I^{2}R's Submission to Blizzard Challenge 2008

Minghui Dong, *Donglai Zhu*, *Bin Ma* and *Haizhou Li*

Human Language Technology Department, Institute for Infocomm Research (I^{2}R),
Agency for Science, Technology and Research (A*STAR), Singapore 138632
{mhdong, dzhu, binma, hli}@i2r.a-star.edu.sg

Abstract

This paper reports the I^{2}R's submission to the Blizzard Challenge 2008. This is our first participation in Blizzard Challenge. In this paper, we describe the approach that we used to build the three required voices. We introduced the acoustic parameters that include MFCC coefficients as spectral parameters in addition to the prosodic parameters for unit selection based speech synthesis. We used regression tree approach to predict the acoustic parameters. The evaluation results show that our approach works well for English speech synthesis. The approach has also shown good performance in keeping the speaker characteristics of the speech database in Mandarin speech synthesis.

Index Terms: speech synthesis, unit selection, spectral and prosodic parameters, parameter prediction

1. Introduction

Blizzard Challenge [1] provides a good way to evaluate different speech synthesis methods on the same datasets. In this year's Blizzard Challenge, two databases are provided to the participants. The first database is a British English database, and the second one is a Mandarin Chinese speech database. Each participant is asked to generate two English voices and one mandarin voice. The first English voice (Voice A) is generated with the full English database, while the second English voice (Voice B) should be generated with the ARCTIC subset of the English database. Participants are requested to use the same approach to generate the three voices.

2. Overview of Our Approach

The system of I^{2}R adopted the unit selection based approach[2][3][4][5]. Unit selection approach to speech synthesis has been shown to be one of the best approaches currently. In a unit selection based speech synthesis system, there is a large unit database which consists of many instances of the same unit as candidate units. The units in the database are designed to cover the variants of the unit as much as possible. During synthesis, a proper unit will be selected from all the candidates of the target unit. Finally the selected units are concatenated to form a speech utterance.

To maintain the naturalness of the synthetic speech, prosody needs to be predicted for each unit. The prosody model predicts the prosodic parameters, which normally describe the pitch, duration and energy of speech units. These prosodic parameters are used as part of the criteria for unit selection.

Besides the normally used prosody parameters, to make sure the selected units have good spectrum appropriateness in the synthetic speech, the spectrum information should also be taken into account in the synthesis process. In our work, we introduced MFCC coefficients as spectral parameters. By using MFCC coefficients, the target spectrum of units can be predicted using statistical models and be applied in unit selection process. We included both prosodic parameters and spectral parameters into our unit selection criteria, and used statistical model to predict the parameters.

3. Speech Database Processing

In this part, we explain how we generate unit labels and calculate speech features.

3.1. Unit Labeling

To build the unit database for the unit selection based synthesis, we first need to extract the unit, ie, identifying the start and end points of each unit in the speech utterances. In this work, the labeling of the phone-sized unit is done with HTK automatically. 39 dimensional MFCC feature is used for the training of the phone models. The frame size is 0.025 second and the frame shift is 0.01 second. Three states are defined for each context independent HMM model for each phone. The phone models are first trained with the speech corpus. Unit boundaries are then obtained by force alignment of speech with its phonetic sequence. The MFCC features that are used for the alignment will further be used in the later stage of the unit selection based synthesis process.

3.2. F0 Calculation

F0 feature is one of the most important features of prosody of speech. In this work, F0 of speech utterance is calculated with the Praat software [6]. Same as MFCC, we use a frame size of 0.01 second. The F0 values of every 0.01 second interval are calculated. For unvoiced part, interpolation is done to give a none-zero F0 value for each interval. Then we apply a simple smoothing process to this F0 sequence. The smoothing is done with moving average represented with the following formula:

$$p'_i = (p_{i-1} + p_i + p_{i+1})/3 \tag{1}$$

where p_i is the F0 value of the i-th frame.

