Multi-tier Non-uniform Unit Selection for Corpus-based Speech Synthesis

Jin-Hui Yang, Zhi-Wei Zhao, Yuan Jiang, Guo-Ping Hu, Xiao-Ru Wu

iFLYTEK Research, Hefei, China
{jhyang,zwzhao,yuanjiang,gphu,xrwu}@iflytek.com

ABSTRACT

In this paper, a corpus-based speech synthesis system KB2006 was developed using the speech database provided by Blizzard Challenge 2006. We proposed a novel unit selection method called multi-tier non-uniform unit selection in our corpus-base speech synthesis system. Non-uniform unit (NUU) in our system was defined as a unit sequences that contains one or more joint phoneme units. By using CART algorithm, NUUs with the same phoneme sequence in the inventory were clustered into different classes according to their prosody and acoustic difference. In the unit selection stage, a multi-tier NUUs selection algorithm was adopted by treating different NUUs with several criterions. With the discrimination, proper candidate units that close to the target unit can be selected for speech concatenation.

1. INTRODUCTION

In recent years, corpus-based speech synthesis has been very popular for its high quality and naturalness speech output [1] [2] [3]. In this method, synthesized speech was produced by concatenating units that were selected from a speech corpus, which contains a larger number of speech units with various prosody and spectral characteristics. The selection process retrieves units of speech from the corpus that best match the target features predicted by the TTS front-end component. The retrieved units should be individually close to their targets and adjacent units should be mutually compatible. To achieving this purpose, target costs of every candidate unit were calculated, and then Viterbi searching algorithm was adopted using concatenation costs to find a best path through the candidate units that minimizes the total cost.

In order to produce high quality speech, a large amount of speech is used to construct the speech corpus, just as Blizzard Challenge 2006 organizers have provided. This makes it more likely that the efficiency of the synthesis system is greatly decreased through the cost calculating and Viterbi searching. Most unit selection systems, using larger speech corpus, have to pre-select candidate units using data-mining technologies or statistical methods [3] [4].

In the paper, we provide a unit pre-selection method called multi-tier NUU selection for the corpus-based speech synthesis. The instants of each NUU in the corpus with the same phoneme sequence were clustered into classes using CART technology according to their prosody and spectral characteristics. Each NUU was described by a Non-uniform Context Dependent Feature Vector (NCDFV). At synthesis time, a series of target non-uniform units were generated from the symbols predicted by the TTS Front-end, each with a NCDFV description. A multi-tier architecture of NUU selection can be obtained with a criterion for good NUU. Using this method, N-best candidate NUUs can be selected into the candidate list for further cost based unit selection process.

The framework of our system is described as Fig.1. This paper is mainly deals with the TTS back-end part, and is organized as follows. Section 2 describes the process of building NUU clustering trees. Section 3 describes unit selection process based on multi-tier NUU selection algorithm. Finally, the evaluation is discussed for future improvement.

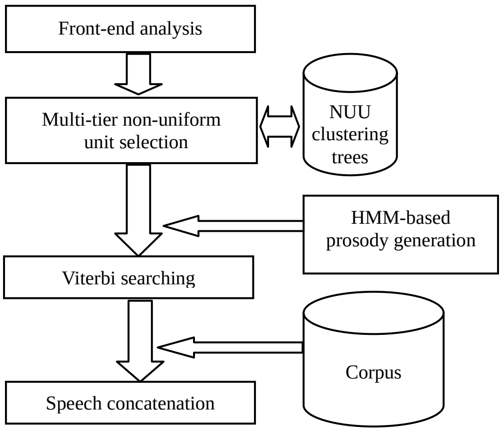

Figure 1: Overview of KB 2006 synthesis system

2. NON-UNIFORM TREE CLUSTERING

The Blizzard Challenge 2006 organizers have released a speech database consists of 4273 phonetically balanced utterances spoken by an English-native male speaker. The database includes speech waveforms recorded at 16 kHz, phoneme segmentations, utterance information files and pitch marks in the Festvox style.

