Decision Tree-based Clustering with Outlier Detection for HMM-based Speech Synthesis

Kyung Hwan Oh, June Sig Sung, Doo Hwa Hong, Nam Soo Kim

School of Electrical Engineering and INMC,
Seoul National University, Seoul, Korea
{khoh, jssung, dhhong}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In order to express natural prosodic variations in continuous speech, sophisticated speech units such as the context-dependent phone models are usually employed in HMM-based speech synthesis techniques. Since the training database cannot practically cover all possible context factors, decision tree-based HMM states clustering is commonly applied. One of the serious problems in a decision tree-based method is that the criterion used for node splitting and stopping is sensitive to irrelavant outlier data. In this paper, we propose a novel approach to removing outliers during the decision tree growing phase. Experimental results show that removing of outlying models improves the quality of the synthesized speech, especially for sentences which originally demonstrated poor quality.

Index Terms: HMM-based speech synthesis, decision tree-based clustering, outliers

1. Introduction

Hidden Markov model (HMM)-based parametric speech synthesis techniques have been developed over the past two decades [1] and can generate speech of acceptable quality [2] with flexible variation of speech characteristics [3, 4, 5]. In HMM-based speech synthesis techniques, spectrum, excitation, and state duration are modeled simultaneously in a unified framework [6]. In order to account for variability of the extracted features, context-dependent models are usually employed. In this work, we apply tri-phone models considering prosodic and linguistic contexts such as accent, stress, part of speech and position in a sentence to represent suprasegmental information of speech. However, it is practically impossible to prepare speech database which covers all the possible context factors. To alleviate this problem, decision tree-based HMM state clustering techniques are adopted. With this clustering procedure, we can somewhat overcome data sparseness and predict unseen context-dependent models at the synthesis stage [7].

In general, decision tree-based clustering for HMM states tying is a top-down approach. The construction of a decision tree is achieved based upon a proper node splitting criterion and stopping criterion. In conventional HMM-based speech synthesis techniques, maximum likelihood (ML) or minimum description length (MDL) criterion is used for the construction of a decision tree [8]. One of the significant drawbacks of the ML or MDL criterion is that it is sensitive to outlier data which usually results in degraded quality of the synthesized speech. If the speech database contains discrepancies such as the unexpected noise, poorly pronounced speech, mis-transcription and mis-segmentation, it makes it difficult for the tree growing algorithm to determine optimal clusters.

In this paper, we propose an algorithm to detect and remove outliers in decision tree-based clustering techniques for HMM states tying. Using a common distance-based outlier detection algorithm, we can detect outlying HMM states. By this step, only similar HMM states are merged into the same cluster, resulting in a robust decision tree. To evaluate the performance of the proposed technique, both objective and subjective tests are performed. The experimental results show that the proposed technique yields improved synthesized speech quality especially for sentences that originally demonstrated poor quality.

In the following Section 2, we briefly review the decision tree-base clustering approach in HMM-based speech synthesis techniques. Section 3 presents outlier detection method in detail which we propose to add in the construction of a decision tree. Experimental results are given in Section 4. Concluding remarks are given in Section 5.

2. Decision tree-based HMM states clustering

Fig. 1 shows an example of a decision tree where each node consists of a number of HMM states. For node splitting, we apply the MDL criterion which accounts for both the model specificity and complexity [9]. Let U denote the set of leaf nodes in a decision tree. The description length of U is given by

$$D(U) \equiv -L(U) + KM \log G + C \quad (1)$$

where $L(U)$ is the log-likelihood of the model U, K is the dimensionality of an observation vector, M is the number of leaf nodes, $G = \sum_{m=1}^{M} \Gamma_m$ with Γ_m denoting the summation of the state occupancy probabilities at node S_m and C is the code length which is here assumed to be constant. It is noted that the description length consists of a likelihood term and a penalty for model complexity. Now suppose that the node S_m is split into S_{mqy} and S_{mqn} according to a binary (yes or no) question q. U' is the set of leaf nodes obtained after splitting the node S_m into two child nodes. Let $\delta_m(q)$ represent the difference of description length before and after node splitting with the question q, i.e., $\delta_m(q) = D(U') - D(U)$. The optimal question $\widehat{q}$ is selected to minimize $\delta_m(q)$ and node splitting is stopped when $\delta_m(q) > 0$ for all the possible questions. This method provides an efficient way for clustering HMM states [9].