3.3. Unit Filtering

Although the speech corpus is carefully designed and recorded, it is inevitable that some speech units may sound not as good as other units. It is expected that these unit should be excluded in the speech synthesis process. To filter out these units, we have done a speech recognition process to recognize the speech segments of the unit obtained from forced alignment. The units that are not on the top positions of the recognition result are marked and would be excluded from the speech synthesis process.

3.4. The databases

English Database: The English speech corpus in Blizzard challenge is a British English corpus that is provided by University of Edinburgh. The released part of the corpus consists of 15 hours speech in 9,509 utterances, which cover children stories, isolated words, emphasis carrying sentences, news articles, etc[7][8]. The corpus was designed to cover the variants of diphone as much as possible. The corpus comes with transcriptions, which are contained in files of festival utterance format. The RP phone set [9] is used to define the pronunciations of the utterances. There are 50 different phonemes in the corpus.

Mandarin Database: The mandarin speech corpus consists of about 5 hours’ speech in 4500 utterances. The text transcription of the corpus comes from news corpus. The corpus was designed to cover variants of Chinese pronunciations. We defined 43 different phonemes for our task.

4. Prosody Model

In this part, we describe how the prosody model of the speech synthesis system is built. As we have included MFCCs into the parameter set, it may be more proper to use the name acoustic parameters rather than prosodic parameters.

4.1. The Acoustic Parameters

The acoustic parameters we define here are used as criteria for selecting the best units. Normally, the parameters are prosodic parameters that describe pitch and duration of unit. However, the use of prosody alone does not take into account the spectral mismatches. To better describe the unit for unit selection, we include both spectral parameters and prosodic parameters in our models. In this work, we use MFCC as our parameters to represent spectral information. The 13 dimensional basic MFCC coefficients, delta MFCC and delta-delta MFCC form a 39-dimensional vector.

We use phone as the basic synthesis unit. The speech signal of each unit is separated into 3 segments, each corresponding to one of 3 HMM states in forced alignment. We use mean values of MFCC vectors for each speech segment to represent the spectral information of the HMM state. For prosodic parameters, we only consider mean value of F0 and duration for each HMM state. Therefore, the 39 MFCC coefficients, the F0 and duration values together form a 41-dimensional vector for each state. For each unit, there are three vectors to represent the three states.

The acoustic parameters for each unit can be represented as the following:

$$X = (X_1, X_2, X_3) \tag{2}$$

where X_i is a 41-dimensional vector for state i (i = 1, 2, 3).

4.2. Linguistic Features

The linguistic features are derived from input text. They are used for predicting the acoustic parameters. Due to the difference of language and the information availability, we defined different linguistic features for English and Chinese.

The English corpus comes with the utterance structure for each speech file. We define the features following those that are used in HTS system [10]. We derived the following linguistic features from the utterance files:

- Context units: phone identities of the previous 2 and next 2 units. (4)
- Syllable information: Stress, accent, length of the previous, current and next syllables. (9)
- Syllable position information: syllable position in word and phrase, stressed syllable position in phrase, accented syllable position in phrase, distance from the stressed syllable, distance from the accented syllable, and name of the vowel in the syllable. (13)
- Word information: length and part-of-speech of the previous word, current word and next word, position of the word in phrase. (12).
- Phrase information: Lengths (in number of words and syllables) of previous phrase, current phrase and next phrase, position of the current phrase in major phrase, boundary tone of the current phase. (8)
- Utterance information: Lengths in number of syllables, words and phrases. (3)

Putting all the features together, we form an input linguistic feature vector of 53 elements for English.

For Mandarin corpus, we defined less linguistic features as the components for calculating some features are not available. The features we used include:

- Context units: phone identities of the previous 2 and next 2 units. (4)
- Tone information: The tones of the current, previous and next syllables. (3)
- Prosodic word information: Whether the syllable is the start or end position in prosodic word. (2)

Altogether, we have a linguistic feature vector of 9 elements for Mandarin.

