MINIMUM GENERATION ERROR TRAINING WITH WEIGHTED EUCLIDEAN DISTANCE ON LSP FOR HMM-BASED SPEECH SYNTHESIS

Ming Lei, Zhen-Hua Ling, Li-Rong Dai

iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, China
leiming@mail.ustc.edu.cn, zhling@ustc.edu, lrdai@ustc.edu.cn

ABSTRACT
This paper presents a minimum generation error (MGE) training method using weighted Euclidean distance measure on line spectral pairs (LSP) for HMM-based speech synthesis. In this paper, weighted Euclidean distance on LSP is introduced as the measurement of generation error to improve the consistency between the model training criterion and the subjective perception on the distortion of synthetic speech. Several common weighting techniques are investigated and compared within the MGE training framework. The experimental results show that the formant bounded weighting (FBW) method achieves the best performance, which improves the naturalness of synthetic speech significantly compared with Euclidean distance measure. Compared with the MGE training using log spectral distortion (LSD) measure, the FBW criterion can achieve similar performance on naturalness with much less computation complexity of model training.

Index Terms— Speech synthesis, hidden Markov model, minimum generation error, line spectral pairs

1. INTRODUCTION
In recent years, hidden Markov model (HMM) based speech synthesis approach has been proposed and developed into a mainstream method for building natural sounding synthetic voices with high flexibility [1]. In the training procedure of this method, speech features including spectrum, pitch and duration are modeled simultaneously in a statistical framework [2]. At synthesis stage, speech features are predicted from the trained models by maximum likelihood parameter generation (MLPG) algorithm [3] and sent to the parametric synthesizer to reconstruct speech waveforms.

There exists two issues in the conventional HMM-based speech synthesis framework where maximum likelihood (ML) criterion is adopted for model training [4]. They are the inconsistency between the model training criterion and the application of speech synthesis, and the ignorance of constraints between static and dynamic features during model training. To solve these issues, minimum generation error (MGE) training was proposed [4], which estimates model parameters by minimizing the Euclidean distance between the predicted and the natural spectral parameters. Furthermore, the log spectral distortion (LSD) was introduced to replace the Euclidean distance in the definition of generation error for line spectral pairs (LSP) to achieve better consistency between the model training criterion and the subjective perception on synthetic speech [5, 6]. However, the MGE-LSD method leads to heavy computation cost in model training because of the complex function used to convert LSP into spectral envelope for calculating the LSD and the gradient in the probabilistic descent (PD) based model updating.

On the other hand, due to the wide application of LSP feature in speech coding, many research work has been conducted to develop a weighted Euclidean distance on LSP to measure the spectral distortion caused by vector quantization (VQ) [7–11], where the weighting coefficients are defined in different ways to approximate the distortion of spectral envelope or subjective perception. In this paper, we introduce the weighted Euclidean distance measure on LSP into the framework of MGE training and several weighting methods are investigated. Similar to MGE-LSD, this method adopts a more subjective perception-related definition of generation error for LSP than the MGE training using Euclidean distance measure. Further, the computation complexity of proposed method is much lower than that of MGE-LSD training because of the simple form of weighted Euclidean distance function.

This paper is organized as follows. Section 2 reviews the MGE criterion for HMM training. In Section 3, the proposed MGE training method with weighted Euclidean distance on LSP is presented. Four weighting methods are introduced in Section 4. And Section 5 describes some objective and subjective experiment results. Finally, our conclusion is given in Section 6.

2. MINIMUM GENERATION ERROR TRAINING
In the HMM-based parametric speech synthesis, MLPG algorithm [3] is adopted to predict speech parameters. The MGE criterion [4] is designed to optimize the models by minimizing the generation error between predicted and natural speech parameters in the training set.

2.1. Maximum likelihood parameter generation
For a given HMM $\lambda$ and state sequence $q$, the parameter generation algorithm is to predict the speech parameter vector sequence $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_T]$ which maximizes $P(\sigma | \lambda, q)$. In order to get a smooth sequence of generated speech parameters, dynamic features are incorporated to define $\sigma_t = [c_t, \Delta c_t, \Delta^2 c_t]^T$, where $c_t$ is the static component, $\Delta c_t$ and $\Delta^2 c_t$ are the velocity and acceleration components respectively. Further, there exists constraints between the static feature sequence $c = [c_1, c_2, \ldots, c_T]^T$ and $\sigma$ as $\sigma = W c$, where $W$ is determined by the velocity and acceleration calculation functions. By setting $\partial P(\sigma | \lambda, q) / \partial c = 0$, we could obtain the predicted static feature sequence $\tilde{c} = [\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_T]^T$ as

$$\tilde{c} = R_q^{-1} r_q \quad (1)$$

$$R_q = W^T \Sigma_q^{-1} W \quad (2)$$

$$r_q = W^T \Sigma_q^{-1} \mu_q \quad (3)$$

where $\mu_q$ and $\Sigma_q$ are the mean vector and covariance matrix decided by the state sequence $q$.

