BUILDING HMM BASED UNIT-SELECTION SPEECH SYNTHESIS SYSTEM USING SYNTHETIC SPEECH NATURALNESS EVALUATION SCORE

Heng Lu, Zhen-Hua Ling, Li-Rong Dai, Ren-Hua Wang

iFLY Speech Lab, University of Science and Technology of China, Hefei, China

luhenglh@mail.ustc.edu.cn, zhling@ustc.edu, lrdai@ustc.edu.cn, rhw@ustc.edu.cn

ABSTRACT

This paper proposes a unit-selection and waveform concatenation speech synthesis system based on synthetic speech naturalness evaluation. A Support Vector Machine (SVM) and Log Likelihood Ratio (LLR) based synthetic speech naturalness evaluation system was introduced in our previous work. In this paper, the evaluation system is improved in three aspects. Finally, a unit-selection and concatenation waveform speech synthesis system is built on the base of the synthetic speech naturalness evaluation system. Optimum unit sequence is chosen through the re-scoring for the N-best path. Subjective listening tests show the proposed synthetic speech evaluation based speech synthesis system significantly outperforms the traditional unit-selection speech synthesis system.

Index Terms— synthetic speech evaluation, unit-selection speech synthesis, kernel Fisher Discriminant, support vector machine

1. INTRODUCTION

Nowadays, there are mainly two mainstream speech synthesis systems. One is the HMM based parametric speech synthesis system [1][2], while the other is the unit-selection and waveform concatenation [3][4] speech synthesis system. In HMM based parametric speech synthesis, Maximum Likelihood (ML) or Minimum Generation Error (MGE) [5][6] are introduced as criteria in the acoustic model training and parameter generation [7]. In unit-selection and waveform concatenation speech synthesis, unit sequence with the minimum costs including the "target" cost and "concatenation" cost is chosen and concatenated to generate synthetic speech. However, none of these speech synthesis criterions directly involves any human perception in the system construction. In [8], we have proposed an SVM and LLR based synthetic speech naturalness evaluation system to evaluate the naturalness of the synthetic speech. And in this paper, we continue to improve the accuracy of the speech naturalness evaluation system in three aspects. Firstly, posterior probability instead of LLR is employed as SVM classification feature in evaluation system. Secondly, the decision tree scale for acoustic model clustering is optimized through the Kernel Fisher Discriminant (KFD) [9] analysis for classification features. Thirdly, Cross-Validation (CV) technique is used to estimate the overall classification accuracy of each parameter combination in the specified range which helped to decide the best parameters set for SVM classifier. Finally, on the base of the finely tuned synthetic speech naturalness evaluation system, a unit-selection speech synthesis system with N-best route re-scoring is constructed. Subjective listening tests show that, proposed system significantly outperforms the baseline HMM-based unit-selection and waveform concatenation speech synthesis system.

The rest of the paper is organized as follows. Section 2 reviews the SVM and LLR based speech naturalness evaluation system proposed in [8]. Section 3 introduces the new improvements on the previous evaluation system. In section 4, an unit-selection and waveform concatenation speech synthesis system is built using the evaluation score given by evaluation system. Objective experiments and subjective listening tests are shown in Section 5. Conclusions and future work are presented in Section 6.

2. SYNTHETIC SPEECH NATURALNESS EVALUATION SYSTEM

We have proposed a SVM and LLR based synthetic speech naturalness evaluation system to evaluate the naturalness of the unit-selection and waveform concatenation synthetic speech in [8]. The framework of the system is shown in Fig. 1. The system construction mainly consists of three steps. 1) Training data set preparation by manual *natural* and *unnatural* labeling for synthetic speech. In [8], 5 native speakers have labeled a total 169,492 names of places generated by a unit-selection speech synthesis system provided by the iFlytek company for GPS navigation application. And prosody word is utilized as the basic unit for labeling, either prosody word is label as *natural* or *unnatural*. 2) Acoustic feature extraction and context-dependent acoustic model training. In this step, pronunciation of the synthetic speech is divided into two spaces [10], the *natural* and *unnatural* pronunciation space. And context dependent acoustic model is trained for the two spaces separately. 3) SVM classifier training. LLR is calculated from acoustic features of the synthetic speech given the two spaces acoustic model trained in step 2 and combined to compose the feature for SVM classification for each basic labeling unit. Finally, SVM classier is trained to evaluate the naturalness for the synthetic speech.

