MINIMUM GENERATION ERROR TRAINING BY USING ORIGINAL SPECTRUM AS REFERENCE FOR LOG SPECTRAL DISTORTION MEASURE

Yi-Jian Wu, Keiichi Tokuda

Nagoya Institute of Technology, Nagoya, Japan

yjwu@sp.nitech.ac.jp, tokuda@nitech.ac.jp

ABSTRACT

This paper improves a minimum generation error (MGE) based HMM training technique for HMM-based speech synthesis by directly using the original spectrum instead of line spectral pairs (LSPs) as reference spectrum for log spectral distortion (LSD) measure. Two types of original reference spectra for LSD calculation are investigated, including the spectrum extracted from speech waveform by STRAIGHT, and the short-time FFT spectrum calculated from speech waveforms. Since only the harmonics of the FFT spectrum are coincident with the underlying spectral envelope, the LSD between generated LSPs and original FFT spectrum is calculated by sampling at the harmonic frequencies, and a weighting function is designed to simulate the sampling strategy on LSPs. From the experimental results, the MGE-LSD training using the FFT spectrum as reference spectrum achieved the best performance.

Index Terms— Speech synthesis, HMM, minimum generation error, log spectral distortion

1. INTRODUCTION

Speech synthesis systems have been under development for decades, and many research efforts have been made to improve the quality and flexibility of synthesized speech. In recent years, HMM-based speech synthesis [1] had been proposed, and shown its potential to realize a speech synthesis system with high quality and flexibility [2]. In this method, the spectrum, pitch and duration are modeled simultaneously in a unified framework [3], and the parameter sequence is generated by maximizing the likelihood of the HMMs related to the parameter sequence under the constraint between static and dynamic features [4]. Comparing to other synthesis methods, this method can learn salient statistical properties (such as speakers, speaking styles, emotions, and so on) from speech data, and generate smooth and stable speech under a small footprint.

In the conventional HMM-based speech synthesis framework, Maximum Likelihood (ML) criterion was adopted for HMM training. However, there are two issues [5] related to the ML-based HMM training for speech synthesis, including the mismatch between training and application of HMM, and the ignorance of constraint between static and dynamic features. In order to resolve these two issues, a minimum generation error (MGE) criterion [5] had been proposed for HMM training, where a generation error function using Euclidean distance was defined, and the HMM parameters were optimized so as to minimize the total generation errors of training data. Furthermore, a log spectral distortion (LSD) was adopted to replace the Euclidean distance to define the generation error between the original and generated line spectral pairs (LSPs) [6] in MGE training, and the quality of synthesized speech was improved [7].

In previous MGE-LSD training, the LSPs extracted from original speech waveforms were used as the reference for spectral distortion measure. In this paper, we continue to improve the MGE-LSD training by directly using the original spectrum for spectral distortion measure. We firstly adopt the spectral envelope extracted from original speech waveform by STRAIGHT [8] as the reference to calculate the LSDs, and define the generation error function. However, the speech waveforms are the actual target signals we want to simulate. The STRAIGHT-based spectral analysis basically can be regarded as a process to recover the spectral envelope from the short-time FFT spectrum calculated from speech waveform, and this process itself may loss some information in speech waveform. Therefore, we directly use the short-time FFT spectrum calculated from speech waveforms as the original reference spectrum for LSD calculation. Since only the harmonics of the FFT spectrum are coincident with the underlying spectral envelope, the LSD between generated LSPs and original FFT spectrum is calculated by sampling at the harmonic frequencies. The MGE-LSD training with FFT spectrum can be regarded as a unified training framework by incorporating spectral analysis and parameter generation into model training. It has a similar concept to the analysis-by-synthesis in speech coding and the closed-loop training [9] for concatenative speech synthesis .

The rest of this paper is organized as follows. In section 2, we briefly review the MGE criterion for HMM training. In section 3, we present the details of MGE-LSD training, including three types of original reference spectra for LSD calculation. In section 4, we describe experiments to evaluate the effectiveness of the MGE-LSD training with different types of LSD calculation, and present the results. Finally, our conclusion are given in section 5.

2. MINIMUM GENERATION ERROR CRITERION

The basic concept of MGE criterion is to calculate the generation errors by incorporating the parameter generation into training process, and then optimize the HMM parameters so as to minimize the total generation errors of training data.

