HMM-BASED HIERARCHICAL UNIT SELECTION COMBINING KULLBACK-LEIBLER DIVERGENCE WITH LIKELIHOOD CRITERION

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 hidden Markov model (HMM) based unit selection method using hierarchical units under statistical criterion. In our previous work we tried to use frame sized speech segments and maximum likelihood criterion to improve the performance of traditional concatenative synthesis system using phone sized units and cost function criterion. In this paper, hierarchical units which consist of phone level units and frame level units are adopted to achieve better balance between the coverage rate of candidate unit and the number of concatenation points during synthesis. Besides, Kullback-Leibler divergence(KLD) between candidate and target phoneme HMMs is introduced as a part of the final criterion for unit selection. The listening result proves that these two approaches can improve the performance of synthetic speech effectively.

Index Terms— Speech Synthesis, HMM, KLD

1. INTRODUCTION

The HMM-based speech synthesis method has made significant progress in the last decade [1-3]. In this method, a unified framework of hidden Markov model is used to model the spectrum, pitch and duration simultaneously [1]. During synthesis, the speech parameters are generated from HMMs using dynamic features [2] and sent to a parametric synthesizer to reproduce speech signal. This method has been proved to be able to synthesize highly intelligible and smooth speech flexibly but its speech quality suffers from the unnatural output of parametric synthesizer greatly. So a HMM-based unit selection method has been proposed in our previous work [4], where frame is used as the base unit and the likelihood of sentence HMM is used as the criterion to guide unit selection. Experiment has proved that this method achieves better performance than state sized and traditional cost function based system using our 1000 sentences database. However, there still exist several problems for this method:

1) The discontinuities and noises caused by unit concatenation decrease the quality of synthesized speech.
2) Because only neighboring three frames are considered during dynamic programming search for unit selection, long-term smoothness can not be promised.
3) The complexity of unit selection is too high for practical application.
4) There is no explicit evidence to prove that the best perceptual performance can be achieved by only maximizing the likelihood of sentence HMM. So the optimal criterion for HMM based unit selection still needs further investigation.

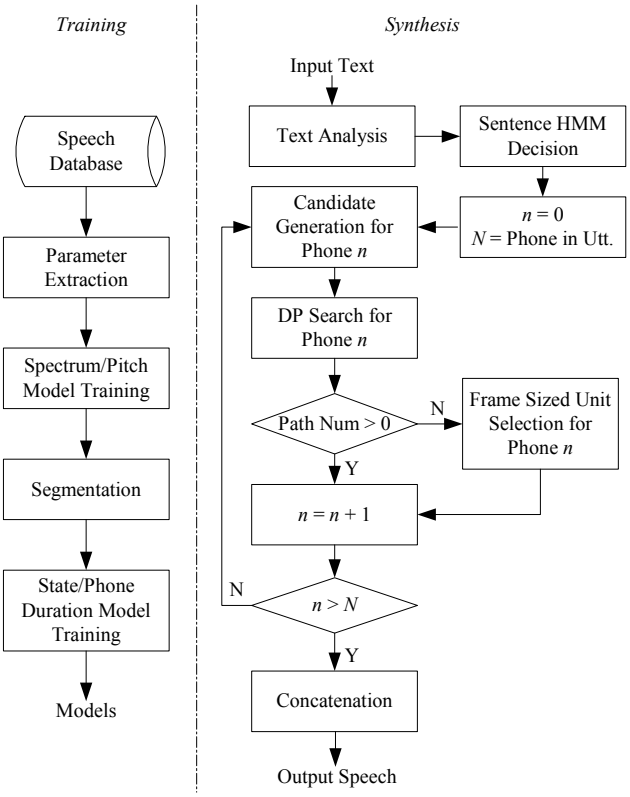

Figure 1: Flowchart of proposed method

In this paper, some new alternatives are proposed. First, hierarchical units are introduced instead of frame sized base units. At the higher level, phones are used as the base unit for unit selection to reduce the search space and provide better long-term smoothness inside a phone. When no appropriate phone candidates or dynamic programming paths can be found for a certain target phone, frames are used as the lower level base units to synthesize this phone. Compared with only phone sized base unit or only frame sized base units, hierarchical units can achieve better balance among continuity, flexibility and