3. IMPROVEMENTS ON SYNTHETIC SPEECH NATURALNESS EVALUATION SYSTEM

3.1. Posterior probability SVM classification features

The combination of posterior probability instead of LLR is employed as the SVM classification feature. Assuming a prosodic word consists of N syllables[1] and the extracted acoustic feature has M dimensions, an $N * M$ dimensional vector $\mathbf{\Omega}$ is composed as

$$\mathbf{\Omega} = \left[\omega_1^1, ..., \omega_1^M, \omega_2^1, ..., \omega_2^M, ..., \omega_n^m, ..., \omega_N^1, ..., \omega_N^M\right]^\top, \quad (1)$$

[1]Syllable is assumed to be the basic unit of trained acoustic models here to facilitate the introduction.

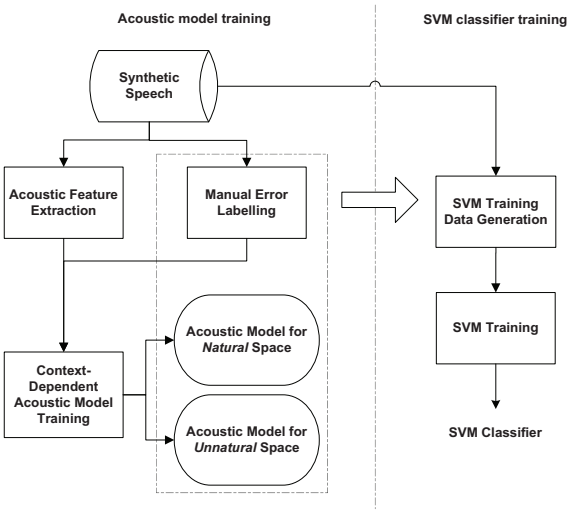

Fig. 1. Flowchart of the synthetic speech evaluation system

and

$$\omega_n^m = \frac{P_m(\boldsymbol{o}_n|\lambda_U, c_n)}{P_m(\boldsymbol{o}_n|\lambda_N, c_n) + P_m(\boldsymbol{o}_n|\lambda_U, c_n)}, \quad (2)$$

where $n = 1, ..., N; m = 1, ..., M$; $\boldsymbol{o}_n$ stands for the acoustic features belong to the n-th syllable in the word; λ_N and λ_U denote the trained acoustic models for *natural* and *unnatural* space respectively; c_n means the context information of syllable n; $P_m(\boldsymbol{o}|\lambda, c)$ indicates the component of likelihood calculated for the m-th dimension of feature $\boldsymbol{o}$ given model λ and context c.

Compared with LLR, the dynamic range of posterior probability is normalized to a small and fixed ($0 \sim 1$) range. Therefore, different dynamic range for different natural/unnatural decision tree cluster is avoided. And posterior probability is a more appropriate feature for SVM classifier training for its fixed dynamic range.

3.2. Decision tree scale optimization with KFD analysis

The the overall performance of the synthetic speech naturalness evaluation system is closely related to its classification features. In our system, the classification features consist of posterior probability. Since the classification features is generated from the clustered acoustic model, we can optimized the discrimination of the classification features through tuning the scale of the *natural* and *unnatural* acoustic model. However, since the SVM training is time consuming and the performance of SVM is also affected by other factors such as misclassification penalty parameter C selection, KFD analysis is chosen to analyze the discrimination of the classification features. KFD is a powerful discriminant analysis method to efficiently compute the degree of discrimination in feature space with kernel trick. The definition and calculating equation of KFD are given in [9].

And Minimum Description Length (MDL) [11][12] criterion is utilized to conduct context dependent acoustic model clustering in the *natural* and *unnatural* acoustic model training process. However, the MDL criterion is derived based on an asymptotic assumption and the assumption fails when there is not enough training data [13]. Since the amount of training data in our experiment is much smaller than speech recognition tasks, MDL criterion may not work successfully. Therefore, the decision tree scale needs to be tuned to get the optimum evaluation system performance. MDL factor α is a parameter set to control the scale of the binary decision tree in the MDL clustering process. Small α value lead to large decision tree scale and large α lead to small tree scale. In MDL criterion, the default α value is 1.

