MINIMUM UNIT SELECTION ERROR TRAINING FOR HMM-BASED UNIT SELECTION SPEECH SYNTHESIS SYSTEM

Zhen-Hua Ling, Ren-Hua Wang

iFlytek Speech Laboratory
University of Science and Technology of China, Hefei, Anhui, P.R.China
zhling@ustc.edu, rhw@ustc.edu.cn

ABSTRACT

This paper presents a minimum unit selection error (MUSE) training method for HMM-based unit selection speech synthesis system, which selects the optimal phone-sized unit sequence from the speech database by maximizing the combined likelihood of a group of trained HMMs. Under MUSE criterion, the weights and distribution parameters of these HMMs are estimated to minimize the number of different units between the selected phone sequences and the natural phone sequences for the training sentences. The optimization is realized by discriminative training using generalized probabilistic descent (GPD) algorithm. Results of our experiment show that this proposed method is able to improve the performance of the baseline system where model weights are set manually and distribution parameters are trained under maximum likelihood criterion.

Index Terms— Speech synthesis, unit selection, HMM, minimum unit selection error, discriminative training

1. INTRODUCTION

At current stage, unit selection and waveform concatenation synthesis [1] and HMM-based parametric synthesis [2] are two main speech synthesis methods. Each of these two methods has its advantages. For unit selection and waveform concatenation method, the original waveforms are preserved and better naturalness can be obtained especially given a large database. On the other hand, HMM-based parametric synthesis provides better smoothness, robustness, flexibility and automation in system building.

In order to integrate the advantages of these two methods, an HMM-based unit selection speech synthesis system was proposed in our previous work [3,4] and satisfactory performance was achieved. In this method [4], following a Kullback-Leibler divergence based unit pre-selection, the optimal candidate phone sequence was searched out from the speech database by maximizing the combined likelihood of a group of HMMs, including spectral model, F0 model, phone duration model and so on. These models are trained under maximum likelihood criterion. Then the waveforms of selected unit sequence are concatenated to produce synthesized speech. The advantage of this method over conventional unit selection method is that statistical criterions are introduced into the calculation of target and concatenation cost, so the synthesis system can be more robust with little human intervention during system building.

However, there are still two problems with the training of this system. First, the weights for combining different models can not be trained automatically under the maximum likelihood criterion which is designed to train the parameters of different models separately. Second, there is no obvious consistency between ML criterion used in model training and the purpose of a unit selection synthesis system. In order to solve these two problems, we try to introduce some objective criterions into model training that are able to evaluate the overall performance of a unit selection system on the training set. Here, we start our work from the simplest one that is to evaluate the synthesized speech by counting how many phones in the selected unit sequence are different from the natural sequence when synthesizing a sentence in the training database. Then model weights and parameters are estimated to minimize such unit selection error. Similar discriminative training method for MCE criterion [5] in speech recognition is adopted here to realize the optimization of model weights and parameters.

This paper is organized as follows. In section 2, the baseline HMM-based unit selection system and ML-based model training is described. Section 3 introduces the discriminative training method under MUSE criterion. Section 4 and 5 are the experiment and conclusion.

2. BASELINE SYSTEM

2.1. HMM-based unit selection

In our previous work [4], the likelihoods of a group of trained HMMs are combined with some weights to guide the phone-sized unit selection as shown in Eq. (1).

$$\boldsymbol{U}^* = \arg\max_{U} \sum_{m=1}^{M} w_m \log P\left(X(\boldsymbol{U},m), q(\boldsymbol{U},m) \mid \Lambda_m, F\right) \quad (1)$$

where $\boldsymbol{U} = (u_1, ..., u_N)$ denotes a candidate phone sequence for the input sentence with N phones and $\boldsymbol{U}^*$ is the optimal one; M represents the number of phone HMM sets for different features used in our system; Λ_m and w_m mean the m-th HMM set and its weight; F denotes the contextual information of a input sentence given by the result of text analysis, which is used to decide the sentence HMM from the trained model sets; $X(\boldsymbol{U}, m)$ and $q(\boldsymbol{U}, m)$ denotes the feature sequence and state sequence of $\boldsymbol{U}$ that correspond to the m-th model. For example, if Λ_1 presents the HMM set for mel-cepstrum parameters, then $X(U, 1)$ means each frame's mel-cepstrum of unit sequence $\boldsymbol{U}$ and $q(\boldsymbol{U}, 1)$ means the state that each frame's mel-cepstrum features belongs to. In Eq.(1) only one state path is used to calculate the HMM likelihood in order to simplify the computation. Eq.(1) can be further rewritten into traditional format of the sum of *target cost* and *concatenation cost* [4]. Then dynamic programming search can be applied to select the optimal unit sequence. In order to reduce the computation cost of unit selection, a KLD-based unit pre-selection method [4] is carried out before the DP search.