To use the database, LSF parameters were extracted from the waveforms using method described in [5]. The F0 parameters of each syllable were normalized into 15-point samples using interpolation and polynomial curve fitting technology. Each phoneme unit in the database was represented with a feature vector consists of spectrum, F0, duration, and other context dependent features.

2.1 Non-uniform unit definition

The non-uniform unit in our system was defined as one or more joint phonemes that concurrent in the same sentence either from the corpus or from the target symbols to be synthesized. From the definition, we can see that a non-uniform unit can has as many phonemes as possible if the sentence is long enough, but in practice, the number of phonemes in non-uniform unit is limited because of the corpus capability and practicability. The more the number of phonemes a NUU includes, the fewer the number of instants a NUU has in the corpus. As for the Blizzard Challenge 2006 system, we limited the max number of phonemes in a non-uniform unit to 6 for CART training.

2.2 Similar non-uniform units clustering

The fundamental problem in non-uniform units clustering is the definition of similarity measure between all instants of the same non-uniform unit in the corpus. The measure should reflect the subjective distance between them. In this paper, both objective acoustic distance measure and subjective rule-base prosody distance measure were employed for similarity measure as described in Eq.1, where λ is the weight for the two distances.

$$D(A,B) = D(A,B)_o + \lambda \bullet D(A,B)_s \qquad (1)$$

The objective acoustic measure defined as the acoustic distance between units, such as the distances of pitch, duration and spectrum, is described as follows:

$$D(A,B)_o = W_{f_0} D(A,B)_{f_0} + W_{dur} D(A,B)_{dur} + W_{lsf} D(A,B)_{lsf} \qquad (2)$$

where W_{f0}, W_{dur}, W_{lsf} are the weights for the three distances.

$$\begin{aligned} D(A,B)_{f0} &= \left|PMA - PMB\right| + \partial \bullet D(A,B)_{f0delta} \\ D(A,B)_{f0delta} &= [\sum_{i=1}^{N-1}(PA_{i+1} - PA_i - PB_{i+1} + PB_i)^2]^{1/2} \end{aligned} \qquad (3)$$

Eq.3 defines the distance of pitch between two phonemes A and B. The pitch value is the normalized N-points pitch of A and B. The PMA and PMB are individually mean value of the normalized N-point pitch data. PA and PB are pitch data of units A and B. ∂ is the weight to adjust the proportion between the pitch mean distance and the pitch delta distance.

$$D(A,B)_{dur} = | DA - DB | \qquad (4)$$

The duration distance is defined as the duration dispersion between phoneme unit A and B.

$$D(A,B)_{lsf} = DTW(LSFA, LSFB) \qquad (5)$$

The spectrum distance between phoneme unit A and B is based on dynamic time warping (DTW) [6] with type I constraints, which generates an alignment path without right angles and whose end-points are mutual comparisons of the initial and final frames.

The local distance in the DTW alignment is defined as the Euclidian distance of 12 LSF coefficients and its delta coefficient between two phoneme units A and B. Eq.5 describes the spectrum (LSF and LSF delta) distance between phoneme units A and B.

The subjective distances between phonemes are defined as rule based context dependent attributes cost that used as a part of the target cost calculation. If we describe the context dependent attributes of A as $(a_1, a_2, a_3, ..., a_M)$, B as $(b_1, b_2, b_3, ..., b_M)$, then we get

$$D(A,B)_s = \sum_{i=1}^{M} d(a_i, b_i) \qquad (6)$$

In Eq.6 the distances between each attribute $d(a_i, b_i)$ are all manually defined according to the human experience.

Since NUU may contain more than one phoneme, the distances between two candidate non-uniform units are defined as the sum of distances between each corresponding phonemes in that NUU.