2.2. Minimum generation error training
MGE criterion is to optimize the model $\lambda$ by minimizing a defined generation error between the natural speech parameters $c$ and the
predicted parameters \( \bar{c} \), which is
\[ \hat{\lambda} = \arg \min D(c, \bar{c}). \] (4)
Conventionally, the generation error is defined in the form of Euclidean distance as
\[ D(c, \bar{c}) = \|c - \bar{c}\|^2 = \sum_{t=1}^{T} \sum_{i=1}^{N} (\bar{c}_{t,i} - c_{t,i})^2 \] (5)
where \( T \) is the length of the feature sequence; \( N \) is the dimension of static feature component; \( \bar{c}_{t,i} \) and \( c_{t,i} \) are the \( i \)-th dimension of generated and natural static features at frame \( t \). Then, a probabilistic descent (PD) algorithm is applied to update the model parameters iteratively [4].

Furthermore, in the MGE-LSD training for LSP parameters, the generation error is defined as
\[ D(c, \bar{c}) = \sum_{t=1}^{T} d_L(c_t, \bar{c}_t) = \sum_{t=1}^{T} \frac{1}{\pi} \int_{0}^{\pi} \frac{|H_{c_t}(\omega)|^2}{|H_{\bar{c}_t}(\omega)|^2} d\omega \] (6)
where \( H_{c_t}(\omega) \) is the spectrum recovered from generated LSP parameters, and \( H_{\bar{c}_t}(\omega) \) is the spectrum recovered from natural LSP parameters [5] or extracted from natural speech waveforms directly [6].

3. MGE TRAINING WITH WEIGHTED LSP DISTANCE

The definition of generation error in Eq.(4) plays an important role in the MGE training. Basically, it is expected to be able to present the subjective perception on the distortion of generated spectral parameters compared with the natural ones. Previous research work on efficient VQ of LSP in speech coding [8–11] informs us that weighted Euclidean distance is more appropriate than the Euclidean distance measure as Eq.(5) to define the generation error for LSP features. Therefore, we introduce the weighted Euclidean distance on LSP into the MGE training and define
\[ D_w(c, \bar{c}) = \sum_{t=1}^{T} \sum_{i=1}^{N} w_{t,i}(\bar{c}_{t,i} - c_{t,i})^2 \] (7)
where \( w_{t,i} \) is the weighting coefficient for the LSP feature of dimension \( i \) at frame \( t \). The above equation can be rewritten as
\[ D_w(c, \bar{c}) = (\bar{c} - c)^T \Gamma (\bar{c} - c) \] (8)
where
\[ \Gamma = \text{diag}\{\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(T)}\}, \]
\[ \Gamma^{(t)} = \text{diag}\{w_{t,1}, w_{t,2}, \ldots, w_{t,N}\} \quad 1 \leq t \leq T \] (9)
is the weighting matrix. The different methods to define the weighting matrix will be discussed in the next section. Then, in the iterative model updating using probabilistic descent, we have
\[ \lambda(\tau + 1) = \lambda(\tau) - \epsilon_{\tau} \frac{\partial D_w(c, \bar{c})}{\partial \lambda} \bigg|_{\lambda=\lambda(\tau)} \] (10)
where \( \tau \) is the number of iteration and \( \epsilon_{\tau} \) is the updating step size for the \( \tau \)-th iteration.

The derivative of generation error with respect to model parameters can be derived from Eq.(1) and (8) as
\[ \frac{\partial D_w}{\partial \mu_q} = 2 \Sigma_q^{-1} W R_q^{-1} (\bar{c} - c), \] (11)
\[ \frac{\partial D_w}{\partial \Sigma_q^{-1}} = 2(\mu_q - W \bar{c})(\bar{c} - c)^T \Gamma R_q^{-1} W^T. \] (12)
Therefore, the model updating formulas are
\[ \mu_q(\tau + 1) = \mu_q(\tau) - 2 \epsilon_{\tau} \Sigma_q^{-1}(\tau) W R_q^{-1} \Gamma (\bar{c} - c), \] (13)
\[ \Sigma_q^{-1}(\tau + 1) = \Sigma_q^{-1}(\tau) - 2 \epsilon_{\tau}(\mu_q(\tau) - W \bar{c})(\bar{c} - c)^T \Gamma R_q^{-1} W^T. \] (14)

Eq.(13) and (14) are much simpler than the updating formulas in MGE-LSD [5]. Thus, the proposed method has much lower computation complexity than the MGE-LSD training.