In (1), $L(U)$ indicates the sum of the log-likelihoods for generating observations at the nodes of U. Let $\{s_1^m, s_2^m, ..., s_{L_m}^m\}$ represent the set of HMM states merged into node S_m where L_m indicates the total number of states. The

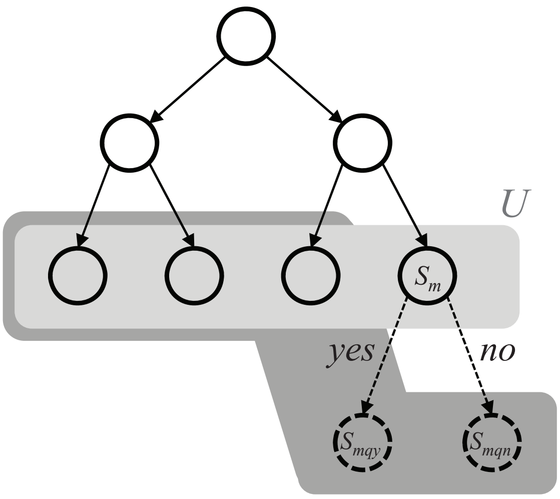


Figure 1: *Node splitting of decision tree.*

log-likelihood at node S_m is given by:

$$L(S_m) = -\frac{1}{2}\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t)[(o_t-\mu_m)'\Sigma_m^{-1}(o_t-\mu_m) \\ + K\log 2\pi + \log(det(\Sigma_m))]. \quad (2)$$

where the prime denotes the transpose of a vector or a matrix, o_t is the observation vector at time t, T_l denotes the number of data frames for state s_l^m, $\gamma_l^m(t)$ denotes the a posteriori probability of the state s_l^m at the t-th frame and μ_m and Σ_m are the mean vector and covariance matrix of the Gaussian distribution at node S_m, respectively.

Let μ_l^m and Σ_l^m be the mean vector and covariance matrix of state s_l^m. Then, μ_m and Σ_m are given as follows:

$$\mu_m = \frac{\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t)o_t}{\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t)} = \frac{\sum_{l=1}^{L_m}\Gamma_l^m\mu_l^m}{\sum_{l=1}^{L_m}\Gamma_l^m} \quad (3)$$

$$\Sigma_m = \frac{\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t)(o_t-\mu_l^m)(o_t-\mu_l^m)'}{\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t)} \\ = \frac{\sum_{l=1}^{L_m}\Gamma_l^m(\Sigma_l^m+\mu_l^m(\mu_l^m)')}{\sum_{l=1}^{L_m}\Gamma_l^m} - \mu_m\mu_m{}' \quad (4)$$

where $\Gamma_l^m = \sum_{t=1}^{T_l}\gamma_l^m(t)$. To reduce computational complexity, it is assumed that the covariance matrix is diagonal resulting in

$$\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t)(o_t-\mu_m)'\Sigma_m^{-1}(o_t-\mu_m) = K\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t). \quad (5)$$

Considering (5), the log-likelihood of the model U can be rewritten as follow:

$$L(U) \simeq -\frac{1}{2}\sum_{m=1}^{M}\sum_{l=1}^{L_m}\sum_{t=1}^{T_l}\gamma_l^m(t) \\ \cdot (K + K\log 2\pi + \log(det(\Sigma_m))). \quad (6)$$

3. Outlier detection and removal

If a node in the decision tree contains outlying HMM states which have extraneous model parameters, the likelihood term of (1) would be distorted. In order to overcome the problems arising from outlying data in the conventional decision tree-based clustering method, we add a step which detects and removes outlying HMM states in the construction of a decision tree. The outlying HMM states have extraneous model parameters compared to the rest of HMM states at the same node. When treating the mean vector of the output probability distribution of an HMM state as an observation, we have to detect anomalous observations by means of an outlier detection algorithm for multivariate data.