complexity. Second, the Kullback-Leibler divergence is combined with likelihood as the final criterion for phone level unit selection. The selected phone sequence is required to maximize the model likelihood and minimize the KL divergence between candidate and target phone HMMs at the same time. KL divergence has shown its effectiveness as a cost measurement that can be generated automatically for unit selection [5]. Compared with likelihood, it can give better description for the similarity between the contextual factors of candidate and target units. In our experiment, the effects of combining KL divergence with likelihood criterion were tested. The flowchart of proposed method is show in Fig. 1.

This paper is organized as follows. Section 2 introduces the details of proposed method. Section 3 presents the experiments and related results. Section 4 is the conclusion.

2. METHOD

2.1. Model Training

In training stage, a set of context-dependent HMMs are estimated according to the acoustic features and label information of training database. The feature vector is composed of spectrum part and F0 part. The spectrum part consists of mel-cepstrums [6], their delta and delta-delta coefficients and is modeled by a continuous probability distribution. The F0 part consists of a logarithm of F0, its delta and delta-delta coefficients and is modeled by multi-space probability distribution (MSD) [7]. A decision tree based model clustering technique is applied after contextual dependent HMM training to improve the robustness of estimated models. During training, the state transition probability matrices for all contextual dependent HMMs with the same monophone label are tied. Then each sentence in the speech database is segmented into states using the trained HMMs. At last, the state duration model and phone duration model [3] are trained.

2.2. Phone Level Unit Selection

2.2.1. Unit Selection Combining Likelihood and KLD

Assuming the number of phones in the sentence for synthesis is N. For phone n, $n = 1,...,N$, the contextual dependent acoustic model and phone duration model determined by clustered HMMs and decision trees are λ_n and λ_n^{dur}. One candidate unit for phone n is $\boldsymbol{u}_n = \{u_{n,1},...,u_{n,T_n}\}$, where $u_{n,i}$ is the i th frame of unit $\boldsymbol{u}_n$ and T_n is the length of $\boldsymbol{u}_n$. The corresponding acoustic model of candidate unit $\boldsymbol{u}_n$ is λ_n^c. $\boldsymbol{o}_n = \{\boldsymbol{o}_{n,1},...,\boldsymbol{o}_{n,T_n}\}$ presents the acoustic feature vectors of unit $\boldsymbol{u}_n$ which consist of static and dynamic features for each frame. The dynamic features of current frame are calculated using the static features of previous, current and next frames [4]. For a whole sentence, the phone candidate sequence can be written as $\boldsymbol{u} = \{\boldsymbol{u}_1,..., \boldsymbol{u}_N\}$ and the optimal one $\boldsymbol{u}^*$ is determined using Eq.(1),

$$\boldsymbol{u}^* = \arg\max_{\boldsymbol{u}} \sum_{n=1}^{N} [LL(\boldsymbol{u}_n, \lambda_n) - KLD(\lambda_n^c, \lambda_n)] \quad (1)$$

$$LL(\boldsymbol{u}_n, \lambda_n) = \log P(\boldsymbol{o}_n \mid \lambda_n, Q_n) + \log P(T_n \mid \lambda_n^{dur}) \quad (2)$$

where $LL(\boldsymbol{u}_n, \lambda_n)$ measures the likelihood of unit $\boldsymbol{u}_n$; $KLD(\lambda_n^c, \lambda_n)$ measures the KL divergence between candidate and target phone models, which will be discussed in the next section. Ignoring the influence of state transition probability and assuming that the state allocation Q_n for unit $\boldsymbol{u}_n$ is the same as the alignment between $\boldsymbol{u}_n$ and λ_n^c which is given by segmentation in training stage, Eq.(1) can be rewritten as Eq.(3) with some weights for different components.