2.1. Parameter generation

For a given HMM λ and the state sequence $\boldsymbol{q}$, the parameter generation algorithm [4] is to determine the speech parameter vector sequence $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top]^\top$ which maximizes $P(\boldsymbol{o}|\boldsymbol{q}, \lambda)$. In HMM-based speech synthesis, $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \Delta^{(2)}\boldsymbol{c}_t^\top]^\top$ includes not only static but also dynamic features. The constraint between static and dynamic feature vectors can be formulated as $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_T^\top]^\top$, and $\boldsymbol{W}$ is a regression matrix [4] for calculating dynamic features.

Under this constraint, parameter generation is equivalent to determining $\boldsymbol{c}$ to maximize $P(\boldsymbol{o}|\lambda, \boldsymbol{q})$. By setting $\partial P(\boldsymbol{o}|\lambda, \boldsymbol{q})/\partial \boldsymbol{c} = 0$, we obtain

$$\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{R}_{\boldsymbol{q}}^{-1}\boldsymbol{r}_{\boldsymbol{q}}, \tag{1}$$

where

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}, \quad \boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}, \tag{2}$$

and $\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_T^\top\right]^\top$ and $\boldsymbol{\Sigma}_{\boldsymbol{q}} = \mathrm{diag}\left(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_T\right)$ are the mean vector and covariance matrix related to $\boldsymbol{q}$, respectively.

2.2. MGE criterion with Euclidean distance measure

In previous MGE criterion [5], an Euclidean distance was adopted to measure the distortion between the original and generated feature vectors, which is calculated as

$$D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}) = \| \boldsymbol{c} - \bar{\boldsymbol{c}}_{\boldsymbol{q}} \|^2 . \tag{3}$$

Although the posterior probability can be used to weight the distances for all possible state sequences, it is computationally expensive for this direct calculation. Therefore, the representative n-best paths can be used to approximate the generation error. In our current implementation, only the optimal state sequence is used, and the generation error is defined as

$$e(\boldsymbol{c}, \lambda) = D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\hat{q}}), \tag{4}$$

where $\hat{\boldsymbol{q}}$ is the optimal state sequence for $\boldsymbol{o}$. This refers to a Viterbi-type MGE training. In the rest of the paper, we use $\boldsymbol{q}$ to denote $\hat{\boldsymbol{q}}$.

Based on the generation error measure, the parameter generation process is incorporated into HMM training for calculating the total generation errors for all training data, which is

$$E(\lambda) = \sum_n e(\boldsymbol{c}_n, \lambda). \tag{5}$$

Finally, the objective of MGE criterion is to optimize the model parameters so as to minimize the total generation errors, i.e.,

$$\hat{\lambda} = \arg\min E(\lambda). \tag{6}$$

As direct solution for Eq. (6) is mathematically intractable, a probabilistic descent (PD) [11] method was adopted for parameter optimization. The details of updating rules for mean and variance parameters in MGE training can be found in [5].

3. MGE TRAINING WITH LOG SPECTRAL DISTORTION

We use a log spectral distortion (LSD) instead of the Euclidean distance to define the generation errors for generated LSPs. Three types of original reference spectra are used to calculate the LSD, which includes the spectrum derived from original LSPs, the spectrum extracted from speech waveform by STRAIGHT, and the short-time FFT spectrum directly calculated from speech waveforms.

3.1. LSD between generated and original LSPs

This paper adopts line spectral pairs (LSP) as the spectral feature for HMM modeling. LSPs are derived from linear prediction coefficients (LPC). For a given p-th order LPC filter $A_p(z)$, two artificial $(p+1)$-th order polynomials can be constructed, which are

$$\begin{cases} P(z) \\ Q(z) \end{cases} = A_p(z) \pm z^{p+1} A_p(z^{-1}). \tag{7}$$

The LSP coefficients are related to the roots of the LSP polynomials. Let us denote $e^{j\omega_i}$ and $e^{-j\omega_i} (i = 1, \ldots, p)$ as the roots of a LSP polynomial, where ω_i are the LSP coefficients. Without loss of generality, we assume $\boldsymbol{p}$ is even number in the rest of the paper.