A simple way to detect outliers for multivariate data is to calculate the distance from each data point to the centroid of the cluster. A data point with a distance larger than a predetermined threshold would be a possible outlier. In this work, the distance from a data point x_i to the centroid of a cluster of which mean and covariance are respectively μ and Σ is defined by:

$$d(x_i) = (x_i-\mu)'\Sigma^{-1}(x_i-\mu). \quad (7)$$

This quadratic form is often called the Mahalanobis distance which is a useful metric for determining the dissimilarity between two sample data [10]. Since, however, some of the data points in the cluster are outliers, it is not easy to obtain robust estimates for the mean, μ and covariance, $\sum$. For that reason, we apply the minimum covariance determinant (MCD) method which estimates the cluster mean and covariance such that they are resistant to outliers. The MCD method is described as follows [11, 12]:

Consider a data set $X_n = \{x_1, ..., x_n\}$ of K dimensional observations. Let the number of the robust observations be h. The h is set to $[n+K+1]/2$ or any integer satisfying $[n+K+1]/2 \le h \le n$.

1. Set $H_1 \subset \{1, ..., n\}$ with $|H_1| = h$ with $|\cdot|$ denoting the cardinality of a set and calculate

$$\hat{\mu}_1 = \frac{1}{h}\sum_{i\in H_1} x_i, \;\; \widehat{\Sigma}_1 = \frac{1}{h}\sum_{i\in H_1}(x_i-\mu_1)(x_i-\mu_1)'. \quad (8)$$

2. If $\det(\Sigma_1) \ne 0$, compute distances $d_1(i)$ for $i = 1, ..., n$

$$d_1(i) = (x_i-\hat{\mu}_1)'\widehat{\Sigma}_1^{-1}(x_i-\hat{\mu}_1) \quad (9)$$

3. Sort distances and take the h minimum observations into a new set H_2.
4. Compute $\hat{\mu}_2$ and $\widehat{\Sigma}_2$ with H_2 according to (8).
5. Iterate 1 to 4 until the estimates of the mean and covariance of the cluster no longer change.

After completing the above procedure, the contribution of outliers to computing the parameters of each cluster is removed or decreased. As a result, we can get more robust models for speech synthesis.

4. Experiments

In the experiments, we applied an English speech database spoken by a male and a female speaker. The speech database used for training the HMM parameters consists of 300 phonetically balanced sentences. We used 41 phonemes including silence as the basic units of speech synthesis and incorporated the following contextual factors to construct context-dependent models:

- preceding, current, and succeeding phonemes
- the accent and stress in the preceding, current, and succeeding syllables
- the part of speech of the preceding, current, and succeeding words
- the number of syllables in the preceding, current, and succeeding words
- the position of the current syllable in the word
- the position of the current syllable in the phrase
- the position of the current word in the phrase
- the numbers of syllables and words in the current phrase
- the punctuation symbol of current word

Speech signals were sampled at 16 kHz and windowed by a 25 ms Hamming window with a 5 ms shift. The acoustic feature vectors for HMM training included logarithm of the fundamental frequency, spectral parameters and their first- and second-order derivatives. As for the spectral parameters, we applied 25th-order mel-cepstral coefficients including the zeroth gain coefficients derived from STRAIGHT analysis [13]. A 5-state left to right structure with no skips was adopted to represent each context-dependent phone model. Furthermore, we used hidden semi-Markov model (HSMM) with explicit state duration distribution [14].

We compared the performances of two different models. One was trained by the conventional training technique, and the other by the proposed technique where the outlier removal algorithm was applied at each node of the decision tree. We did not remove outliers of the HMM states until the tree growing reached the third level of nodes because early removal would also get rid of important models.

One of the significant advantages of the proposed technique is to maintain stable synthesized speech quality by removing outlying models. For this, it is necessary to measure not only the average speech quality but also the worst case quality. First, we synthesized 100 sentences using the HMM trained by the conventional technique. Next, the 100 synthesized sentences were divided into two classes depending on the subjective speech quality. The first class consists of 50 sentences with fair speech quality while the other set, consisting of 50 sentences, collects the sentences of poor quality. We performed objective and subjective evaluations for these two classes separately.