4.3. Acoustic Parameter Prediction

The acoustic parameter prediction process calculates the parameters from the linguistic features. The prediction can be represented with the following formula:

$$y_i = F_i(X) \tag{3}$$

where y_i is the i-th parameter for the unit, X is the linguistic feature vector for the unit.

In this work, with the linguistic features being the predictors and the acoustic parameters (3 vectors for 3 states) being the responses, we build our models using CART [11] approach. Each parameter is predicted separately with a CART tree. This is different from the prediction of the vector in [12], where CART is used to predict a vector. We try to use individual trees to predict a more accurate value for each individual parameter.

Regression tree is used to model the parameters. Each acoustic parameter is predicted separately with an individual tree. For each parameter, a CART tree is first trained with the training data. Then it is tested with 10-fold cross-validation method on the training data to find the best sub-tree. Finally, the tree is pruned to the best sub-tree. In total, 123 trees are trained for the 3 vectors of a unit.

For voice A (full database), the corpus consists of 9,509 utterances. Without loss of generality, in our task, we only use part of them for our training. We selected 2,000 utterances from different categories of the utterances. Among the units contained in the utterances, we randomly selected 35,200 units as our training data. For voice B, the databases consists of 35952 units, we use all of the data as our training data. For Mandarin voice, we selected 36000 units as our training data using the same procedure as voice A.

5. Unit Selection

We use unit selection approach as our synthesis method. In this part, we describe how we define the cost function.

5.1. Cost Function

The unit selection process is based on the cost function that consists of two parts (1) a target cost to measure the difference between the target unit and the candidate unit. (2) a join cost to measure the acoustic smoothness between the concatenated units.

Our target cost further consists of two parts (1) the cost of acoustic parameters and (2) the cost of context linguistic features. The target cost c_t is defined as the following:

$$c_t = w_{ta}c_{ta} + w_{tl}c_{tl} \tag{4}$$

where, c_{ta} and c_{tl} are cost of acoustic parameters and cost of linguistic features, w_{ta} and w_{tl} are the weights respectively.

The reason why we use two cost components here is that each one of them alone is not sufficient to describe the target cost. The cost of linguistic feature is to ensure the general spectral and prosodic correctness of the candidate unit. However, due to the variations of speech, using this cost alone may easily leads to extreme cases (abnormal spectrum). The use of cost for acoustic parameter can avoid the selection of the extreme cases, because the statistical models favor the average values. However, the use of acoustic parameter alone is also not enough because the accuracy of prediction model is always limited.

The cost of acoustic parameters c_{ta} is defined as the squared value of Mahalanobis distance [13] (i.e we did not take the square root in the following formula) between the target unit and the candidate unit is as the following:

$$c_{ta} = \sum_{i=1}^{3}((U_i - V_i)^T W^{-1}(U_i - V_i)) \tag{5}$$

where W is the covariance matrix, U_i and V_i (i = 1, 2, 3) are predicted parameter vectors for target unit and the actual parameter vector for candidate unit. The reason why we use Mahalanobis distance is that it takes into account the correlations between the each element in the vector.

The cost of context linguistic features c_{tl} is defined according to the difference between the features of the target unit and those of the candidate units. When the feature is different, a cost value is given. The total cost is the sum of all the costs for each individual features. In this function, we give higher cost value to the mismatch of important factors (e.g. the identities of immediate previous unit and immediate next unit, the accent of the unit, the stress of the unit, etc).

Join cost c_j measures the mismatch between two units that will be concatenated. It is defined as the squared value of Mahalanobis distance between the vector of the end frame of the previous unit E_{i-1} vector of the start frame of the current unit S_i.

$$c_j = (E_{i-1} - S_i)^T W^{-1}(E_{i-1} - S_i) \tag{6}$$

where W is the covariance matrix.