The log spectral distortion (LSD) between original and generated LSP feature vectors is calculated as

$$D_L(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{\pi} \int_0^{\pi} \left[\log |A_{\boldsymbol{c}}(\omega)| - \log |A_{\bar{c}}((\omega)| \right]^2 d\omega, \tag{8}$$

where $A_{\boldsymbol{c}}(\omega)$ and $A_{\bar{c}}(\omega)$ are the spectra related to $\boldsymbol{c}_t$ and $\bar{\boldsymbol{c}}_t$, respectively. Based on the definition of LSP, the power spectrum corresponding to a set of LSP coefficients can be calculated as

$$|A_{\boldsymbol{c}}(\omega)|^2 = \frac{1}{4} \left[|P_{\boldsymbol{c}}(\omega)|^2 + |Q_{\boldsymbol{c}}(\omega)|^2 \right], \tag{9}$$

where

$$|P_{\boldsymbol{c}}(\omega)|^2 = 4 \cos^2 \frac{\omega}{2} \prod_{i=1}^{\frac{p}{2}} 4 \left(\cos\omega - \cos c_{2i-1} \right)^2, \tag{10}$$

$$|Q_{\boldsymbol{c}}(\omega)|^2 = 4 \sin^2 \frac{\omega}{2} \prod_{i=1}^{\frac{p}{2}} 4 \left(\cos\omega - \cos c_{2i} \right)^2 . \tag{11}$$

From Eqs. (8)-(11), it is difficult to formulate the direct solution for the integration in Eq. (8). An alternative is to use a numerical integration to approximate the integral, which is calculated by accumulating the values of integrand at certain sampling points. Then Eq. (8) can be rewritten as

$$D_L(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{N_s} \sum_{j=1}^{N_s} \left[\log |A_{\boldsymbol{c}}(\omega_j)| - \log |A_{\bar{c}}(\omega_j)| \right]^2 , \tag{12}$$

where ω_j is the location of each sampling point and N_s is the total number of sampling points.

Two sampling strategies were investigated in [7], which includes the equidistance sampling, i.e.,

$$\omega_j = \frac{(2j-1)\pi}{2N_s}, \quad j = 1, 2, \ldots, N_s, \tag{13}$$

and the sampling at LSP frequencies, i.e.,

$$\omega_j = c_{t,j}, \quad j = 1, 2, \ldots, p, \tag{14}$$

where $c_{t,j}$ is the j-th coefficient of the original LSP vector $\boldsymbol{c}_t$. Compared to the equidistant sampling, the advantage of the latter sampling strategy is that it implicitly puts more weights on spectral peaks, and less weights on spectral valleys, which is due to one of the properties of LSP that there are more LSPs around spectral peaks. This is coincident with the human perception, which is more sensitive on spectral peaks than spectral valleys.

3.2. LSD between generated LSPs and spectrum extracted from original waveform

Previously we adopted the spectrum derived from original LSPs as the reference spectrum for LSD calculation. Since the LSPs are extracted from the original speech/spectrum, we can directly calculate the LSD between generated LSPs and original spectrum if the original spectrum are available. In this study, we use STRAIGHT [8] to extract the spectral envelope from the original speech waveform, and then use the extracted spectrum as the original reference spectrum to calculate the LSD for generated LSPs, i.e.,

$$D_S(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \int_0^{\pi} \left[\log |A_S(\omega)| - \log |A_{\bar{c}}(\omega)| \right]^2 d\omega, \tag{15}$$

where $A_S(\omega)$ is the spectrum extracted from original speech and $A_{\bar{c}}(\omega)$ is the spectrum related to the generated LSP vector $\bar{\boldsymbol{c}}_t$, respectively. Similarly, we can formulate the equations of numerical integration for this LSD function using the equidistance sampling and the sampling on LSPs, which are similar to Eq. (12)-(14). The only difference is to replace $A_{\boldsymbol{c}}(\omega)$ by $A_S(\omega)$.

3.3. LSD between generated LSPs and short-time FFT spectrum calculated from original waveform

Actually, the speech waveforms are the target signals we want to simulate. The STRAIGHT-based spectral analysis can be regarded as a process to recover the spectral envelope from the short-time FFT spectrum calculated from speech waveform, and this process itself may loss some information in speech waveform. Therefore, we directly adopt the short-time FFT spectrum calculated from speech waveform as the original reference spectrum for LSD calculation. Since only the spectral values of the FFT spectrum at the harmonic frequencies are reliable for estimating the spectral envelope, we need to calculate the LSD between generated LSPs and original FFT spectrum by sampling at the harmonic frequencies. The related numerical integration of LSD function is calculated as

$$D_F(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{N_h} \sum_{j=1}^{N_h} \left[\log |A_F(\omega_j)| - \log |A_{\bar{\boldsymbol{c}}}(\omega_j)|\right]^2, \tag{16}$$

where

$$\omega_j = 2j\pi f_0 / F_s, \quad j = 1, 2, \ldots, N_h, \tag{17}$$

$$N_h = \left[\frac{F_s}{2f_0}\right], \tag{18}$$

$A_F(\omega)$ is the FFT spectrum calculated from original speech waveform, f_0 is the fundamental frequency, F_s is the sampling rate of waveform, and N_h is the number of harmonics.

