interspeech 2008 incorporating SST 08

Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis

Yi-Jian Wu, Keiichi Tokuda

Nagoya Institute of Technology, Nagoya, Japan
yjwu@sp.nitech.ac.jp, tokuda@nitech.ac.jp

Abstract

A minimum generation error (MGE) criterion had been proposed to solve the issues related to maximum likelihood (ML) based HMM training in HMM-based speech synthesis. In this paper, we improve the MGE criterion by imposing a log spectral distortion (LSD) instead of the Euclidean distance to define the generation error between the original and generated line spectral pair (LSP) coefficients. Moreover, we investigate the effect of different sampling strategies to calculate the integration of the LSD function. From the experimental results, using the LSDs calculated by sampling at LSPs achieved the best performance, and the quality of synthesized speech after the MGE-LSD training was improved over the original MGE training.

Index Terms: Speech synthesis, HMM, minimum generation error, log spectral distortion, line spectral pairs

10.21437/Interspeech.2008-170

1. Introduction

Speech synthesis has been studied for several decades, and many effective methods and techniques had been developed. In recent years, HMM-based speech synthesis was proposed [1]. In this method, the spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [2], 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 [3]. Under its statistical training framework, it can learn salient statistical properties of speakers, speaking styles, emotions, and etc., from the speech corpus. The recent improvements and implementations [4, 5, 6] showed its potential to realize a speech synthesis system with high quality and flexibility.

In the conventional HMM-based speech synthesis framework, Maximum Likelihood (ML) criterion was adopted for HMM training. Although its performance is quite good, there are two issues [7] related to ML-based HMM training, 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 had been proposed for HMM training [7], where a generation error function was firstly defined, and the HMM parameters were optimized by using probabilistic descent (PD) [8] method so as to minimize the total generation errors of training data. Furthermore, it had been applied to the tree-based clustering for context dependent HMMs and the whole HMM training procedure [9].

In original MGE criterion, the Euclidean distance was adopted to measure the distortion between the original and generated acoustic features. Although we used line spectral pairs (LSP) [10] as the spectral feature for HMM modeling, the Euclidean distance between two LSPs is not so convincing as a spectral distortion measure. In fact, there are many meaningful spectral distortion measures which were popularly used, such as log spectral distortion (LSD), Itakura-Saito distortion, and etc [11]. In this paper, we adopt the LSD to replace the Euclidean distance for generation error definition in MGE criterion, and reformulate the updating rules for model parameters. Because the integration in LSD calculation cannot be solved directly, we need to compute it by numerical integration, where the integral is approximated by accumulating the values of integrand at certain sampling points. In addition, we investigate the effects of two sampling ways for LSD calculation, including equidistant sampling and sampling at LSP frequencies.

The rest of the paper is organized as follows. In section 2, we briefly review the parameter generation algorithm, the MGE criterion for HMM training, and the properties of LSP. In section 3, we present the details of imposing the LSD to define the generation error for LSPs in MGE criterion, and formulate the related updating rules for model parameters. In section 4, we describe the experiments to evaluate the effectiveness of the MGE training with the LSD, and show the results. Finally, our conclusions are given in section 5.

2. Related techniques

2.1. Parameter generation algorithm

For a given HMM λ and the state sequence $\boldsymbol{q}$, the parameter generation algorithm 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 order to keep the smooth property of the generated parameter sequence, the dynamic features including delta and delta-delta coefficients $\Delta^{(n)}\boldsymbol{c}_t (n = 1, 2)$ are used, i.e., the parameter vector can be rewritten as

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \Delta^{(2)}\boldsymbol{c}_t^\top\right]^\top. \tag{1}$$

The constraints between static and dynamic feature vector can be formulated as $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_T^\top\right]^\top$. Due to limited space, here the details of $\boldsymbol{W}$ are not given, which can be found in [3, 7].

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

$$\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{R}_{\boldsymbol{q}}^{-1}\boldsymbol{r}_{\boldsymbol{q}}, \tag{2}$$

where

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}, \quad \boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}, \tag{3}$$

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. Minimum generation error criterion

In previous MGE criterion [7], the Euclidean distance was adopted to measure the distortion between the original and generated feature vectors, which is calculated as

$$D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}) = \| \boldsymbol{c} - \bar{\boldsymbol{c}}_{\boldsymbol{q}} \|^2 . \quad (4)$$

Although the posterior probability $P(\boldsymbol{q}|\lambda, \boldsymbol{o})$ can be used to weight the distance for all possible state sequence $\boldsymbol{q}$, it is computationally expensive for this direct calculation. Therefore, the representative n-best paths can be used to approximate the generation error. In the real implementation, only the optimal state sequence is used and the generation error is defined as

$$e(\boldsymbol{c}, \lambda) = D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\hat{\boldsymbol{q}}}), \quad (5)$$

where $\hat{\boldsymbol{q}}$ is the optimal state sequence for $\boldsymbol{o}$. In fact, this refers to a Viterbi-type MGE training. In the following part of this paper, we use $\boldsymbol{q}$ to denote $\hat{\boldsymbol{q}}$ by default.