$$\begin{aligned} \boldsymbol{u}^* = \arg\min_{\boldsymbol{u}} \sum_{n=1}^{N} [& W_{cmp} \bullet \frac{T_n^p}{T_n} \bullet \sum_{i=1}^{T_n} (\boldsymbol{o}_{n,i} - \boldsymbol{m}_{n,i})^T \boldsymbol{\Sigma}_{n,i}^{-1} (\boldsymbol{o}_{n,i} - \boldsymbol{m}_{n,i}) \\ & + W_{dur} \bullet \frac{(T_n - m_n^{dur})^2}{\sigma_n^{dur\,2}} + W_{kld} \bullet KLD(\lambda_n^c, \lambda_n)] \end{aligned} \quad (3)$$

where the likelihood of acoustic model is normalized by the candidate phone duration T_n and predict phone duration T_n^p; $\boldsymbol{m}_{n,i}$ and $\boldsymbol{\Sigma}_{n,i}$ are the mean vector and covariance matrix for the observation Gaussian PDF of frame i in $\boldsymbol{u}_n$ decided by λ_n and Q_n; $\lambda_n^{dur} = \mathcal{N}(m_n^{dur}, \sigma_n^{dur\,2})$; W_{cmp}, W_{dur} and W_{kld} are some weights that are set manually. In order to facilitate unit search progress, Eq.(3) can be converted to the traditional form of a sum of "target cost" and "concatenation cost" as Eq.(4) considering the calculation of dynamic features for frames at phone boundaries.

$$\boldsymbol{u}^* = \arg\min_{\boldsymbol{u}} \{\sum_{n=1}^{N} TC(\boldsymbol{u}_n) + \sum_{n=2}^{N} CC(\boldsymbol{u}_{n-1}, \boldsymbol{u}_n)\} \quad (4)$$

where $TC(\boldsymbol{u}_n)$ and $CC(\boldsymbol{u}_{n-1}, \boldsymbol{u}_n)$ denote the target cost of unit $\boldsymbol{u}_n$ and the concatenation cost of units $\boldsymbol{u}_{n-1}$ and $\boldsymbol{u}_n$ respectively, given as

$$\begin{aligned} TC(\boldsymbol{u}_n) = & W_{cmp} \bullet \frac{T_n^p}{T_n} \bullet \sum_{i=2}^{T_n - 1} (\boldsymbol{o}_{n.i} - \boldsymbol{m}_{n,i})^T \boldsymbol{\Sigma}_{n.i}^{-1} (\boldsymbol{o}_{n,i} - \boldsymbol{m}_{n,i}) \\ & + W_{dur} \bullet \frac{(T_n - m_n^{dur})^2}{\sigma_n^{dur\,2}} + W_{kld} \bullet KLD(\lambda_n^c, \lambda_n) \end{aligned} \quad (5)$$

$$\begin{aligned} CC(\boldsymbol{u}_{n-1}, \boldsymbol{u}_n) = & W_{cmp} \bullet \frac{T_n^p}{T_n} \bullet (\boldsymbol{o}_{n,1} - \boldsymbol{m}_{n,1})^T \boldsymbol{\Sigma}_{n,1}^{-1} (\boldsymbol{o}_{n.1} - \boldsymbol{m}_{n,1}) \\ & + W_{cmp} \bullet \frac{T_{n-1}^p}{T_{n-1}} \bullet (\boldsymbol{o}_{n-1,T_{n-1}} - \boldsymbol{m}_{n-1,T_{n-1}})^T \boldsymbol{\Sigma}_{n-1,T_{n-1}}^{-1} (\boldsymbol{o}_{n-1,T_{n-1}} - \boldsymbol{m}_{n-1,T_{n-1}}) \end{aligned} \quad (6)$$

Dynamic programming search can be realized using Eq.(4)~(6). Compared with conventional definition of target cost and concatenation cost, these costs given here are derived automatically and few manual designing and tuning is necessary.