Based on the above calculation of LSD, the related MGE-LSD training can be regarded as a unified training framework by incorporating spectral analysis and parameter generation into model training process. Actually, these three processes can be combined in another way with different focus, where a statistical spectral analysis method was proposed in [10] by incorporating model training and parameter generation into spectral analysis process.

3.3.1. F0 optimization

Since a small error of F0 will result in a large difference in high-frequency harmonics, the accuracy of F0 is critical for the LSD calculation by harmonic sampling. After F0 extraction, we refine the F0 for each frame by searching a nearby F0 value to maximize the accumulated log spectral value on harmonic frequencies, which is

$$\hat{f}_0 = \arg\max_{f_0} \sum_{j=1}^{N_h} \log |A_F(2j\pi f_0 / F_s)|, \tag{19}$$

3.3.2. Weighting function from LSPs

As we mentioned, the sampling on LSPs for LSD calculation is equivalent to putting more weights on the region with dense LSPs and less weights on the region with sparse LSPs. From this point, we design a weighting function to simulate the sampling on LSPs for the above integration by the harmonic sampling.

For a given sequence of sampling points $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_N]$, we define a related effective region for each point ω_i, whose left and right boundaries are

$$\omega_i^{(l)} = (\omega_{i-1} + \omega_i)/2, \tag{20}$$

$$\omega_i^{(r)} = (\omega_{i+1} + \omega_i)/2. \tag{21}$$

Lets assume i' and i'' satisfy $\omega_i^{(l)} \in (c_{i'-1}, c_{i'})$ and $\omega_i^{(r)} \in (c_{i''-1}, c_{i''})$. Finally, the weight for the sampling point ω_i is defined as the ratio of the number of LSP points covered by $[\omega_i^{(l)}, \omega_i^{(r)}]$, i.e.,

$$\varphi_i = \frac{1}{p+1}\left[\frac{c_{i'} - \omega_i^{(l)}}{c_{i'} - c_{i'-1}} - \frac{c_{i''} - \omega_i^{(r)}}{c_{i''} - c_{i''-1}} + (i'' - i')\right]. \tag{22}$$

3.4. Parameter updating

Here we use the LSD between original and generated LSPs as example to formulate the updating rules in MGE-LSD training. For the cases of using other two LSDs, we only need to replace $A_{\boldsymbol{c}}(\omega)$ with $A_S(w)$ or $A_F(w)$ in the formulation. Under the MGE criterion, we minimize the total generation errors

$$E'(\lambda) = \sum_n e'(\boldsymbol{c}_n, \lambda) = \sum_n \sum_{t=1}^{T} D_L(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t), \tag{23}$$

with respect to

$$\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \ldots, \boldsymbol{\mu}_K^\top\right]^\top, \tag{24}$$

$$\boldsymbol{U} = \left[\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, \ldots, \boldsymbol{\Sigma}_K^{-1}\right]^\top, \tag{25}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k-th Gaussian component, and K is the total number of Gaussian components in the model set λ.

The PD method [11] is adopted for parameter optimization. For each training utterance $\boldsymbol{c}_\tau$, the parameter set is updated as

$$\lambda_{\tau+1} = \lambda_\tau - \epsilon_\tau \boldsymbol{H}_\tau \left.\frac{\partial e'(\boldsymbol{c}_\tau, \lambda)}{\partial \lambda}\right|_{\lambda=\lambda_\tau}, \tag{26}$$

where $\boldsymbol{H}_\tau$ is a positive definite matrix, and ϵ_τ is a learning rate that decreases when utterance index τ increases.