For the training process, all the instances belong to the same NUU are gathered as a training dataset. The CART algorithm is adopted to cluster those instants into classes based on their distances described above, using NCDFV as question set for decision tree splitter. After this procedure, instants in the same leaf node are assumed to be having similar prosody and spectrum characteristics according to the distance defined and the CART split criterion.

Nine context dependent features in NCDFV are used in the paper, they are:

- Boundary type in phoneme level: the boundary type before (or after) the first (or last) phoneme in the non-uniform unit, it has four categories.
- Boundary type in syllable level: the boundary type before (or after) the syllable that contains the first (or last) phoneme in the current NUU. It has three categories.
- Maximum internal boundary type: the maximum boundary type between the phonemes of the current NUU. It has four categories
- Position in major phrase: the syllable number between the current NUU and previous (or next) major phrase boundary.
- Phoneme context: categories of phoneme name of the neighbored phoneme before (or after) the current NUU.

For every non-uniform unit in the corpus, clustering trees are built using the same method discussed above. Not all non-uniform clustering trees are built using automatic training method, for some non-uniform units, the instants in the corpus are so sparse that we can’t build any reliable decision trees. In this case, rule based method is used to build those non-uniform units trees from the knowledge of experts.

3. MULTI-TIER NON-UNIFORM UNIT SELECTION AND SPEECH SYNTHESIS

3.1 Multi-tier non-uniform unit pre-selection

In the speech synthesis process, input texts are converted into a sequence of phonetic transcriptions with high-level prosodic descriptions, such as the prosodic hierarchies, stress, accent and breaks, etc. Then, a series of target non-uniform units can be generated from those phonetic transcriptions with different phoneme number and different joint phoneme name. In the unit pre-selection process, N-best candidate units for each target phoneme should be selected from the corpus that best match the prosody and spectrum characteristics of the target synthetic speech. Good non-uniform units, that can match the target ultimately, should be selected ahead.

In the paper, several criterions for good non-uniform unit are described as follows:

- **Criterion 1**: Non-uniform unit that contains more phonemes in it. This makes it more likely that large continuous speech segments can be selected from corpus directly to produce the synthetic speech without suffering from distortion caused by mismatch between concatenated units.
- **Criterion 2**: Non-uniform unit that begins and ends at a large boundary such as syllable boundary, word boundary and major phrase boundary. This makes it possible that whole syllable even whole word can be selected into the candidate list, which makes the system have a chance to synthesis speech with little concatenation and makes synthetic speech more fluent and more natural.
- **Criterion 3**: Non-uniform unit without large boundary type between phonemes in it. In most situations, it is very difficult to select qualified candidate non-uniform units with large boundary type inside.

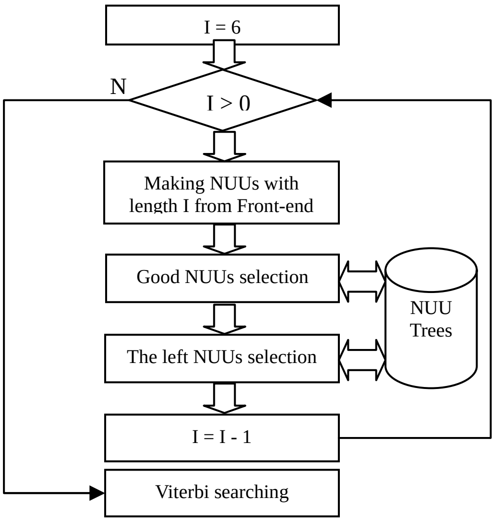

Figure 2: Flowchart of the selection algorithm

The non-uniform unit selection algorithm has a top to bottom architecture according to the criterions described above. In the top level of the process, we prefer to select non-uniform units that have many phonemes. If there are not enough long non-uniform units, we turn to the underside level for short non-uniform units. Finally the bottom level is the non-uniform units that have only one phoneme inside. For each level, the non-uniform units have the same length are grouped into two categories by the criterion 2 and 3. The good non-uniform units were selected at the first step, and the left units were selected last. The method is shown in Fig.2.