2.2.2. Calculation of KLD between HMMs

KL divergence is popularly used to measure the similarity between two probabilistic distributions. However, for two HMMs there is no closed form solution for calculating the KLD between them. One alternative way is to estimate it by sampling using Monte-Carlo methods, but it will lead to very

high complexity. Here, the upper bound of KLD between two left-to-right HMMs [8] is adopted as Eq.(7).

$$KLD(\lambda,\tilde{\lambda}) \le \sum_{i=1}^{S}\{\frac{D(\mathcal{N}(\boldsymbol{m}_i,\boldsymbol{\Sigma}_i)\,||\,\mathcal{N}(\tilde{\boldsymbol{m}}_i,\tilde{\boldsymbol{\Sigma}}_i))}{1-a_{ii}} + \frac{D(\mathcal{N}(\tilde{\boldsymbol{m}}_i,\tilde{\boldsymbol{\Sigma}}_i)\,||\,\mathcal{N}(\boldsymbol{m}_i,\boldsymbol{\Sigma}_i))}{1-\tilde{a}_{ii}} + \frac{(a_{ii}-\tilde{a}_{ii})\log(a_{ii}/\tilde{a}_{ii})}{(1-a_{ii})(1-\tilde{a}_{ii})}\} \quad (7)$$

where S is the number of states in a model; $\mathcal{N}(\boldsymbol{m}_i,\boldsymbol{\Sigma}_i)$ and $\mathcal{N}(\tilde{\boldsymbol{m}}_i,\tilde{\boldsymbol{\Sigma}}_i)$ present the observation PDF of state i for model λ and $\tilde{\lambda}$; a_{ii} and $\tilde{a}_{ii}$ present the state transition probability for λ and $\tilde{\lambda}$. Because λ and $\tilde{\lambda}$ must present the same monophone in our system and the transition probability matrix is tied, $a_{ii}=\tilde{a}_{ii}$ and Eq.(7) can be simplified as

$$KLD(\lambda,\tilde{\lambda}) \le \sum_{i=1}^{S}\frac{1}{1-a_{ii}}\{D(\mathcal{N}(\boldsymbol{m}_i,\boldsymbol{\Sigma}_i)\,||\,\mathcal{N}(\tilde{\boldsymbol{m}}_i,\tilde{\boldsymbol{\Sigma}}_i)) + D(\mathcal{N}(\tilde{\boldsymbol{m}}_i,\tilde{\boldsymbol{\Sigma}}_i)\,||\,\mathcal{N}(\boldsymbol{m}_i,\boldsymbol{\Sigma}_i))\} \quad (8)$$

For each state, the KLD between two D-dimension single mixture Gaussian distributions can be calculated as [9]

$$D(\mathcal{N}(\boldsymbol{m}_i,\boldsymbol{\Sigma}_i)\,||\,\mathcal{N}(\tilde{\boldsymbol{m}}_i,\tilde{\boldsymbol{\Sigma}}_i)) = \frac{1}{2}\ln(\frac{|\tilde{\boldsymbol{\Sigma}}_i|}{|\boldsymbol{\Sigma}_i|}) - \frac{D}{2} + \frac{1}{2}tr(\tilde{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\Sigma}_i) + \frac{1}{2}(\tilde{\boldsymbol{m}}_i-\boldsymbol{m}_i)^T\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\boldsymbol{m}}_i-\boldsymbol{m}_i) \quad (9)$$