Therefore through calculating the KFD for classification features with different MDL factor α value, the scale of the binary decision tree can be optimized automatically and the discrimination in feature space is maximized through tuning the scale of the decision tree.

3.3. Cross-Validation for parameter tuning in SVM training

In our synthetic speech synthesis naturalness evaluation system, Radial Basis Function (RBF) is chosen as the kernel function in SVM classifier training. RBF kernel is shown in equation 3. There is a multiplier γ in the RBF definition.

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = e^{-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2}, \quad (3)$$

In [8], the SVM training tool LIBSVM [14] is used to train the SVM classifier. And the misclassification penalty parameter C and γ are set to default value ($C = 1$,$\gamma = 1$) in RBF SVM training process. In this paper, CV technique is conducted to estimate the classification accuracy of each parameter combination (C,γ) in the specified range. The overall classification accuracy is employed as the criterion for parameter tuning.

4. BUILDING UNIT-SELECTION SPEECH SYNTHESIS SYSTEM WITH EVALUATION SCORE

A unit-selection and waveform concatenation speech synthesis system is constructed on the based of the synthetic speech naturalness evaluation system. Other than the only one path with the minimum "target" and "concatenation" cost chosen using Viterbi scoring, N-best unit sequences are preserved for re-scoring by synthetic speech naturalness evaluation system. The re-scoring strategy is as follows. We set a fixed threshold manually for synthetic speech naturalness evaluation score, then score the one best path with the minimum "target" and "concatenation" cost given in previous selection using proposed synthetic speech naturalness evaluation system. As higher evaluation score represents better synthetic speech naturalness, when the score of the path is higher than the threshold, the one best path given in previous selection is regarded as natural and chosen as output. Otherwise, it is regarded as unnatural and the path with the highest synthetic speech evaluation score is chosen from the N-best paths as the optimum unit sequence for further concatenation.

5. EXPERIMENTS

5.1. Improvements on synthetic speech synthesis naturalness evaluation system

The same evaluation task and data set as in [8] is employed in this paper. Since pitch unnaturalness is the most frequent error type that happens in Mandarin and pitch error harm the overall synthetic speech naturalness most, we continue to focus on detecting the synthetic errors caused by unnatural pitch patterns in Mandarin in our experiments. A text set consisting of 169,492 names of places was used in our experiments. Theses texts were synthesized using the synthesis system provided by iFLYTEK company for GPS navigation application, and the basic unit for concatenation is phone. And the synthetic results were manually labeled as *natural* and *unnatural* in the data set preparation step. In our experiment, each sentence

MDL factor α	KFD value
0.1	0.2403
1	0.1619
5	0.1388
10	0.1383

Table 1. KFD value varies according to the variation of MDL factor α (3-syllable SVM)

consisted of only one prosodic word for the text of place names. Among all labeled synthetic speech, 54,355 prosodic words were labeled with synthetic errors on pitch perception, while the other 117,397 were labeled as natural ones. And 90% of these synthetic sentences were randomly selected to construct a training set, while the remaining ones were used as a test set. In the acoustic model training step, syllable is employed as the basic unit for long-term, average and concatenating characteristic pitch feature extraction and acoustic model training. The same 13 dimensional acoustic feature as in [8] is extracted for each syllable. Using these features, two context-dependent HMM models are trained for the *natural* and *unnatural* pronunciation space separately. SVM vector for classification is calculated as in Section 3.1 for every prosody word in the training set. For prosody words containing different amount of syllables, SVM classifiers are trained separately for the different size of dimensions of the SVM feature vectors. We take the SVM for 3-syllable word and 2-syllable word as an example in the following introduction. For 3-syllable SVM, 31,114 natural words and 15,770 unnatural words are in the training data set, and 3,458 natural and 1,753 unnatural in testing set. For 2-syllable SVM classifier, 10,249 natural 4,189 unnatural and 1,139 natural 466 unnatural prosody words are contained by the training and testing data set respectively.

5.1.1. Decision tree scale optimization with KFD analysis

KFD analysis is implemented for word SVM training classification feature generated with different decision tree scale. RBF kernel is chosen in accord to the RBF SVM classifier used in evaluation system. Table 1 shows the the KFD analysis results according to different MDL factor α values for the 3-syllable SVM classifier. Parameter μ [9] is set to 0.001 and γ is set to its default value 1 in the KFD analysis process.