3.2 Prosody generation and cost-based unit selection

HMM-based technology [7] were used for prosody parameters prediction, includes pitch and duration. The phoneme duration was modeled by a phoneme duration model combined with a state duration model [8].The pitch was modeled by a multi-space probability distribution (MSD) [9], considering the static and dynamic components of pitch.

During synthesis, pitch parameters are predicted by ML-based parameter generation method [7], the predicted results are one point per 5ms, and are normalized into 15 points per syllable.

After unit pre-selection and prosody parameter prediction, Viterbi algorithm was adopted for best path searching through candidate units. The target costs for each candidate were calculated as follows:

$$c_{tar} = \sum_{i=1}^{M} w_i d(c_i, t_i) \qquad (7)$$

where $d(c_i, t_i)$ is the distance of *i-th* attribute between candidate unit and target unit, the attribute sequence is a M-dimension feature vector includes prosody parameters and context dependent features. w_i is the cost weight of *i-th* attribute which is manually defined according to the human experience.

The joint costs are calculated using following equation:

$$c_{join} = c_{pen} + w_{f0} d_{f0} + w_{\Delta f0} d_{\Delta f0} \qquad (8)$$

The non-continuity penalty value c_{pen} is zero when two phonemes are continuous in the source corpus or otherwise a constant integer. d_{f0} is the pitch distance at the joint boundary between two candidate units to be concatenated, $d_{\Delta f0}$ is the pitch delta distance between two candidate units at joint boundary. w_{f0} and $w_{\Delta f0}$ are the cost weights for the two distances.

The total cost of each candidate for Viterbi searching is:

$$c_{total} = c_{tar} + w_{con} c_{join} \qquad (9)$$

where w_{con} is a weight for target costs and joint cost [10].

4. EXPERIMENTS AND EVALUATIONS

Two systems were built for evaluation and comparison of the effect of method proposed in this paper. They are as follows:

- MTS_TTS: multi-tier non-uniform unit selection based speech synthesis system proposed in this paper.
- CVN_TTS: conventional decision-based phoneme unit selection system without multi-tier non-uniform unit pre-selection algorithm.

The same testing texts, which composed of 100 sentences, were respectively synthesized with the two TTS systems. We logged the finally selected units with their target information, and saved the synthetic speech. Both objective evaluation and subjective evaluation were taken to evaluate the synthetic speech.

In the objective evaluation, the total average costs of finally selected units were logged. Since MOS score is relative with the cost functions used in the Viterbi searching unit selection [11]. Table 1 shows the result of average cost for each selected phonemes.

From table 1, we can see that the MTS_TTS system gets the less cost compared to the CVN_TTS, which indicates that the units selected from MTS_TTS are more close to the target unit than that of CVN_TTS.

Table 1: Average costs of the two systems.

system	cost
MTS_TTS	72.3
CVN_TTS	196.4

In the subjective evaluation, five English-native speakers were asked to give the preference of each sentence from different systems. The result is shown in Fig.3.

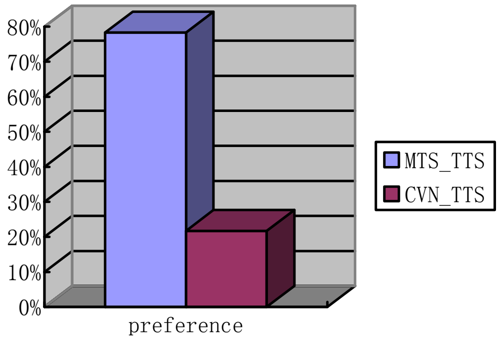

Figure 3: Listener preference of the two systems.

