HMM-based Unit Selection Using Frame Sized Speech Segments

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 for concatenative speech synthesis system. Frame sized waveform segments are adopted as basic synthesis units here to increase the coverage rate of candidate units and the chance of finding appropriate ones. In training stage, a set of contextual dependent HMMs are trained with static and dynamic acoustic features. When synthesizing a sentence, the optimal frame sequence is searched out from speech corpus by maximizing the output probability of a sentence HMM constructed according to the contextual information of input text. Listening test proves that proposed method can achieve better performance of synthesized speech compared with the method using state sized units and cost function criterion.
Index Terms: speech synthesis, unit selection, HMM

1. INTRODUCTION

The Hidden Markov Model (HMM) had been widely used in speech recognition field. Meanwhile, the HMM based speech synthesis also made significant progress in the last decade[1,2]. In this method, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs[1] and the parameters are generated from HMMs by using the dynamic features[2]. Then parametric synthesizer is used to synthesize speech signal based on generated parameters. This method is able to synthesize highly intelligible and smooth speech sounds flexibly but its performance suffers from the unnatural output of parametric synthesizers greatly.

Also, HMM-based unit selection method was proposed [3-5]. In this method, speech inventory is constructed automatically, which contains state sized speech units segmented using trained acoustic HMMs. Target acoustic features of the sentence for synthesis is also predicted by the models or by other means. Then unit selection is realized by dynamic programming (DP) search to minimize an overall costing function, which is defined as a weighted sum of target costs and concatenation costs on various features, such as pitch, duration, energy and spectrum. For this method how to keep the continuity at concatenation boundaries is significant especially when using speech corpus with limited size and it is tricky to modulate the weight between target cost and concatenation cost.

This paper presents an alternative method for HMM-based unit selection speech synthesis. In this method, frame sized speech segments are used. Smaller units have shown the effectiveness in improving the coverage of candidate units and simplifying segmental concatenation in some systems[6,7]. Besides, probabilistic criterion[8] is applied here and replaces cost function to make use of not only the predicted value but also the distribution property of acoustic features at each frame provided by the sentence HMM. Listening test proves that proposed method can produce more natural speech than the method using state sized segment and cost function criterion.

This paper is organized as follows. Section 2 gives a brief overview of proposed method and the details about maximum likelihood(ML) based unit selection are described in section 3. Section 4 introduces some techniques to reduce the computation complexity and section 5 presents the results of experiment and evaluation. Section 6 is the conclusion.

2. METHOD

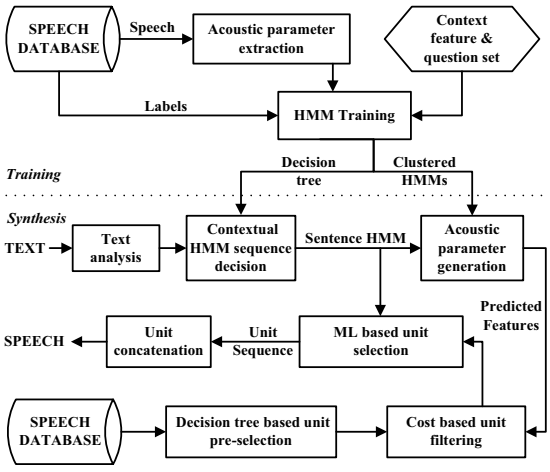

Figure 1: Flowchart of proposed method.

The flowchart of proposed method is shown in Fig.1. In training stage, a set of contextual 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[9], 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)[1]. A decision tree based model clustering technique is applied after contextual dependent HMM training to improve the robustness of estimated models. Then each sentence in the speech database is segmented into

states according to the trained HMMs to facilitate unit pre-selection which will be discussed in section 4.