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

$$E(\lambda) = \sum_{n=1}^{N} e(\boldsymbol{c}_n, \lambda), \quad (6)$$

where N is the total number of training utterances.

Finally, the object of MGE criterion is defined, which is to optimize the parameters of HMMs so as to minimize the total generation errors

$$\hat{\lambda} = \arg\min E(\lambda). \quad (7)$$

As direct solution for Eq. (7) is mathematically intractable, probabilistic descent (PD) [8] method was adopted for parameter optimization. The details of updating rules for mean and variance parameters in MGE training can be found in [7].

2.3. Line spectral pairs

In this paper, line spectral pairs (LSP) is adopted as the spectral feature for HMM modeling. Here we review some properties of LSP. LSP is derived from LPC (linear prediction coefficient) filter. An LPC filter is defined as

$$H_p(z) = \frac{G}{A_p(z)}, \quad (8)$$

$$A_p(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-p}, \quad (9)$$

where p is the order of LPC filter, and G is the gain.

For a given p-th order LPC filter, we can construct two artificial $(p+1)$-th order polynomials, which are

$$P(z) = A_p(z) + z^{p+1} A(z^{-1}), \quad (10)$$

$$Q(z) = A_p(z) - z^{p+1} A(z^{-1}). \quad (11)$$

The LSP coefficients are related to the roots of the LSP polynomials. Lets denote $e^{j\omega_i}$ and $e^{-j\omega_i} (i = 1, \ldots, p)$ as the roots of LSP polynomial, where ω_i are the LSP coefficients. Without loss of generality, we assume the order $\boldsymbol{p}$ **is even number** in the rest of the paper. The polynomials $P(z)$ and $Q(z)$ can be rewritten as

$$P(z) = (z+1) \prod_{i=1}^{\frac{p}{2}} (z - e^{j\omega_{2i-1}})(z - e^{-j\omega_{2i-1}}), \quad (12)$$

$$Q(z) = (z-1) \prod_{i=1}^{\frac{p}{2}} (z - e^{j\omega_{2i}})(z - e^{-j\omega_{2i}}). \quad (13)$$

The LSP coefficients have several useful properties. Firstly, the LSP has good interpolation property, which is suitable for HMM modeling and generation. Secondly, a cluster of adjacent LSPs characterizes a formant frequency, and the bandwidth of a given formant depends on the closeness of the corresponding LSPs. Furthermore, the spectral sensitivities of LSPs are localized, i.e., a change in a given LSP produces a change in the LPC power spectrum only in its neighborhood.

3. MGE with log spectral distortion

3.1. Log spectral distortion for LSPs

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

$$D_{lsd}(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{\pi} \int_0^{\pi} [\log |A_{\boldsymbol{c}}(\omega)| - \log |A_{\bar{\boldsymbol{c}}}((\omega)|]^2 \, d\omega. \quad (14)$$

where $A_{\boldsymbol{c}}(\omega)$ and $A_{\bar{\boldsymbol{c}}}(\omega)$ are the spectra related to $\boldsymbol{c}_t$ and $\bar{\boldsymbol{c}}_t$, respectively. Based on the definition of LSP in Eqs. (10)-(13), the power spectrum corresponding to a set of LSP can be calculated as

$$|A(\omega)|^2 = \frac{1}{4} \left[|P(\omega)|^2 + |Q(\omega)|^2 \right], \quad (15)$$

where

$$|P(\omega)|^2 = 4 \cos^2 \frac{\omega}{2} \prod_{i=1}^{\frac{p}{2}} 4 (\cos\omega - \cos c_{2i-1})^2 , \quad (16)$$

$$|Q(\omega)|^2 = 4 \sin^2 \frac{\omega}{2} \prod_{i=1}^{\frac{p}{2}} 4 (\cos\omega - \cos c_{2i})^2 . \quad (17)$$

From Eqs. (14)-(17), it is difficult to formulate the direct solution for the integration in Eq. (14). An alternative is to use a numerical integration to approximate the integral, which is calculated by accumulating values of integrand at certain sampling points. Then Eq. (14) can be rewritten as

$$D_{lsd}(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{S} \sum_{s=1}^{S} [\log |A_{\boldsymbol{c}}(\omega_s)| - \log |A_{\bar{\boldsymbol{c}}}(\omega_s)|]^2 , \quad (18)$$

where

$$\omega_s = \frac{(2s-1)\pi}{2S}, \quad s = 1, 2, \ldots, S, \quad (19)$$

and S is the number of sampling points. It can be seen that the approximation becomes more accurate when S increases. However, the computational cost increases simultaneously. We need to set an appropriate value for S to balance the accuracy and efficiency.