Larger KFD value stands for better discrimination between the *natural* and *unnatural* classification features. From the resultx we can see that better discrimination is obtained when the scale of the decision tree becomes larger. We choose $\alpha = 0.1$ as the optimum parameter. The KFD value by $\alpha = 0.1$ is significantly higher than the default $\alpha = 1$. And for 2-syllable case, $\alpha = 0.1$ is chosen through KFD analysis.

5.1.2. CV for parameter tuning in SVM training

With the MDL factor α set to 0.1, SVM classifier is trained. CV is engaged to tune the parameter set (C,γ). The total training data set is divided into 5 subset in the CV process. The parameter tuning result is shown in Fig. 2 (3-syllable SVM, MDL=0.1), the overall classification accuracy for the two classes is 74.700%. As a comparison, for the 3-syllable SVM, $\alpha = 1$, the classification accuracy

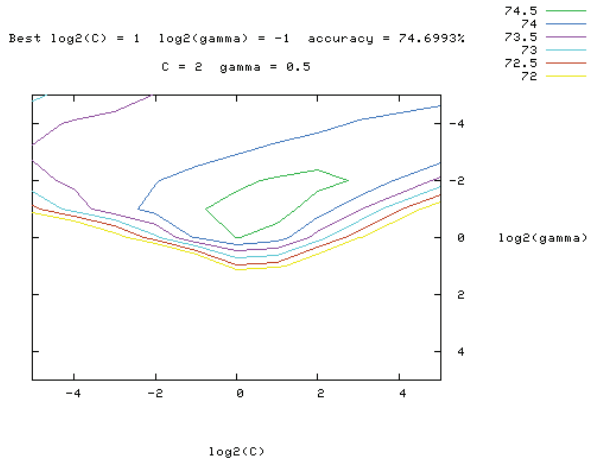

Fig. 2. CV based SVM parameter tuning (3-syllable SVM, $\alpha = 0.1$)

Preference	3-syllable	2-syllable
$A > B$	26.7%	24.2%
$A = B$	22.7%	20.9%
$A < B$	50.7%	55.0%
$A > C$	38.7%	25.0%
$A = C$	26.7%	41.7%
$A < C$	34.7%	33.4%

Table 2. Subjective preference listening test for the 3 and 2 syllable case. A: baseline, B: proposed, C: random selection

is 71.037%. The results show that the $\alpha = 0.1$ SVM outperform the $\alpha = 1$ SVM classifier, and it is consistent with the KFD analysis results. For the and 2-syllable SVM, when $\alpha = 0.1$, the overall classification accuracy is 76.292%.

From the CV tuning results, the optimum parameter set ($C = 2$,$\gamma = 0.5$) is chosen for the 3 and 2 syllable SVM respectively. The tool LIBSVM is employed in the SVM training process.

5.2. Subjective listening tests for proposed speech synthesis system

Given 6874 full sentences (not only name of places) by the same speaker as the GPS navigation system, three speech synthesis systems are constructed as comparison. System A is the system proposed in [15], which serves as the baseline system. In system A, only one route with the minimum cost is chosen. System B is the synthetic speech naturalness evaluation based unit-selection speech synthesis system proposed in Section 4. System C is a random route selection system in which one route is randomly chosen from the N-best unit sequences for unit concatenation. The framework of system B is shown in Fig. 3. In system B, a three level unit selection is conducted to choose the optimum sentence unit sequence. Firstly, Kullback-Leibler divergence (KLD) [16] between the model of the candidate unit and the target model is used to conduct the unit pre-selection to reduce computational cost in the Viterbi unit-selection process. Secondly, after pre-selection, a function consisting of target cost and concatenation cost calculated from the acoustic models