The total cost c is calculated with the following function.

$$c = w_t \sum_{i=0}^{n} c_t(i) + w_j \sum_{i=1}^{n} c_j(i) \qquad (7)$$

where n is number of units in the sequence, $c_t(i)$ is the target cost of unit i, $c_j(i)$ is the join cost between unit i-1 and unit i, w_t and w_j are weights for target cost and join cost respectively.

The best unit sequence is determined by searching for a best path among the candidate unit lattice to minimize the total cost of the selected sequence. Viterbi algorithm is used to find the best sequence.

The weights in the cost function are manually tuned. To tune the weight to the system's best performance, we selected 20 testing sentences. Whenever there is a new setting for the weights, we synthesize the testing sentences to see whether the speech quality is improved. For the weights in Eq 4, we find that that best speech quality is achieved when w_{ta} and w_{tl} are given roughly the same value. For the weights in Eq 7, we also give w_t and w_j roughly the same value.

6. Evaluation Results

The organizer of Blizzard Challenge 2008 has conducted online listening evaluation and released results. This is our first participation in the Blizzard Challenge. We are happy to see some encouraging aspects of our approach from the results.

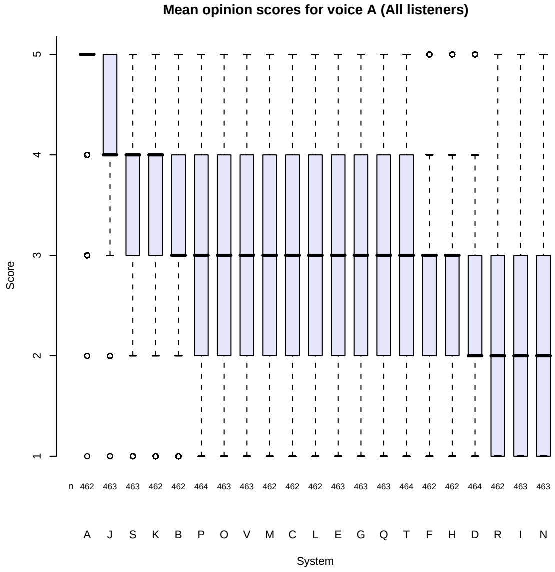

Figure 1. MOS score for English voice A (All listeners)

For English voice A, we achieved a mean MOS score of 3.1 and the similarity score of 3.2. Figure 1 shows the statistics of MOS score for voice A from all listeners' feedback. Our system is O in the Figure. From the figure, we can see that our system achieved a median score of 3. This shows that our method for English voice synthesis is successful.

For Mandarin voice, we achieved a MOS score of 3.3 and the similarity score of 3.5. Our similarity score is the highest among all the Mandarin systems. Figure 2 shows the statistics of similarity score for Mandarin voice from all listeners' feedback. From the figure, we can see that our system achieved a median score of 4 for similarity to original speaker. This shows that our method is able to keep the speaker characteristics very successfully.

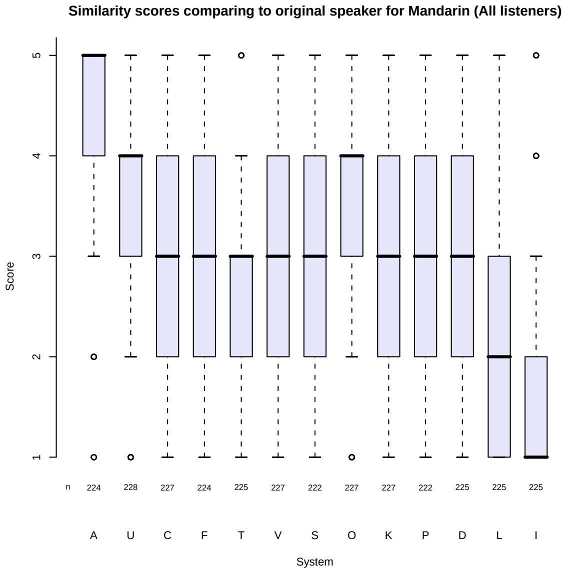

Figure 2. Similarity score for Mandarin voice (All listeners)

7. Discussion

Our system showed good performance in the score of similarity to the original speaker in Mandarin speech synthesis. Although the exact reason is not fully investigated, we think the most likely reason is that we have included MFCC coefficients to characterize the spectral feature in unit selection process.