4. WEIGHTING METHODS FOR LSP DISTANCE

4.1. Inverse harmonic mean weighting

The inverse harmonic mean weighting (IHWM) method [8] was proposed according to the property that if the LSP parameters of adjacent dimensions within a frame come close, its speech spectrum would have a strong resonance near the frequency band of these adjacent LSP dimensions [7]. Considering the high sensitivity of subjective perception to the distortion of spectral band with strong resonance, the IHWM method assigns a large weighting coefficient to the adjacent LSP dimensions which have close values as
\[ w^{(ihm)}_{t,i} = s_i^2 \left( \frac{1}{c_{t,i} - c_{t,i-1}} + \frac{1}{c_{t,i+1} - c_{t,i}} \right), \]
\[ c_{t,0} = 0, c_{t,N+1} = \pi, i = 1, \ldots, N \] (15)
where \( s_i \) is set manually to enhance the weight for low LSP dimensions [8].

4.2. Inverse variance weighting

It was proved that for small variation between two LSP vectors the log spectrum distortion in Eq.(6) can be approximated by a weighted Euclidean distance of LSP with the inverse of the feature variances as weights [9], which is
\[ d_L(c_t, \bar{c}_t) \approx \beta \sum_{i=1}^{N} (\bar{c}_{t,i} - c_{t,i})^2 / \sigma_{\gamma}^2 \] (16)
where \( \beta \) is a constant value and \( \sigma_{\gamma}^2 \) is the variance of the \( \gamma \)-th dimension LSP parameter [9]. Hence, in the inverse variance weighting (IVW) method we get the weighting coefficient as
\[ w^{(ivw)}_{t,i} = \beta / \sigma_{\gamma,i}^2. \] (17)
Thus \( w^{(ivw)}_{t,i} \) is a const value with respect to frame \( t \).

4.3. Gardner weighting

Comparing with IVW which adopts consistent weights for all frames, the Gardner weighting (GW) method [10] approximates the LSP more accurately using a weighted Euclidean distance between LSP vectors with different weights for each frame. Here, the weighting coefficients are defined as
\[ w^{(gw)}_{t,i} = g_{t,i}, \] (18)
\[ (g_{t,i})_{N \times N} = 4 \beta J^T(c_t) R_H(c_t) J(c_t) \]
where \( R_H(c_t) \) is the autocorrelation matrix of the impulse response of the synthesis filter \( H_{c_t}(\omega) \); \( J(c_t) \) is the Jacobian matrix transforming LSP to LPC coefficients; \( \beta \) is a constant value [10].
4.4. Formant bounded weighting

Formant bounded weighting (FBW) [11] is an integration of IHMW and GW methods. In IHMW, the weighting coefficients are designed to enlarge the distortion on the frequency band with formants. While the GW method uses weighted Euclidean distance on LSP to approximate the LSD, FBW takes both of these two aspects into account [11]: it calculates the distance between the adjacent dimensions of $N$-order and $(N-1)$-order LSP parameters to estimate the frequency band with spectral resonance; in order to approximate the LSD with weighted LSP distance, the correlation between the mean weighting coefficient vectors given by FBW and GW is used to guide the calculation of weights in FBW. Therefore, we have

$$w_{t,i}^{(fbw)} = s_t^2 \left( \frac{1}{c_{t,i}^{(N)}} - \frac{1}{c_{t,i-1}^{(N-1)}} + \frac{1}{c_{t,i}^{(N-1)}} - \frac{1}{c_{t,i}^{(N)}} \right),$$

where $c_{t,i}^{(N)}$ denotes the $i$-th dimension of $N$-order LSP parameters at frame $t$; $s_t$ is the same as that in Eq. (15): $\alpha_i = \{\alpha_1, ..., \alpha_K\}$ is optimized to maximize the correlation coefficient between vector

$$w_{t,i}^{(fbw)} = \left[ \bar{w}_{1,i}^{(fbw)}, \bar{w}_{2,i}^{(fbw)}, ..., \bar{w}_{K,i}^{(fbw)} \right]^T,$$

and

$$w_{t,i}^{(gw)} = \left[ \bar{w}_{1,i}^{(gw)}, \bar{w}_{2,i}^{(gw)}, ..., \bar{w}_{K,i}^{(gw)} \right]^T,$$

$$\bar{w}_{t,i}^{(gw)} = \frac{1}{T} \sum_{t=1}^{T} w_{t,i}^{(gw)}, i = 1, ..., K;$$

K is commonly set to $2$ [11], while we use $K = N$ in our implementation to increase the proportion of LSD approximation in FBW.