2.2. ML-based model training

In our system, five different HMM sets for contextual dependent phones are used. They are spectral model, F0 model, phone duration model, concatenative spectral model and concatenative F0 model respectively. These models are trained under maximum likelihood criterion and the weights among them are set manually.

The spectral model and F0 model are trained as different streams in a unified acoustic model. At first, acoustic features are extracted from the speech waveforms of training database. STRAIGHT [6] is used to analyze the spectral envelop and F0 from the waveform of training database. Then mel-cepstrums are derived from the STRAIGHT spectrum for each frame. The final feature vector consists of static, delta and delta-delta components of mel-ceptrums and logarithmized F0. A set of contextual dependent HMMs are estimated according to the acoustic features and label information of the training database under maximum likelihood criterion. The spectral features and F0 are treated as different streams in model training and multi-space probability distribution (MSD) [7] is used to describe the F0 streams. A decision tree based model clustering method is applied after contextual dependent HMM training to improve the robustness of estimated models.

After training of acoustic model, each utterance in the training database is segmented into phones by Viterbi alignment. Based on the results of segmentation, contextual dependent phone duration model, concatenative spectral model and concatenative F0 model are trained with the same decision tree clustering method as acoustic model training. The feature of concatenation model is defined as the differential of mel-cepstrum and F0 between the first frame of current phone and the last frame of previous phone. MSD is also used for the concatenative F0 model.

3. DISCRIMINATIVE TRAINING UNDER MUSE CRITERION

3.1. Optimization criterion

Here we propose the minimum unit selection error (MUSE) training for the following two reasons:

1) To realize fully automatic training. In our baseline system, the model weights w_m can not be estimated because these models are trained under ML criterion separately with different features. Therefore we need a criterion that is able to give a unified evaluation for all models to estimate these model weights automatically.
2) To improve the consistency between the model training criterion and the purpose of a unit selection system that is to provide as much similarity as possible between synthesized speech and the natural one.

A simple evaluation of such similarity is to count how many phone units are different between the selected unit sequence and the natural unit sequence when synthesizing sentences in the speech corpus. We define such difference as unit selection error and expect to optimize model weights and parameters to minimize this error.

The discriminative training method for MCE [5] training in speech recognition is followed to realize the optimization of MUSE training. At first, the *discriminant function* is defined as Eq.(2) for a training sentence with contextual description F given $\boldsymbol{\Lambda}$ and $\boldsymbol{W}$

$$g(F, \boldsymbol{U}; \boldsymbol{\Lambda}, \boldsymbol{W}) = \sum_{m=1}^{M} w_m \log P\left(X(\boldsymbol{U}, m), q(\boldsymbol{U}, m) \mid \Lambda_m, F\right) \quad (2)$$

and Eq. (1) can be rewritten as

$$\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} g(F, \boldsymbol{U}; \boldsymbol{\Lambda}, \boldsymbol{W}) \quad (3)$$

where $\boldsymbol{\Lambda} = \left(\Lambda_1, ..., \Lambda_M\right)$, $\boldsymbol{W} = (w_1, ..., w_M)$ denote all model parameters and weights. In order to describe the decision process as Eq.(3) in a functional form, a *misclassification measure* is introduced

$$d(F; \boldsymbol{\Lambda}, \boldsymbol{W}) = -g(F, \boldsymbol{U}^{(0)}; \boldsymbol{\Lambda}, \boldsymbol{W}) + \log\left\{\frac{1}{r_{\max}} \sum_{r=1}^{r_{\max}} \exp[g(F, \boldsymbol{U}^{(r)}; \boldsymbol{\Lambda}, \boldsymbol{W}) \bullet \eta]\right\}^{1/\eta} \quad (4)$$

where $\boldsymbol{U}^{(r)} = (u_1^{(r)}, ..., u_N^{(r)})$ is the r-th best candidate unit sequence for the input sentence and $\boldsymbol{U}^{(0)}$ represents the natural unit sequence for the training sentence.