For the mean and variance parameters, the gradients of the generation error function are calculated as

$$\frac{\partial e'(\boldsymbol{c}_\tau, \lambda)}{\partial \boldsymbol{\mu}} = 2\boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} \boldsymbol{R}_{\boldsymbol{q}}^{-1} \boldsymbol{\zeta}, \tag{27}$$

$$\frac{\partial e'(\boldsymbol{c}_\tau, \lambda)}{\partial \boldsymbol{U}} = 2\boldsymbol{S}_{\boldsymbol{q}}^\top \operatorname{diag}^{-1}\left(\boldsymbol{W} \boldsymbol{R}_{\boldsymbol{q}}^{-1} \boldsymbol{\zeta} (\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{W} \bar{\boldsymbol{c}}_{\boldsymbol{q}})\right), \tag{28}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} = \operatorname{diag}(\boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{U}), \tag{29}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{m}, \tag{30}$$

$$\boldsymbol{\zeta} = \left[\boldsymbol{\zeta}_1^\top, \boldsymbol{\zeta}_2^\top, \ldots, \boldsymbol{\zeta}_T^\top\right]^\top, \tag{31}$$

$$\boldsymbol{\zeta}_t = \left[\zeta_{t,1}, \zeta_{t,2}, \ldots, \zeta_{t,p}\right]^\top, \tag{32}$$

$$\zeta_{t,i} = \frac{1}{2S} \sum_{s=1}^{S} \frac{\left|X_{\bar{\boldsymbol{c}}}^{(i)}(\omega_s)\right|^2}{\left|A_{\bar{\boldsymbol{c}}}(\omega_s)\right|^2} \frac{\sin \bar{c}_{t,i}}{\cos \omega_s - \cos \bar{c}_{t,i}} \log \left|\frac{A_{\bar{\boldsymbol{c}}}(\omega_s)}{A_{\boldsymbol{c}}(\omega_s)}\right|, \tag{33}$$

$$X_{\bar{\boldsymbol{c}}}^{(i)}(\omega_j) = \begin{cases} P_{\bar{\boldsymbol{c}}}(\omega_j), & i \text{ is odd} \\ Q_{\bar{\boldsymbol{c}}}(\omega_j), & i \text{ is even} \end{cases}. \tag{34}$$

In the above equations, $\boldsymbol{S}_{\boldsymbol{q}}$ is a matrix whose elements are 0 or 1 determined according to the optimal state sequence $\boldsymbol{q}$ for $\boldsymbol{c}_\tau$. The operation of diag$(\cdot)$ is to convert a $3DT \times 3D$ matrix to a $3DT \times 3DT$ block-diagonal matrix with a block size of $3D$, and diag$^{-1}(\cdot)$ is the inverse operation of diag$(\cdot)$.

It should be noted that the above formulation of updating rules are valid for all types of LSDs calculated by the equidistant sampling, the harmonic sampling, and the sampling at LSP frequencies, respectively. The only differences between them are the number of sampling points N_s and the position of each sampling point ω_j.

4. EXPERIMENTS

4.1. Experimental conditions

We used the phonetically balanced 503 sentences from ATR Japanese speech database (B-set, MHT) in the experiment. The first 450 sentences were used as training data, and the remaining 53 sentences were used for evaluation. Speech signals were sampled at a rate of 16kHz. The acoustic features include F0 and LSP coefficients, where LSP coefficients were calculated based on spectra extracted by STRAIGHT [8]. The feature vector consists of static features (including 24-th LSP coefficients, logarithm of gain and logarithm of F0), and their delta and delta-delta coefficients. A 5-state left-to-right no-skip HMM structure was used, and MSD-HMM [12] was adopted for F0 modeling. In synthesis, the STRAIGHT synthesis filter was used to synthesize the speech waveform.

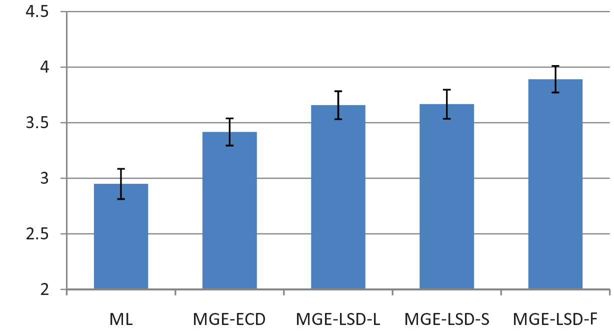

Fig. 1. MOS scores for ML training and MGE training with different spectral distortion measures