2.2.3 Implementation

For each phone in the target sentence, all the phones in the database with the same monophone label are used as the candidate for calculating target cost according to Eq.(5). Then the best K_{phone} units are kept for dynamic programming search using the concatenation cost defined in Eq.(6). Besides, some constraints are used to guarantee the confidence of selected phone units. For any candidate unit $\boldsymbol{u}_n$, if

$$\begin{aligned} &(\boldsymbol{o}_{n,i}-\boldsymbol{m}_{n,i})^T\boldsymbol{\Sigma}_{n,i}^{-1}(\boldsymbol{o}_{n,i}-\boldsymbol{m}_{n,i}) > THRES_{cmp} \\ &or \quad (T_n-m_n^{dur})^2/\sigma_n^{dur\,2} > THRES_{dur} \\ &or \quad KLD(\lambda_n^c,\lambda_n) > THRES_{kld} \end{aligned} \quad (10)$$

this unit will be discarded. So it is possible that no candidate units or dynamic programming path can be found for a certain target phone. At this time, frame sized units are used to synthesize this phone. By modifying the thresholds in Eq.(10), the proportion between synthesized phones using phone sized units and frame sized units can be controlled.

Because multi-space probability distribution (MSD) [7] is used to model the F0 streams in our system, some related processes are carried out. First, in order to calculate the likelihood of F0 features in section 2.2.1, the voiced/unvoiced space decision is made according to the voicedness of each frame in candidate $\boldsymbol{u}_n$ and then simplify the calculation to a single space problem. Second, in KL divergence calculation, the weight of voice space for each state in the candidate and target phone models are required to be larger or smaller than 0.5 at the same time, otherwise the KLD between these two models will be set to infinite.

2.3. Frame Level Unit Selection

When no appropriate candidate units or paths can be found during dynamic programming search for a certain target phone, frame sized units are used to synthesize this phone using almost the same method proposed by our previous work [4]. The differences are that here the target for synthesis is a phone not a sentence and the candidate frames are restricted to come from the same leaf node of the decision tree for spectral model clustering as the target frame. Therefore the KLD for spectral parameters between candidate and target frames is zero. After frame level unit selection for a certain phone, the phone level unit selection goes on for the next target phone as shown in Fig.1.

2.4 Concatenation

At last, the result of hierarchical unit selection can be presented by a list of candidate frames no matter they are selected by phone or by frame. Then the same cross-fade technique for frame sized unit concatenation [10] is used to generate the speech waveform.

3. EXPERIMENTS

3.1. Experiment Conditions

The database used for HMM training and unit selection was the same as our previous work [4], which consisted of 1000 phonetically balanced Chinese sentences. 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 for each initial/final in Chinese.

For phone level unit selection, W_{cmp} was set to 1/39 for spectral part and 1/9 for F0 part; W_{dur} was set to 100; $THRES_{kld}$, $THRES_{cmp}$ and $THRES_{dur}$ were set to 20, 20 and 5; K_{phone} was set to 1000. For frame level unit selection, the same settings used in our previous frame sized unit selection system (ML_DP2_1) [4] was adopted.

For comparison purpose, ML_DP2_1 was adopted as the baseline system using only frame sized base unit. In order to test the effectiveness of combining KLD and likelihood criterion, three hierarchical unit systems with W_{kld} = 0, 5 and 1e+9 were constructed, which present likelihood-based, combined and KLD-based criterion respectively.

3.2. Complexity Evaluation

10 sentences out of the training set containing 495 phones were synthesized by the baseline system and proposed systems with different W_{kld}. After using hierarchical units, the numbers of target phones that were synthesized by phone sized unit and frame sized unit are shown in Table 1.

The compute time to real time ratio for these systems using hierarchical units is about 6~7 on our PC platform with 2.4GHz CPU. Because the setting of K_{phone} is quite high here, the

reduction of complexity is not so significant compared with the baseline system whose real time ratio is about 12.

Table 1: The number of synthesized phones using phone level units or frame level units

W_{kld}	*Total Phone Num*	*Phone Level Synthesis*	*Frame Level Synthesis*
0	495	454	41
5	495	455	40
1e+9	495	453	42

3.3. Subjective Evaluation

In our subjective evaluation, the 40 sentences synthesized by 4 systems were evaluated by 10 listeners. Each listener was required to gives two opinion scores from 1(bad) to 5 (good) for each sentence. One score is for "naturalness" which evaluates the segmental articulation and prosodic fluency of the whole sentence and the other is for "speech quality" which measures the noise and discontinuity caused by concatenation. The final average scores for each system are shown in Fig.2.