$$\boldsymbol{U}^{(r)} = \arg\max_{\boldsymbol{U} \neq \boldsymbol{U}^{(0)}, ..., \boldsymbol{U}^{(r-1)}} g(F, \boldsymbol{U}; \boldsymbol{\Lambda}, \boldsymbol{W}) \quad (5)$$

$r_{\max}$ defines how many candidate unit sequence are taking into calculation for misclassification measure. In our

implementation, $r_{\max}$ is set to 1 and Eq.(4) can be simplified as

$$d(F;\boldsymbol{\Lambda},\boldsymbol{W}) = -g(F,\boldsymbol{U}^{(0)};\boldsymbol{\Lambda},\boldsymbol{W}) + g(F,\boldsymbol{U}^{(1)};\boldsymbol{\Lambda},\boldsymbol{W}) \quad (6)$$

Then the *loss function* is defined in a smoothed zero-one form as Eq.(7) to evaluate the performance of the unit selection for one sentence where γ controls the smoothness of the sigmoid function.

$$l(F;\boldsymbol{\Lambda},\boldsymbol{W}) = \frac{1}{1+e^{-\gamma d(F;\boldsymbol{\Lambda},\boldsymbol{W})}} \quad (7)$$

For a speech corpus with I sentences, the criterion of MUSE training is to minimize the overall empirical loss $L(\boldsymbol{\Lambda},\boldsymbol{W})$, which is calculated as

$$L(\boldsymbol{\Lambda},\boldsymbol{W}) = \frac{1}{I}\sum_i l(F_i;\boldsymbol{\Lambda},\boldsymbol{W}) \quad (8)$$

Actually, what Eq.(7) measures is the sentence-level string error rate, not the phone-level unit error rate as we expect. However, the result of our experiment shows that we can also get the reduction of unit selection error by minimizing the string error rate according to above criterion.

3.2. Parameter estimation using GPD

The generalized probabilistic descent (GPD) [8] algorithm is adopted here to realize the optimization of Eq.(8) by following iterative updating

$$\boldsymbol{W}(i+1) = \boldsymbol{W}(i) - \varepsilon_i \nabla l(F_i;\boldsymbol{\Lambda},\boldsymbol{W})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}(i),\boldsymbol{W}=\boldsymbol{W}(i)} \quad (9)$$

$$\boldsymbol{\Lambda}(i+1) = \boldsymbol{\Lambda}(i) - \varepsilon_i \nabla l(F_i;\boldsymbol{\Lambda},\boldsymbol{W})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}(i),\boldsymbol{W}=\boldsymbol{W}(i)} \quad (10)$$

where F_i denotes the input contextual description of the i-th sentence and ε_i is the step size of each adjustment.

3.2.1. Update of model weights

Given a training sentence, the update of model weight w_m follows

$$w_m(i+1) = w_m(i) - \varepsilon_i \left.\frac{\partial l(F_i;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial w_m}\right|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}(i),\boldsymbol{W}=\boldsymbol{W}(i)} \quad (11)$$

where

$$\frac{\partial l(F;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial w_m} = \frac{\partial l(F;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial d(F;\boldsymbol{\Lambda},\boldsymbol{W})} \bullet \frac{\partial d(F;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial w_m} \quad (12)$$

$$\frac{\partial l(F;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial d(F;\boldsymbol{\Lambda},\boldsymbol{W})} = \gamma \bullet l(F;\boldsymbol{\Lambda},\boldsymbol{W})\left(1 - l(F;\boldsymbol{\Lambda},\boldsymbol{W})\right) \quad (13)$$

$$\frac{\partial d(F;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial w_m} = -\frac{\partial g(F,\boldsymbol{U}^{(0)};\boldsymbol{\Lambda},\boldsymbol{W})}{\partial w_m} + \frac{\partial g(F,\boldsymbol{U}^{(1)};\boldsymbol{\Lambda},\boldsymbol{W})}{\partial w_m} \quad (14)$$

$$\frac{\partial g(F,\boldsymbol{U};\boldsymbol{\Lambda},\boldsymbol{W})}{\partial w_m} = \log P\left(X(\boldsymbol{U},m), q(\boldsymbol{U},m) \mid \Lambda_m, F\right) \quad (15)$$