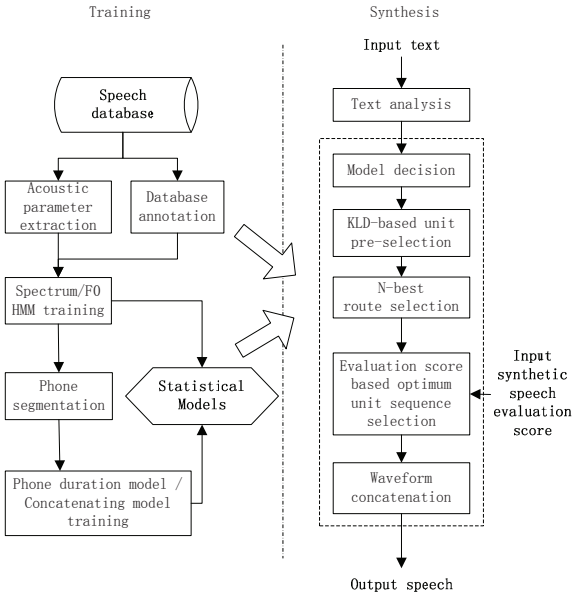

Fig. 3. Building HMM based unit-selection speech synthesis system using evaluation score

are combined with the KLD is minimized to select the optimum unit sequence. Other than the only one route chosen as in [15], N-best unit sequences are preserved for further selection. Finally, the final one best unit sequence is find out using the strategy introduced in Section 4 from the N-best route.

For the 3-syllable SVM classifier, we synthesize 100 name of places with 3-syllable by each system with $N = 200$, and ask native speaker to label the 100 sentences. Among 100 synthetic sentences, 40 sentences are labeled as completely natural and the other 60 sentences are regarded unnatural with different extent. By setting evaluation threshold to 0.5670, the accuracy for the unnatural unit is 65.8%, and recall 80.0%. Preference tests are conducted between system A,B and system A,C. Five natives are asked to give preference score on randomly chosen 15 sentences from the 100 synthetic speech. The result is shown in Table 2.

The results prove that our proposed evaluation score based HMM based unit-selection speech synthesis system significantly outperform the baseline speech synthesis system A. System C is comparable to system A shows an important knowledge that when the cost is small enough (N-best), minimum cost criterion can no longer discern unnatural from natural speech. In this case, criterion concerns with human perception is a necessary for further selection.

6. CONCLUSIONS AND FUTURE WORK

In this paper, synthetic speech naturalness evaluation system proposed in our previous work is improved in three directions. And a evaluation score based HMM based unit-selection speech synthesis system is proposed. Subjective listening test show that proposed evaluation score based HMM based unit-selection speech synthesis system significantly outperform the baseline system without evaluation score. Our future work includes employing the synthetic naturalness evaluation system to continuous long sentences, and to include some spectrum and duration feature into SVM classifier.

7. REFERENCES

[1] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," *Proc.ICASSP*, vol. 4, pp. 1229–1232, 2007.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, Nov. 2009.

[3] N. Iwahashi, N. Kaiki, , and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," *IEICE Trans. Fundamentals*, vol. E76-A, no.11, pp. 1942–1948, Nov. 1993.

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

[5] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM based speech synthesis," *Proc.ICASSP*, p. 89C92, 2006.

[6] Y.-J. Wu, H. Zen, Y. Nankaku, and K. Tokuda, "Minimum generation error criterion considering global/local variance for HMM-based speech synthesis," *Proc.ICASSP*, p. 4621C4624, 2008.

[7] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc.ICASSP*, pp. 660–663, 1995.

[8] H. Lu, Z.-H. Ling, S. Wei, L.-R. Dai, and R.-H. Wang, "Automatic error detection for unit selection speech synthesis using log likelihood ratio based SVM classifier," *Proc.Interspeech*, p. 162C165, 2010.

[9] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers, "Fisher discriminant analysis with kernels," *Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pp. 41–48, Aug 1999.

[10] S. Wei, G.-P.. Hu, Y. Hu, and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no 10, pp. 896–905, Oct. 2009.

[11] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn*, vol. 21, 2, pp. 1, 2000.

[12] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," *EuroSpeech97*, pp. 99–102, 1997.

[13] J. Rissanen, "Stochastic complexity in stochastic inquiry," *World Scientific Publishing Company*, 1980.

[14] "http://www.csie.ntu.edu.tw/scjlin/libsvm/," .

[15] Z.-H. Ling and R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," *Proc. ICASSP*, pp. 1245–1248, 2007.

[16] P. Liu and F.K. Soong, "Kullback-Leibler divergence between two hidden Markov models," *Microsoft Research Asia, Tech. Rep.*, 2005.