5. EXPERIMENTS

5.1. Experimental conditions

A female Chinese speech database containing 1,000 phonetically balanced sentences was used in our experiments. The sample rate of recorded speech waveforms was 16kHz. 950 sentences were selected randomly from the database for model training, and the remaining ones were used as a development set to control the number of iteration in MGE training.

The acoustic features used for model training included F0 and spectral parameters, which are 40-order LSP and an extra gain dimension derived from the spectral envelope provided by STRAIGHT analysis. The frame shift was set to 5ms. A 5-state left-to-right HMM structure with no skips was adopted to train context-dependent phone models. Here, the context features and the question sets in decision-tree-based model clustering were designed considering the characteristics of Chinese. MGE training was conducted only for the model parameters of spectral features and the output of ML training was used as the initial model for the iterative updating. After each iteration in MGE training, the reduction of LSD between natural LSP and generated LSP on the development set was calculated and the iterative updating stopped if the reduction of LSD was smaller than a threshold. In order to get a reliable calculation of the weight matrix, the natural, not the generated LSP parameters are used in Eq. (15)-(20) in our implementation. That also means it is not necessary to update the weighting matrix in the iterative model updating of MGE training. And the weighting matrix will be calculated only once for each weighting method, which can be implemented before modeling training procedure, and has no relationship with updating of model parameters. Therefore, the MGE training using weighted Euclidean distance (MGE-WED) will not increase the complexity of conventional MGE training using Euclidean distance (MGE-ECD) significantly, and will have much less complexity than MGE-LSD which will re-calculate LSD and gradient for each frame in each iteration. Four models using the different weighting methods described in Section 4 were trained. We also trained the models by ML, MGE-ECD [4] and MGE-LSD [5] for comparison.

5.2. Objective evaluation

The average LSDs between the generated and natural speech spectrums on the development set for different model training methods are compared in Fig.1. It shows that MGE-GW achieves the smallest LSD after MGE training among the four MGE-WED methods and its performance is almost equal to MGE-LSD. This is reasonable because the definition of weighting matrix in GW method is to get an accurate approximation to LSD. In Fig.1, the LSDs of MGE-IHW and MGE-FBW are slightly higher than MGE-GW because the approximation of LSD in IVW is not as accurate as that in GW and FBW method can only partly represent the LSD. Different from other MGE-WED methods, the LSD of MGE-IHMW is higher than MGE-ECD because the IHMW method is designed based on the sensitivity of subjective perception, not the approximation to LSD.

5.3. Subjective evaluation

Firstly, a subjective evaluation was conducted to compare the performance of MGE-GW and MGE-FBW which are the two best ones among the four MGE-WED methods in our objective evaluation. 20 sentences out of the database were synthesized using these two models respectively. For each sentence, subjects, including 5 Chinese native listeners, were presented a pair of synthetic speech from these two models in random order, and asked which speech sounded more natural. Fig.2 shows the subjective preference scores with 95% confidence interval for the two methods. We can see that naturalness of MGE-FBW is better than that of MGE-GW significantly because MGE-FBW takes the property of subjective perception into account besides LSD approximation in the definition of weighting matrix. This also demonstrates the inconsistency between the objective LSD measure and the subjective perception on naturalness.
Furthermore, the MGE-FBW method was compared with the ML and MGE-ECD training methods subjectively. Another 20 test sentences out of the training and development sets were synthesized using these three methods respectively. 5 Chinese-native listeners participated in this test. They were presented pairs of synthetic speech in random order and asked which one sounded better. Fig. 3 shows the preference scores with 95% confidence interval for these three methods. We can see that the naturalness of synthetic speech using MGE-ECD criterion is better than that of ML training significantly and weighted Euclidean distance measure on LSP can improve the subjective performance further.

Another subjective evaluation was conducted to compare the performance of MGE-FBW and MGE-LSD methods. 20 test sentences were synthesized and 5 listeners took part in the preference test. The result is shown in Fig. 4 and the difference is not statistically significant.

### 6. CONCLUSION

In this paper, the weighted Euclidean distance measure on LSP is introduced into the MGE training for HMM-based speech synthesis. Four weighting matrix calculation methods that are originally proposed for VQ of LSP in speech coding are compared within the framework of MGE training. The results of objective and subjective evaluation shows that formant bounded weighting (FBW) is the best among the four weighting method we investigate. The naturalness of synthetic speech after MGE-FBW training gets significantly improved comparing with the conventional MGE training using Euclidean LSP distance. Besides, the subjective performance of MGE-FBW training is not worse than that of MGE-LSD and the computation complexity of MGE-FBW training is much lower.

### 7. REFERENCES