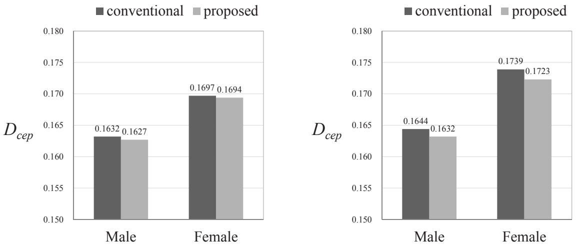
4.1. Objective evaluation

We used the following objective measures to calculate the differences between the generated and reference parameters of the spectrum, f0 and state duration, respectively:

1) mel-cepstral distance

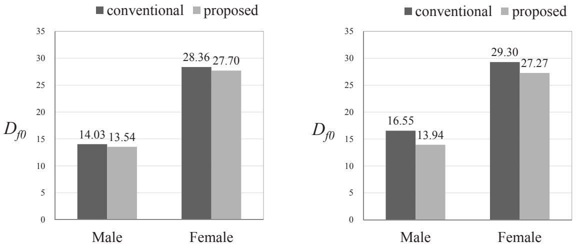
$$D_{cep} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{N_{order}}\sum_{i=1}^{N_{order}}(c_{ref}(t,i) - c_{syn}(t,i))^2} \qquad (10)$$

where T is the total number of frames, N_{order} denotes the order of mel-cepstral coeffiecients and $c(t,i)$ is the i-th mel-cepstral coeffiecient at the t-th frame where the subscript ref indicates the mel-cepstrum extracted from the actual speech waveform and syn means that acquired from the synthesized ones,



(a) *Fair quality set.* (b) *Poor quality set.*

Figure 2: *Mel-cepstral distance for sentences which originally demonstrated fair and poor quality.*



(a) *Fair quality set.* (b) *Poor quality set.*

Figure 3: *RMSE of f0 for sentences which originally demonstrated fair and poor quality.*

2) root mean square error (RMSE) of f0

$$D_{f0} = \sqrt{\frac{1}{T_{voiced}}\sum_{t=1}^{T_{voiced}}(f_{ref}(t) - f_{syn}(t))^2} \qquad (11)$$

where T_{voiced} means the total number of voiced frames and $f(t)$ is the fundamental frequency at the t-th frame with the subscripts ref and syn respectively denoting the reference and synthesized speech, and

3) root mean square error (RMSE) of state duration

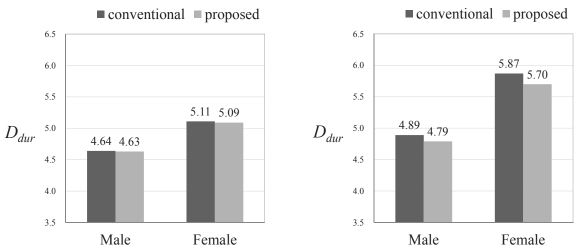
$$D_{dur} = \sqrt{\frac{1}{S_{total}}\sum_{s=1}^{S_{total}}(d_{ref}(s) - d_{syn}(s))^2} \qquad (12)$$

where S_{total} is the total number of states and $d(s)$ is state duration at the s-th state with the subscripts ref and syn indicating the reference and synthesized speech, respectively.

Figs. 2, 3 and 4 show the objective evaluation results for the fair and poor quality sets. The results obtained from the proposed technique show much better performance than that of the conventional technique especially for the sentences belonging to the poor speech quality set. As for the fair speech quality set, the two techniques show similar performances.