During synthesis, the contextual analysis result of input text is used to decide the sentence HMM according to the clustering decision tree. Then parameter generation algorithm using dynamic features[2] is applied to predict the duration of each state and generate acoustic parameters for each frame, which are used as target values for unit filtering. The optimal sequence of frame sized speech segments is chosen from speech database to maximize the output probability of sentence HMM by searching among the candidates at each frame using multi-stage dynamic programming(DP). In order to reduce the computation complexity, unit pre-selection and filtering are carried out to decrease the number of candidates for DP search at each frame. At last, the optimal sequence of units are concatenated to produce output speech. The details about each step are discussed in the following two parts.

3. ML-BASED UNIT SELECTION

3.1 ML Criterion for Unit Selection

Assuming λ is the concatenated sentence HMM based on the decision tree and contextual information of input text, Q is the state sequence for each frame determined by state duration model[2], $u_i^{(1)}, u_i^{(2)}, ...u_i^{(K)}$ are the K candidate units for frame i, N is the total number of frames for synthesized sentence. Then the optimal sequence of candidate frames is chosen to maximize the likelihood of output probability of the sentence HMM given λ and Q.

$$M^* = \arg\max_M \log P(o(u) \mid \lambda, Q) \tag{1}$$

where $u = u_1^{(m_1)}, u_2^{(m_2)}, ...u_N^{(m_N)}$ is the candidate sequence for sentence determined by unit index path $M = [m_1, m_2, ..., m_N]$, $m_i \in [1, 2, ..., K], i = 1, ..., N$; M^* is the optimal path; $o(u) = [o(u_1^{(m_1)})^T, ..., o(u_N^{(m_N)})^T]^T$ is the observation vectors of candidate sequence u containing both static and dynamic features;

$$o(u_i^{(m_i)}) = [c(u_i^{(m_i)})^T, \Delta c(u_i^{(m_i)})^T, \Delta^2 c(u_i^{(m_i)})^T]^T \tag{2}$$

$c(u_i^{(m_i)}) = [c_1(u_i^{(m_i)}), ..., c_d(u_i^{(m_i)})]^T$ is the static acoustic feature vector of candidate unit $u_i^{(m_i)}$ for frame i, d is the dimension of static features. The dynamic features are calculated as follows

$$\Delta c(u_i^{(m_i)}) = 0.5(c(u_{i+1}^{(m_{i+1})}) - c(u_{i-1}^{(m_{i-1})})) \tag{3}$$

$$\Delta^2 c(u_i^{(m_i)}) = 0.25(c(u_{i+1}^{(m_{i+1})}) + c(u_{i-1}^{(m_{i-1})})) - 0.5c(u_i^{(m_i)}) \tag{4}$$

By introducing dynamic features, the models can describe not only the distribution property of static acoustic parameters but also their correlations among adjacent frames which benefit the continuity of selected unit sequence. The output probability of observation vector at each state of HMMs is presented by a single mixture Gaussian probability distribution function(PDF) and for each frame the PDF is known after state duration generation. Assuming the Gaussian PDF at frame i is $\mathcal{N}(U_i, \Sigma_i), i = 1, ..., N$, where U_i *and* Σ_i present the mean vector and covariance matrix respectively, then Eq. 1 can be rewritten as

$$M^* = \arg\min_M \sum_{i=1}^{N} (o(u_i^{(m_i)}) - U_i)^T \Sigma_i^{-1} (o(u_i^{(m_i)}) - U_i) \tag{5}$$

The optimal fame sequence can be chosen by dynamic programming search according to Eq.5.

3.2 Two-stage DP Search

Compared with common DP method where only the correlation between current unit and previous unit is considered, a two-stage DP algorithm is necessary here because both the feature information of previous and next candidate units are required in order to calculate $o(u_i)$ at each frame according to Eq.2-4. The whole recursive DP search process is introduced in this section.