The above numerical integration can be regarded as an equidistant sampling of power spectrum in the frequency domain. Accordingly, other sampling strategies can also be applied. Here we sample the power spectrum on each LSP frequency, and calculate the integral as

$$D_{lsd}(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{p} \sum_{s=1}^{p} [\log |A_{\boldsymbol{c}}(\omega_s)| - \log |A_{\bar{\boldsymbol{c}}}(\omega_s)|]^2 , \quad (20)$$

where

$$\omega_s = c_{t,s}, \quad s = 1, 2, \ldots, p \quad (21)$$

and $c_{t,k}$ is the k-th coefficient of the original LSP vector $\boldsymbol{c}_t$.

Compared to the equidistant sampling, the advantage of this sampling strategy is that it implicitly puts more weight on spectral peaks, and less weight 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. MGE-based Parameter updating

With the log spectral distortion, we define a new generation error function for the original LSP vector sequence $\boldsymbol{c}$ as

$$e'(\boldsymbol{c},\lambda) = D_{lsd}(\boldsymbol{c},\bar{\boldsymbol{c}}_{\boldsymbol{q}}) = \sum_{t=1}^{T} D_{lsd}(\boldsymbol{c}_t,\bar{\boldsymbol{c}}_t), \tag{22}$$

where $\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \left[\bar{\boldsymbol{c}}_1^\top, \bar{\boldsymbol{c}}_2^\top, \ldots, \bar{\boldsymbol{c}}_T^\top\right]^\top$ is the generated LSP vector sequence. Finally, the new MGE training algorithm is to minimize the total generation errors

$$\hat{\lambda} = \arg\min \sum_{n=1}^{N} e'(\boldsymbol{c}_n,\lambda), \tag{23}$$

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 unique Gaussian component, and K is the total number of Gaussian components in the model set λ.

The PD method [8] is adopted here for parameter optimization. For each training utterance $\boldsymbol{c}_\tau$, the parameter set is updated as

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

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

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

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

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

where

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

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

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

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

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

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

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

It should be noted that the above formulation of updating rules are valid for both LSDs calculated by the equidistant sampling and by sampling at LSP frequencies. The only differences between them are the number of sampling points S and the positions of sampling points ω_s.

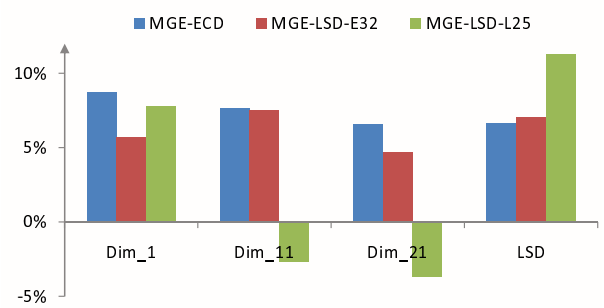

Figure 1: Relative reduction of generation errors after MGE training

4. Experiments

4.1. Experimental setups

We used the phonetically balanced 503 sentences from ATR Japanese speech database (B-set, MHT) in this experiments. The first 450 sentences were used as the training data, and the remaining 53 sentences were used for evaluation. The 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 [13]. 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 was used, and MSD-HMM [12] was adopted for F0 modeling. In synthesis, the STRAIGHT synthesis filter was used to synthesize the speech waveform.

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

a) Original MGE training with Euclidean distance measure (MGE-ECD);
b) MGE training with LSD which is calculated by equidistant sampling, where S were set to 32 and 512;
c) MGE training with LSD which is calculated by sampling at LSP frequencies and zero point, i.e., $S = 25$;

Since we aim to compare the effectiveness of MGE training with different spectral distortion measures, only spectrum part of model parameters were updated in MGE training.