We have also realized that we need to improve our system in a few aspects:

(1) As we use fully automatic way to generate the unit label. There are some errors in the speech unit database, which affects the speech quality of some synthesized utterance. We need to further investigate how to remove the errors in the labeling.

(2) The weights in the cost function are determined manually. The setting of the weights may be improved further by using some automatic approach.

(3) We did not use a pre-classification process to select the unit candidate for unit selection. The speech

synthesis process is a little slow now. We hope to improve it in future.

8. Conclusions

This paper described our speech synthesis approach for Blizzard Challenge 2008. We introduced the acoustic parameters that include MFCC coefficients as spectral parameters in addition to the prosodic parameters for unit selection based speech synthesis. We used regression tree approach to predict the acoustic parameters. The cost function was defined to use the cost of acoustic parameters as one of the components. The cost for acoustic parameters is defined as the Mahalanobis distance between the respective vectors. The evaluation results show that our approach works quite well for English speech synthesis. The approach has also shown good performance in keeping the speaker characteristics of the speech database in Mandarin speech synthesis.

9. References

[1] Robert A. J. Clark, Monika Podsiadlo, Mark Fraser, Catherine Mayo, Simon King, Statistical Analysis of the Blizzard Challenge 2007 Listening Test Results, Proc. Blizzard Challenge Workshop, 2007, Bonn, Germany.

[2] A. W. Black, P. Taylor, “Automatically Clustering Similar Units for Unit Selection in Speech Synthesis,” in Proc. Eurospeech 97, vol 2 pp 601-604, Thodes, Greece.

[3] R. Clark, K. Richmond, V. Strom, S. King, “Multisyn voice for the Blizzard Challenge 2006,” Blizzard Workshop 2006.

[4] M. Schroder, A. Hunecke, S. Krstulovic, “OpenMary – Open Source Unit Selction as the Basic for Research on Expressive Synthesis,” Blizzard Workshop 2006.

[5] M. Chu, H. Peng, Y. Zhao, Z. Niu and E. Chang, “Microsoft Mulan – a Bilingual TTS System”, Proc. of ICASSP 2003, Hong Kong, 2003.

[6] Boersma, Paul, “Praat, a System for Doing Phonetics by Computer.” Glot International 5:9/10, 341-345, 2001.

[7] V. Strom, R. Clark, and S. King, “Expressive Prosody for Unit-Selection Speech Synthesis,” in Proc. Interspeech, Pittsburgh, 2006.

[8] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, D. Jurafsky, “Modelling Prominence and Emphasis Improves Unit-Selection Synthesis,” in Proc. Interspeech, Antwerp, 2007.

[9] S. Fitt, “Documentation and User Guide to UNISYN Lexicon and Post-lexical Rules, Tech. Rep.”, Centre for Speech Technology Research, Edinburgh, 2000.

[10] http://hts.sp.nitech.ac.jp/.

[11] L. Breiman, , J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and Regression Trees”. Monterey, Calif., U.S.A.: Wadsworth, Inc., 1984.

[12] A. W. Black, C. L. Bennett, B. C. Blanchard, J. Kominek, B. Langner, K. Prahallad and A. Toth, “CMU Blizzard 2007, A hybrid Acoustic Unit Selection System from Statistically Predicted Parameters,” Blizzard Challenge workshop 2007.

[13] P.C. Mahalanobis, On the Generalized Distance in Statistics, Proceedings of the National Institute of Science of India 12 (1936) 49-55.