First, define the quasi-likelihood calculated at frame *i* as

$$L_i(m_{i-1}, m_i, m_{i+1}) = (o(u_i^{(m_i)}) - U_i)^T \Sigma_i^{-1} (o(u_i^{(m_i)}) - U_i) \tag{6}$$

Assuming current frame number is I(I=3,…N), the candidate q is selected for current frame and the candidate p is selected for previous frame($m_I = q, m_{I-1} = p$), then the unit index paths and corresponding quasi-likelihood sums until frame I regarding with the combination of p and q can be described as

$$M^{I,p,q} = [m_1^{I,p,q}, ..., m_{I-2}^{I,p,q}] \quad m_i^{I,p,q} \in [1, 2, ..., K] \quad i \in [1, 2, ..., I-2] \tag{7}$$

$$L^{I,p,q} = \sum_{i=1}^{I-3} L_i(m_{i-1}^{I,p,q}, m_i^{I,p,q}, m_{i+1}^{I,p,q}) + L_{I-2}(m_{I-3}^{I,p,q}, m_{I-2}^{I,p,q}, p) + L_{I-1}(m_{I-2}^{I,p,q}, p, q) \tag{8}$$

An illustration for definition of $M^{I,p,q}$ is shown in Fig.2.

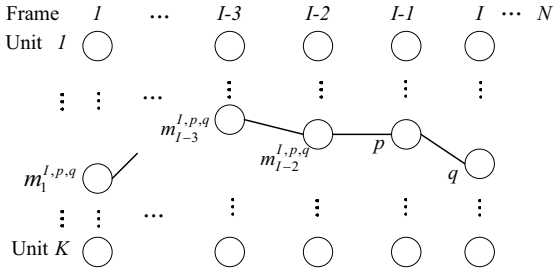

Figure 2: An example of path description for two-stage DP.

The optimal path until frame I considering candidate q for current frame and candidate p for frame I-1 is defined as

$$M^{*I,p,q} = \arg\min_{M^{I,p,q}} L^{I,p,q} \tag{9}$$

$$L^{*I,p,q} = \min_{M^{I,p,q}} L^{I,p,q} \tag{10}$$

It can be solved recursively from frame I-1 to frame I as

$$L^{*I,p,q} = \min_{m_{I-2}^{I,p,q}}(L^{*I-1,m_{I-2}^{I,p,q},p} + L_{I-1}(m_{I-2}^{I,p,q}, p, q)) \quad (11)$$

$$m_{I-2}^{*I,p,q} = \arg\min_{m_{I-2}^{I,p,q}}(L^{*I-1,m_{I-2}^{I,p,q},p} + L_{I-1}(m_{I-2}^{I,p,q}, p, q)) \quad (12)$$

$$M^{*I,p,q} = [M^{*I-1,m_{I-2}^{*I,p,q},p}, m_{I-2}^{*I,p,q}] \quad (13)$$

In order to initiate the recursion, the quasi-likelihood at beginning frame is calculated without dynamic features.

$$M^{*2,p,q} = [\,] \quad (14)$$

$$L^{*2,p,q} = (c(u_1^{(p)}) - U_1^c)^T \Sigma_1^{c-1} (c(u_1^{(p)}) - U_1^c) \quad (15)$$

where U_i^c *and* Σ_i^c are the mean vector and covariance matrix for the Gaussian PDF of only static features at frame *i*. The final search result is given as

$$[p^*, q^*] = \arg\min_{p,q}(L^{*N,p,q} + (c(u_N^{(q)}) - U_N^c)^T \Sigma_N^{c\ -1} (c(u_N^{(q)}) - U_N^c)) \quad (16)$$

$$M^* = [M^{*N,p^*,q^*}, p^*, q^*] \quad (17)$$

3.3 Unit Concatenation

The optimal sequence of frames given by ML-based DP search are concatenated using cross-fade technique following the method described in [6].

4. COMPLEXITY REDUCTION

Because of introducing two-stage dynamic programming search, the computation complexity of above ML-based unit selection method for a sentence of *N* frames is about $O(NK^3)$. Comparing frame sized units with state sized unit, the number of candidates *K* within corpus increases greatly. For example, a one-hour speech database consists of 720,000 units when frame length is set to 5ms. If all of them are used as candidates for DP search, the complexity is unacceptable. So unit pre-selection and filtering methods are used to reduce the searching space for DP and some pruning techniques are also applied.