From Fig.3 we can see that system MST_TTS achieved more preference than system CVN_TTS, which reveals that the multi-tier non-uniform unit selection algorithm is effective in the corpus-based speech synthesis, since it can provide larger continuous speech segments from corpus for speech concatenation synthesis.

5. DISCUSSIONS

For Blizzard Challenge 2006, we found that it's difficult for us to synthesize highly natural speech output from the 1 hour ARCTIC database released, especially in the domain of text for news. With a limited corpus, the corpus-based system has little advantages in synthesized prosody compares with the HMM-based parametric synthesis system, but has advantages in the synthesized speech quality since it do not need much signal processing during waveform generation.

There are also many work to do to promote the effect of our speech system, includes:

- Too many rule-based cost definitions and units' similarity measures were used in the speech synthesis system development. It makes that the system performance were strongly affected by the effectiveness of those weights tuned by human experiences. Automatic training methods combine rules should be adopted in the future work [11-13].
- The intonation should be future studied to make synthetic speech more natural and more expressive. And if needed, signal modifications can be adopted to improve the naturalness and expressiveness of speech output.

6. CONCLUSIONS

In this paper, a novel unit selection method called multi-tier non-uniform unit selection was developed in our corpus-base speech synthesis. Similar non-uniform units with the same sequences of joint phonemes in the corpus were clustered into groups using CART algorithm according to their prosody and acoustic characteristics. In the unit selection stage, a multi-tier NUUs selection algorithm was adopted by treating different NUUs with several criterions. With the discrimination; proper candidate units that close to the target unit can be selected for speech concatenation synthesis. Experiments show the effectiveness of our method described in this paper.

7. REFERENCES

[1] A. Hunt and A. Black, *Unit selection in a concatenate speech synthesis system using a large speech database*, in proc. ICASSP. 1996, pp. 373-376.

[2] Ren-Hua Wang, Zhongke Ma, Wei Li, Donglai Zhu, *A Corpus-based Chinese Speech Synthesis with Contextual Dependent Unit Selection*, ICSLP 2000, vol 2, pp. 391-394.

[3] Zhen-Hua Ling, Yu Hu, Zhi-Wei Shuang, Ren-Hua Wang, *Decision Tree Based Unit Pre-selection in Mandarin Chinese Synthesis*, ISCSLP 2002.

[4] W. Hamza, R. Donovan, *Data-driven segment preselection in the IBM trainable speech synthesis system*, proc. ICSLP 2002, (Denver, USA), 2002.

[5] F. Itakura, *Line spectral representation of linear predictive coefficients*, Journal of Acoustic Society of America, 87(4), pp.1738-1752, 1990.

[6] Lawrence, R, and J. Biing-Hwang, *Fundamentals of speech Recognition*, Prentice Hall Signal Processing Series, Englewood Cliffs NJ, USA, 1993.

[7] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., *Speech parameter generation algorithms for hmm-based speech synthesis*, proc. ICASSP, 2000, vol. 3, pp. 1315-1318.

[8] Yi-Jian Wu, *Research on HMM-based Speech Synthesis*, Ph.D Thesis, University of Science and Technology of China, 2006. [in Chinese]

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

[10] John Kominek, Christina Bennett, Brian Langner, Arthur Toth. *The Blizzard Challenge 2005 CMU Entry. a method for improving speech synthesis systems*, proc. Interspeech 2005, pp. 85-88.

[11] Hu Peng, Yong Zhao and Min Chu, *Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation*, Proc. Of ICSLSP2002, Denver, 2002.

[12] Yong Zhao, Peng Liu, Yusheng Li, Yining Chen, Min Chu, *Measuring Target Cost in Unit Selection with KL-Divergence between Context-dependent HMMS*, proc. ICASSP 2006, Vol I, pp. 725-728.

[13] Ann K. Syrdal, Alistair D. Conkie, *Perceptually-Based Data-Driven Join Costs: Comparing Join types*, proc. Interspeech 2005. pp. 2813-2616.