3.2.2. Update of model parameters

For HMM parameters, only means and variance in state PDFs of each Λ_m are updated. In Eq.(1), we define

- $X(\boldsymbol{U},m) = [\boldsymbol{x}_{m1},...,\boldsymbol{x}_{mT_m}]$, where T_m is the number of frames for feature sequence $X(\boldsymbol{U},m)$;
- $\boldsymbol{x}_{mt} = [x_{mt1},...,x_{mtD_m}]^T$ and D_m is the feature dimension of the m-th HMM;
- $q(\boldsymbol{U},m) = \left((u_{m0},q_{m0}),(u_{m1},q_{m1}),...,(u_{mT_m},q_{mT_m})\right)$, where (u_{mt},q_{mt}) denotes the phone unit index and state index that the t-th frame of $\boldsymbol{U}$ belongs to.

Then the log likelihood in Eq.(1) can be rewritten as

$$\log P\left(X(\boldsymbol{U},m), q(\boldsymbol{U},m) \mid \Lambda_m, F\right) = \sum_{t=1}^{T_m}\left[\log a_{m\,u_{mt}q_{mt-1}q_{mt}} + \log b_{m\,u_{mt}q_{mt}}(\boldsymbol{x}_{mt})\right] + \log \pi_{m\,u_{m0}q_{m0}} \quad (16)$$

where $a_{mjkk'}$ is the transition probability from state k to state k' for the contextual dependent HMM of unit j that belongs to model set Λ_m; $\pi_{m\,jk}$ is the initial probability; $b_{mjk}(\boldsymbol{x}_{mt})$ is the state observation PDF and presented by a normal distribution with diagonal covariance matrix in our system

$$b_{mjk}(\boldsymbol{x}_{mt}) = \mathcal{N}(\boldsymbol{x}_{mt};\boldsymbol{\mu}_{mjk},\boldsymbol{R}_{mjk}) \quad (17)$$

where

$$\boldsymbol{x}_{mt} = \left[x_{mtl}\right]_{l=1}^{D_m}, \boldsymbol{\mu}_{mjk} = \left[\mu_{mjkl}\right]_{l=1}^{D_m}, \boldsymbol{R}_{mjk} = \left[\sigma^2_{mjkl}\right]_{l=1}^{D_m} \quad (18)$$

Similar to the update of model weights, the update of the means and variances follows

$$\mu_{mjkl}(i+1) = \mu_{mjkl}(i) - \varepsilon_i \left.\frac{\partial l(F_i;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial \mu_{mjkl}}\right|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}(i),\boldsymbol{W}=\boldsymbol{W}(i)} \quad (19)$$

$$\frac{\partial g(F,\boldsymbol{U};\boldsymbol{\Lambda},\boldsymbol{W})}{\partial \mu_{mjkl}} = w_m \sum_{t=1}^{T_m}\left[\delta(u_{mt}-j)\delta(q_{mt}-k)\frac{x_{mtl}-\mu_{mjkl}}{\sigma^2_{mjkl}}\right] \quad (20)$$

$$\sigma^2_{mjkl}(i+1) = \sigma^2_{mjkl}(i) - \varepsilon_i \left.\frac{\partial l(F_i;\boldsymbol{\Lambda},\boldsymbol{W})}{\partial \sigma^2_{mjkl}}\right|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}(i),\boldsymbol{W}=\boldsymbol{W}(i)} \quad (21)$$

$$\frac{\partial g(F,\boldsymbol{U};\boldsymbol{\Lambda},\boldsymbol{W})}{\partial \sigma^2_{mjkl}} = w_m \sum_{t=1}^{T_m}\left\{\frac{1}{2}\delta(u_{mt}-j)\delta(q_{mt}-k)\left[\frac{(x_{mtl}-\mu_{mjkl})^2}{\sigma^4_{mjkl}} - \frac{1}{\sigma^2_{mjkl}}\right]\right\} \quad (22)$$

where $\delta(\cdot)$ denotes the Kronecker delta function.

For F0 model and concatenative F0 model with MSD, only the distribution parameters of the voiced space are updated.