4.1 Decision Tree Based Unit Pre-selection

In training stage, each sentence in corpus is segmented into states by Viterbi alignment using clustered contextual dependent HMMs. Before unit selection, the state and corresponding cluster that each target frame in the sentence for synthesis belongs to is known according to the state duration model and clustering decision tree. The unit pre-selection is realized by keeping the frames in the states that share the same leaf node in the clustering decision tree with target frame and discarding all the other candidates in database. A threshold N_{pre} is used and the number of candidates after pre-selection is promised to be greater than N_{pre} otherwise the leaf node is traced back to parent node to get more candidates. In our system, the feature streams for F0 and mel-cepstrums use two different clustering decision trees and here we choose the tree for mel-cepstrums in unit pre-selection.

4.2 Cost Based Unit Filtering

A further processing to reduce the search complexity is realized by cost based unit filtering after pre-selection. We calculate the target cost for each candidate and select the *K* units with minimum costs. The target cost is defined as follows

$$TC = TC_{pitch} \cdot W_{pitch} + TC_{gain} \cdot W_{gain} + TC_{spec} \cdot W_{spec} \quad (18)$$

$$TC_{pitch} = \left|\ln(f_{0cand}) - \ln(f_{0targ})\right| \quad (19)$$

$$TC_{gain} = \left|cep_{0cand} - cep_{0targ}\right| \quad (20)$$

$$TC_{spec} = \sqrt{\sum_{i=1}^{p}(cep_{i\,cand} - cep_{i\,targ})^2} \quad (21)$$

where F_{0cand} and F_{0targ} are F0 for candidate frame and target frame, cep_{icand} and cep_{itarg} are the *i*-th order mel-cepstrum for candidate unit and target unit respectively. *W* are the weights to combine the costs of different acoustic features, which are set manually. The difference between the usage of target cost here and in common cost function based unit selection method is that here the target cost is not a part of criterion for final unit selection but a pre-processing method.

4.3 Search Pruning

Two kinds of pruning strategy are attempted in DP searching:

1) One-stage DP search. The whole model training and unit selection processes are refreshed by using simplified dynamic features as

$$o(u_i^{(m_i)}) = [c(u_i^{(m_i)})^T, \Delta c(u_i^{(m_i)})^T]^T \quad (22)$$

$$\Delta c(u_i^{(m_i)}) = c(u_{i+1}^{(m_{i+1})}) - c(u_i^{(m_i)}) \quad (23)$$

The computation complexity is $O(NK^2)$ if one-stage DP is used.

2) Path pruning. After calculating $L_I^{*p,q}$ using Eq.11, for each candidate unit *p* at frame *I*, only $K'(K' < K)$ candidate units *q* at frame *I*-1 with least $L_I^{*p,q}$ are kept and the other $K - K'$ units are ignored during the following search. It can reduce the complexity to $O(NK'K^2)$. When $K' = 1$, the complexity of two-stage DP search is reduced to as much as one-stage DP search.

5. EXPERIMENTS

5.1 Experiment Conditions

The database used for HMM training consists of 1000 phonetically balanced Chinese sentences pronounced by a female speaker. There are 25,096 syllable initials and 29,942 syllable finals and the total size is 266MB (16kHz sampled, 16bits PCM). Speech signal is analysis at 5 ms frame shift and the mel-cepstrum order is 13 (including 0-order). 5-state left-to-right with no skip HMM structure is adopted for each initial/final in Chinese. Context features and question set for

decision tree clustering are designed considering the characteristic of Chinese.

The length for frame-sized synthesis unit is also 5 ms. N_{pre} is set to 4000 for pre-selection and K, W_{pitch}, W_{gain}, W_{spec} are set to 200, 10, 1 and 1 respectively for unit filtering.

In the experiment, we compare the performance of following five systems:

1) STA_CF: a HMM-based concatenative synthesis system using state-sized units with similar method described in [4].
2) FRM_CF: frame sized units are used but the unit selection is implemented under cost function criterion. The target cost is calculated using Eq.18-21. The concatenation cost is calculated based on the spectral distortion at concatenation points following the method described in [6].
3) ML_DP2_1: proposed method with 2-stage DP, $K'=1$.
4) ML_DP2_10: proposed method with 2-stage DP, $K'=10$.
5) ML_DP1: proposed method with 1-stage DP.