4.2. Experimental results

4.2.1. Effect of sampling strategies

Fig. 1 shows the relative reduction of generation errors on the test data after MGE training, which includes the Euclidean distance between original and generated LSPs (i.e., ECD errors) for several typical dimensions, and the LSDs between original and generated LSPs (i.e, LSD errors). From this figure, although the original MGE-ECD training focus on minimizing the ECD errors, the LSD errors are alleviated in certain extent as a by-product. After the MGE-LSD training, the LSD errors is largely reduced, especially for the case that the LSD is calculated by sampling at LSPs. However, the improvement for the ECD errors is less than that after the MGE-ECD training. and the relative reduction rates for some dimensions (e.g., 11th and

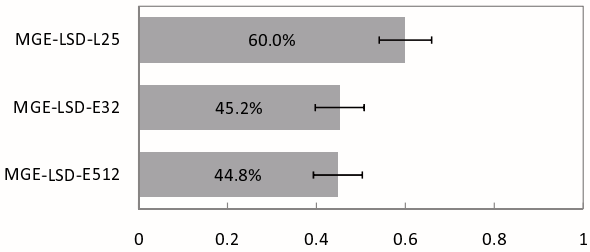

Figure 2: Preference scores for different sampling strategies

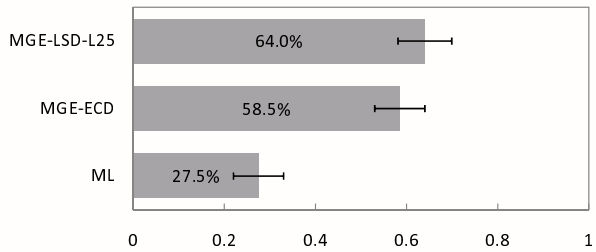

Figure 3: Preference scores for different training procedures

21th) are even less than 0, which means the ECD errors is worse than the baseline ML training. Since the calculation of LSD in Eq. (14) is based on all dimension of LSPs, minimization of the LSD errors does not guarantee the reduction of the ECD errors for each dimensions. From this point, the effect of MGE-LSD training is reasonable.

In order to evaluate the performance of different sampling strategies in MGE-LSD training, we conducted a formal subjective listening test. The quality of synthesized speech was compared by a paired comparison. Three sampling strategies including the equidistant sampling on 32 (LSD-E32) and 512 (LSD-E512) points, and the sampling at LSP frequencies with zero point (LSD-L25). Eight Japanese listeners participated in the test. They were presented pairs of synthesized speech in random order, and asked which one sounded better. For each listener, 25 test sentences were randomly selected from the 53 test sentences.

Fig. 2 shows the preference scores with 95% confidence interval. It is obvious that the sampling at LSP frequencies achieve the best performance, which is coincident with our description of its advantage in Section 3.1. Moreover, by comparing scores of LSD-E32 and LSD-E512, there is no improvement when increasing the number of sampling points from 32 to 512.

4.2.2. Effect of MGE-LSD training

Finally, a subjective listening test was conducted to evaluate the effectiveness of MGE-LSD training by comparing it with ML and MGE-ECD training. The setting of MGE-LSD-L25 was adopted based on the experiment in Sec. 4.2.1. The synthesized speech samples from the models trained by above three training procedures were compared by paired comparison. Other setting of the listening test is the same as the previous one.

The results are shown in Fig. 3. It can be seen that the MGE training significantly improves the quality of synthesized speech over the original ML training, and the performance of MGE-LSD training is better than that of MGE-ECD training. By comparing the synthesized speech after MGE-ECD and MGE-LSD training, we found that the clearness of synthesized speech was improved, and the artificial effect was reduced after MGE-LSD training.

5. Conclusions

In this paper, a log spectral distortion (LSD) is incorporated into MGE training by replacing the Euclidean distance to define the generation error between original and generated LSPs. We investigated the effect of different sampling strategies, including equidistant sampling in frequency domain and sampling at LSP frequencies, to calculate the integration of LSD function. Experiment results showed that using the LSDs calculated by sampling at LSP frequencies achieved the best performance, and the quality of synthesized speech after the MGE-LSD training was improved over the original 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*, 1996, pp. 389–392.

[2] 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*, 1999, vol. 5, pp. 2347–2350.

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

[4] H. Zen, and T. Toda, "An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005," in *Proc. of Eurospeech*, 2005, pp. 93-96, 2005.

[5] 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 *Interspeech 2006 satellite meeting, Blizzard Challenge 2006*.

[6] 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*.

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

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

[9] Y.-J. Wu, R.H. Wang, and F. Soong, "Full HMM training for minimizing generation error in synthesis," in *Proc. of ICASSP*, 2007, pp. 517–520.

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

[11] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition Englewood Cliffs," Prentice-Hall, NJ, USA, 1993.

[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*, 1999, pp. 229–232.

[13] 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.