The HMM training in this experiment was performed as follows. Firstly, the conventional ML-based HMM training was conducted. Then the optimal state alignments for all training data were obtained using the ML-trained HMMs. Finally, the MGE training was applied to re-estimate the parameters of ML-trained HMMs. In the experiments, we conducted the MGE training with different configurations, which are as follows:

a) MGE-ECD: MGE training with Euclidean distance;
b) MGE-LSD-L: MGE training with LSD between generated and original LSPs;
c) MGE-LSD-S: MGE training with LSD between generated LSPs and original STRAIGHT-extracted spectrum;
d) MGE-LSD-F: MGE training with LSD between generated LSPs and original FFT spectrum;

In the configurations of MGE-LSD-L and MGE-LSD-S, the LSDs were calculated by sampling at LSP frequencies. In the configuration of MGE-LSD-F, we adopted Eq. (19) to refine the extracted F0, and the weighting function in Eq. (22) to simulate the sampling at LSP frequencies. Since we aimed to evaluate the effectiveness of different spectral distortion measures for MGE training, only spectrum part of model parameters were updated in MGE training.

4.2. Experimental results

A formal subjective listening test was conducted to evaluate the performances of MGE training with different spectral distortion measures. Five training configurations, including ML, MGE-ECD, MGE-LSD-L, MGE-LSD-S and MGE-LSD-F, were evaluated. Eight listeners participated in the test. Each listener evaluated 15 sets of samples consisting of five synthesized speech samples, and gave the MOS on the naturalness. The speech samples were randomly selected for each listener from the 53 test sentences.

The results are shown in Fig. 1, with vertical lines indicating the 95% confidence intervals. It can be seen that the quality of synthesized speech was gradually improved when we apply MGE-ECD training, and then MGE-LSD training. From this figure, MGE-LSD-L and MGE-LSD-S achieve the similar MOS scores, which can be explained as follows. Although using 24-order LSPs to represent the extracted spectrum introduces a little spectral distortion, the spectrum derived from LSPs usually has sharper formants than the original extracted spectrum, which means the conversion from spectrum to LSPs partially enhances the spectral formants. Therefore, such effect of formant enhancement could compensate the spectral distortion after converting spectrum to LSPs. Among all the configurations of MGE-LSD training, MGE-LSD-F achieved the best performance. As we mentioned in Sec. 3.3, the MGE-LSD training with the FFT spectrum as original reference spectrum for LSD calculation can be regarded as a unified training framework, where spectral analysis and parameter generation was incorporated into model training process. Such unified training framework eliminated the mismatch between these three components, and thus improved the performance.

5. CONCLUSIONS

This paper introduces the MGE training with a log spectral distortion (LSD) for measuring the distortion of generated LSPs. We compared three types of original reference spectra for calculating the LSD, which includes the spectrum derived from original LSPs, the extracted spectrum by STRAIGHT, and the short-time FFT spectrum directly calculated from speech waveforms. Experimental results showed that using the LSDs calculated with the FFT spectrum as reference spectrum achieved the best performance, and the quality of synthesized speech after the MGE-LSD training was significantly improved over the original ML and MGE-ECD training.

6. ACKNOWLEDGEMENTS

This work was partly supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

7. REFERENCES

[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," in *Proc. of ICASSP*, pp. 389–392, 1996.

[2] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system - HTS-2007 system for the Blizzard Challenge 2007", in *Blizzard Challenge 2007*.

[3] 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. of ICASSP*, vol. 5, pp. 2347–2350, 1999.

[4] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. of ICASSP*, pp. 660–663, 1995.

[5] Y.-J. Wu and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. of ICASSP*, vol. 1, pp. 889–892, 2006.

[6] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," in *J. Acoust. Soc. Amer.*, vol. 57, p. 535(a), p. s35(A), 1975.

[7] Y.-J. Wu and K. Tokuda, "Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis," in *Proc. of Interspeech*, pp. 577–580, 2008.

[8] H. Kawahara, I. Masuda-Katsuse and A. deCheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," in *Speech Communication*, vol. 27, pp. 187–207, 1999.

[9] M. Akamine and T. Kagoshima, "Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS drive TTS)," in *Proc. of ICSLP*, 1998.

[10] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM," in *Proc of ICASSP*, pp. 3925–3928, 2008.

[11] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, vol. EC-16, no. 3, pp. 299–307, 1967.

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, pp. 229–232, 1999.