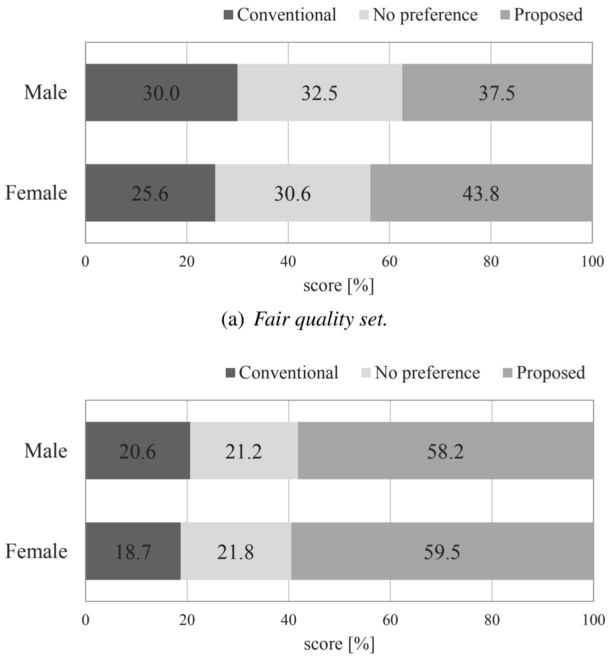
4.2. Subjective evaluation

We also conducted a number of paired preference tests for synthesized speech generated from the conventional and proposed techniques. Test sentences were classified into the fair and poor quality sets in the same way as in the objective evaluation tests. Nine listeners were asked to choose the more natural synthesized speech between two samples or choose “no preference”



(a) *Fair quality set.* (b) *Poor quality set.*

Figure 4: *RMSE of state duration for sentences which originally demonstrated fair and poor quality.*



(a) *Fair quality set.*

(b) *Poor quality set.*

Figure 5: *Results of preference test for sentences which originally demonstrated fair and poor quality.*

if the difference was not distinguishable. Twenty sentences for each experimental condition were used for this preference test and the results are given in Fig. 5. The results show that the synthesized speeches generated from the proposed technique are preferred, especially for the sentences which originally demonstrated poor quality.

5. Conclusions

In this paper, we have proposed an outlier removal algorithm which is applied during the decision tree growing phase of the HMM-based speech synthesizer training. With the use of the proposed method, the decision tree-based clustering algorithm can overcome its weakness arising from the outlier data, and as a result, the decision tree secures the robustness of the speech parameters by removing outliers. In a series of tests, we could confirm that the proposed technique guarantees a well-balanced speech quality irrespective of the given sentence.

6. Acknowledgments

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0083044) and by the National Research Foundation of Korea (NRF) funded by Korean Government (MEST) (NRF-2009-0085162).

7. References

[1] A.W. Black, H. Zen, and K. Tokuda, “Statistical parametric speech synthesis,” *Proc. of ICASSP*, pp.1229-1232, Apr. 2007.

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

[3] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Voice characteristics conversion for HMM-based speech synthesis system,” in *Proc. of ICASSP*, pp.1611-1614, Apr. 1997.

[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” in *Proc of ICASSP*, pp.805-808, May 2001.

[5] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” in *Proc. of Eurospeech*, pp.2523-2526, Sept. 1997.

[6] 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. Eurospeech*, pp. 2374-2350, Sep. 1999.

[7] S. Young, J. Odell, and P. Woodland, “Tree-based state tying for high accuracy acoustic modeling,” in *Proc. ARPA Human Language Technology Workshop*, pp. 307-312, Mar. 1994.

[8] K. Shinoda, and T. Watanabe, “Acoustic modeling based on the mdl principle for speech recognition,” in *Proc. EuroSpeech*, pp. 99-102, Sep. 1997.

[9] J. Yamagishi, *Average Voice Based Speech Synthesis*, Doctoral thesis, Tokyo Institute of Technology, pp.22-26, Mar. 2006.

[10] P. C. Mahalanobis, “On the generalised distance in statistics,” in *Proc. of the National Institute of Sciences of India*, pp. 49-55, Nov. 2008.

[11] J. Hardin, *Multivariate Outlier Detection and Robust Clustering with Minimum Covariance Determinant Estimation and S-Estimation*, Doctoral thesis, Uinversity of Califonia, pp. 1-55, 2000.

[12] P. Rousseeuw, and K. Driessen, “A fast algorithm for the minimum covariance determinant estimator,” in *Technometrics*, pp. 212-223, Aug. 1999.

[13] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited,” in *Proc. ICASSP*, pp. 1303-1306, Apr. 1997.

[14] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” *IEICE Trans. Inf. and Syst.*, pp. 825-834, May 2007.