5.2 Evaluation Results

20 sentences which are not contained within the training set are synthesized by the above 5 systems and evaluated by 10 listeners. Each listener is required to gives an evaluation score from 1(bad) to 5 (good) for each sentence. The final mean opinion score(MOS) for these 5 systems is shown in Fig.3.

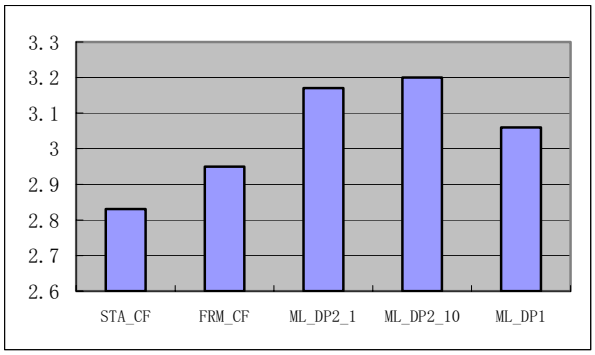

Figure 3: The MOS evaluation results.

From the result, we can see that:

1) After using smaller segments, the performance of cost function based system is improved. By examining the synthesized sentences, the discontinuity at unit boundaries for FRM_CF is less serious and gives better naturalness in perception than STA_CF although the number of concatenation points increases for the former one. It is also interesting to find that for FRM_CF many consecutive frames in the corpus sentences are searched out and the actual number of concatenation points decreases.
2) By introducing ML criterion into unit selection, the performance of synthesized speech improves further because more distribution property of static and dynamic acoustic features is taking into account besides only the predicted values.
3) It is reasonable that pruning for DP search decreases the performance. However comparing ML_DP2_1 and ML_DP1, which have the same computation complexity in unit selection, the former one achieves better performance. This indicates that ML-based unit selection can benefit from finer feature modeling, especially the relationship among consecutive frames.

5.3 Complexity Evaluation

Although some unit pre-selection and pruning techniques have been applied, the complexity of proposed method is still very high. For system ML_DP2_1, the time consumption for synthesis is about 12xRT(real time ratio) on our PC platform with 2.4GHz CPU .

6. CONCLUSIONS

In this paper, an alternative HMM based unit selection method for concatenative speech synthesis system is proposed. The frame sized speech segments are used to increase the coverage rate of candidate units and improve the discontinuity at concatenation boundaries. Besides, probabilistic criterion realized by multi-stage DP search is introduced here to replace the cost function used in common concatenation systems. In order to reduce the computation complexity, some unit pre-selection and search pruning techniques are also applied. Listening test proves that better performance of synthesized speech can be achieved by proposed method. However, the computation complexity of proposed method is still very high. How to find better unit pre-selection and pruning methods and replacing interval-fixed frames with pitch-synchronous frames or variable-length units for concatenation will be the contents of our future work.

7. ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China under grant number 60475015.

8. 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] Huang, X., Acero, A., Hon, H., Ju, Y., Liu, J., Merdith, S. and Plumpe, M., “Recent improvements on Microsoft’s trainable text-to-speech system – Whistler,” in *Proc. of ICASSP*, 1997, pp. 959-962.

[4] Donovan, R. E., “Trainable speech synthesis,” *PhD. Thesis*, Cambridge University Engineering Department, 1996.

[5] Donovan, R. E., and. Eide, E. M., “The IBM trainable speech synthesis system,” in *Proc. of ICSLP*, 1998, pp.1703-1706.

[6] 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.

[7] Kishore, S. P. and Black, A.W., “Unit size in unit selection speech synthesis,” in *Proc. of Eurospeech*, 2003, pp. 1317-1320.

[8] Sakai, S. and Shu, H., “A probabilistic approach to unit selection for corpus-based speech synthesis,” in *Proc. of Eurospeech*, 2005, pp.81-84.

[9] 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.