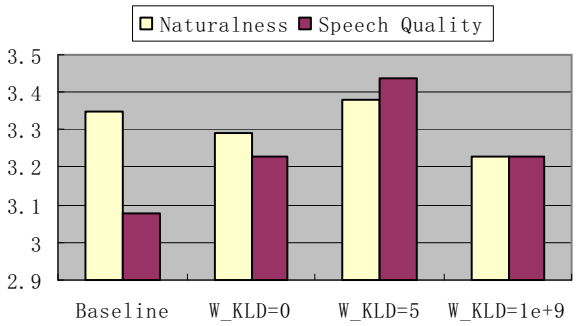

Figure 2: The subjective evaluation results between baseline system and proposed systems with different KLD weight

From these results, we can see that:

1) By introducing hierarchical units, the speech quality of synthesized speech can be improved due to the usage of more naturally continuous frames and the reduction of concatenation points. The speech quality differences between baseline system and proposed system with $W_{kld} = 5$ is significant ($p = 0.00 < 0.05$, paired t-test).
2) On the other hand, using larger base unit may decrease the naturalness of synthesized speech because the coverage rate of base unit is decreased and it is more difficult to find appropriate ones. This is demonstrated by comparing the performance of baseline system with $W_{kld} = 0$. However, the naturalness differences between these four systems are not significant ($p > 0.05$, paired t-test).
3) Compared with only likelihood criterion ($W_{kld} = 0$) or only KLD criterion ($W_{kld} =$ 1e+9), better performance can be achieved when combining them together ($W_{kld} = 5$). The speech quality differences between proposed system with $W_{kld} = 5$ and other two systems are significant ($p = 0.011$ and $0.007 < 0.05$, paired t-test).

4. CONCLUSIONS

In this paper, an HMM-based hierarchical unit selection method is proposed to improve the performance of baseline frame sized unit selection system. KLD is integrated with likelihood to give a better criterion for unit selection. The experiment results show that by using hierarchical units and combing KLD with sentence likelihood appropriately the system can achieve better performance than the baseline system with half of its complexity. However, the method adopted here to combine KLD with likelihood criterion is quite simple and the weights still needs further tuning if the optimal performance is expected. More reasonable statistical criterions for unit selection and automatic training of the thresholds and weights used in our system will be the task of our future work.

5. ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

6. REFERENCES

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

[2] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, 2000, vol. 3, pp. 1315-1318.

[3] Zhenhua Ling, Yijian Wu, Yuping Wang, Long Qin, and Renhua Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method", in *ICSLP Satellite Workshop, Blizzard Challenge*, 2006.

[4] Zhenhua Ling, and Renhua Wang, "HMM-based unit selection using frame sized speech segments", in *Proc. of ICSLP*, 2006, pp. 2034-2037.

[5] Yong Zhao, Peng Liu, Yusheng Li, Yining Chen, and Min Chu, "Measuring target cost in unit selection with KL-divergence between context-dependent HMMs", in *Proc. of ICASSP*, 2006, pp. I-725-I-728.

[6] Fukada, T., Tokuda, K., Kobayashi, T. and Imai, S., "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. of ICASSP*, 1992, vol.1, pp.137-140.

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

[8] Peng Liu, and Frank K. Soong, "Kullback-Leibler Divergence between Two Hidden Markov Models", *Microsoft Research Asia, Technical Report*, 2005.

[9] Christopher Rozell, "Information-Theoretic Analysis of Neural Responses in the Frequency Domain", *Rice University, Techinical Report*, 2003.

[10] Hirai, T., and Tenpaku, S., "Using 5 ms segments in concatenative speech synthesis," in *Proc. of 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, USA, 2004, pp. 37-42.