4. EXPERIMENTS

4.1. Experiment conditions

A Chinese speech database containing 1000 phonetically balanced sentences pronounced by a professional female speaker was used in our experiment. The total size of the waveform is 266MB (16kHz sampled, 16bits PCM). Speech signal was analysis at 5 ms frame shift and the mel-cepstrum order was 13 (including 0-order). 5-state left-to-right with no skip HMM structure was adopted. In Chinese each syllable has a [Consonant]+Vowel+[Nasal] structure and can be split into *initial* part ([Consonant]) and *final* part (Vowel+[Nasal]). Each *initial*/*final* is treated as a phone unit in our experiment. The training of the baseline system followed the introduction of section 2 and the model weights w_m are all set to 1.

In MUSE training, 800 sentences are selected randomly from the database for training and the remaining sentences are used as testing set for unit selection error rate evaluation. The initial values of model weights are all set to 1, which is the same as the baseline system. The clustered contextual dependent HMMs given by ML training in the baseline system are used as the initial model for mean and variance updating according to Eq.(19)-(22). 10 iterations using the training set are carried out and the convergence of unit selection error rate on the training set and testing set is show in Fig. 1. The error rate is calculated as the number of unit selection error divided by the number of all synthesized phone units.

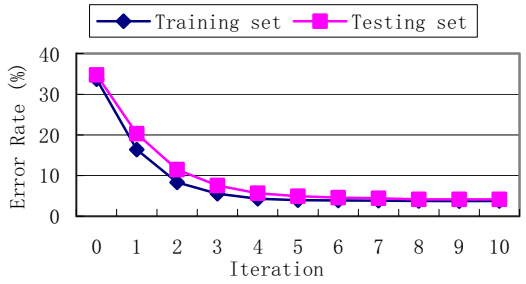

Figure 1: The convergence of unit selection error rate for MUSE training on the training set and testing set

4.2. Subjective evaluation

40 sentences out of the training set were synthesized by the baseline system and MUSE training system. These sentences were tested by 5 listeners pair by pair. The listener was asked to tell which sentence in each pair is better. Then we calculated the preference score of these two systems by collecting the evaluation of all listeners.

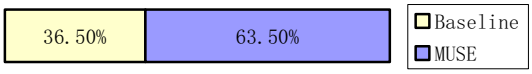

Figure2: The preference score of subjective evaluation for the baseline and MUSE system

The result is shown in Fig.2 which reflects the improvement of MUSE training compared with the baseline system.

5. CONCLUSION

This paper makes some preliminary exploration in adopting task specific criterion for unit selection synthesis into the model training of a statistical model based unit selection speech synthesis system. A discriminative training method under MUSE criterion is proposed which helps us to estimate the model weights that can not be estimated under ML criterion and improve the performance of synthesized speech. However, MUSE is still far from an ideal criterion to evaluate the performance of a unit selection system. To design more reasonable criterion and to improve the optimization method will be the goals of our future work.

6. ACKNOWLEDGEMENTS

This work was partially supported by Hi-Tech Research and Development Program of China (Grant No. 2006AA01Z137, 2006AA010104) and National Natural Science Foundation of China (Grant No. 60475015).

7. REFERENCES

[1] Hunt, A.J.,and Black, A.W., "Unit selection in a concatenative speech synthesis system using large database", *Proc. of ICASSP*, 1996, pp.373-376.

[2] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. of Eurospeech*, 1999, vol. 5, pp. 2347-2350.

[3] Zhenhua Ling, and Renhua Wang, "HMM-based Hierarchical Unit Selection Combining Kullback-Leibler Divergence with Likelihood Criterion", *Proc. of ICASSP*, 2007, pp. 1245-1248.

[4] Zhenhua Ling, Long Qin, Heng Lu, Yu Gao, Lirong Dai, Renhua Wang, Yuan Jiang, Zhiwei Zhao, Jinhui Yang, Jie Chen, and Guoping Hu,, "The USTC and iflytek speech synthesis systems for Blizzard Challenge 2007", *Blizzard Challenge Workshop*, 2007.

[5] Juang B, Chou W, and Lee C., "Minimum classification error rate methods for speech recognition", *IEEE Transactions on Speech and Audio Processing*, 1997, vol. 5, pp.257-265.

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

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

[8] Blum, R.J., "Multidimensional stochastic approximation method", *Ann. Mat. Stat.*, 1954, vol. 25, pp. 737-744.