction on assessing the validity of this parameterisation in speech applications with statistical processing, e.g., an HMM-based speech synthesis system. The findings that were obtained from the investigation will be useful for such practical applications.

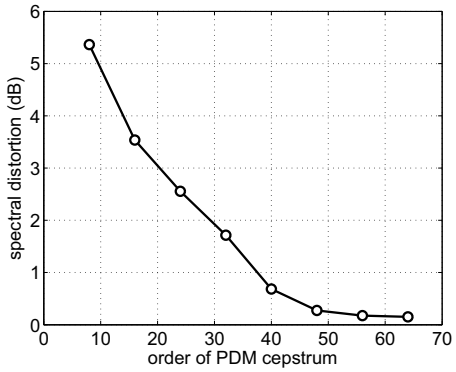


Figure 8: Spectral distortion versus the order of PDM cepstrum

5. Acknowledgements

The author would like to thank Shinsuke Sakai, Tomoki Toda and Keiichi Tokuda for their stimulating and helpful discussions; further, he would also like to thank Hisashi Kawai and Satoshi Nakamura for their support in carrying out this study.

6. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH'99*, Budapest, Hungary, Sep. 1999, pp. 2347–2350.

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

[3] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[4] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *IEICE Transactions*, vol. E91-D, no. 6, pp. 1764–1773, Jun. 2008.

[5] Y. Shiga and S. King, "Accurate spectral envelope estimation for articulation-to-speech synthesis," in *Proc. 5th ISCA Speech Synthesis Workshop*, CMU, Pittsburgh, Jun. 2004, pp. 19–24.

[6] J. Wouters and M. W. Macon, "Spectral modification for concatenative speech synthesis," in *Proc. ICASSP2000*, Istanbul, Turkey, 2000, pp. 941–944.

[7] D. T. Chappell and J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication*, vol. 36, no. 3, pp. 343–374, 2002.

[8] L. Qin, Y.-J. Wu, Z.-H. Ling, and R.-H. Wang, "Improving the performance of HMM-based voice conversion using context clustering decision tree and appropriate regression matrix format," in *Proc. Interspeech*, Pittsburgh, PA, Sep. 2006, pp. 2250–2253.

[9] Y. Shiga, "Pulse-density-modulated spectrum for statistical speech processing," in *Proc. 13$^{\text{th}}$ International Conference on Speech and Computer*, St. Petersburg, Russia, Jun. 2009.

[10] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. IEEE Press, 2005.

[11] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, CMU, Pittsburgh, Jun. 2004, pp. 223–224.

[12] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. ICASSP97*, vol. 2, Munich, Germany, Apr. 1997, pp. 1303–1306.