Hierarchical Non-uniform Unit Selection based on Prosodic Structure

Jun Xu[1,2], *Dezhi Huang*[2], *Yongxin Wang*[1,2], *Yuan Dong*[2,3], *Lianhong Cai*[1], *Haila Wang*[2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Speech and Natural Language Processing Unit, France Telecom R&D Beijing, China
[3]Beijing University of Posts and Telecommunications, Beijing, China
{xujun00, wangyongxin}@mails.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn
{dezhi.huang, haila.wang}@orange-ftgroup.com, yuandong@bupt.edu.cn

Abstract

In speech synthesis systems based on wave concatenation, using longer units can generate more natural synthetic speech. In order to improve the usage of longer units in the corpus, this paper proposed a hierarchical non-uniform unit selection framework. Each layer included in the framework is an independent searching procedure which searches for different sized units and adopts suitable naturalness measuring functions related to the unit type. We have applied it to our Mandarin speech synthesis system according to the Chinese prosodic structure with respect to the statistical result in our corpus. Experiment result shows it outperforms our previous system.

Index Terms: speech synthesis, unit selection, non-uniform, prosodic structure

1. Introduction

Nowadays, speech synthesis based on concatenating speech segments from a large scale corpus has become the state-of-the-art technology in most TTS systems [1] [2] [3]. And the size of the segment used in the system, although language dependent, would greatly affect the naturalness of the generated speech. As a result of using longer units, the prosodic and acoustic features at segment boundaries will be well-kept to reduce perceptual discontinuities. However, to deal with the variations of any possible segments and prosodic environments, an unrealistic corpus would be required. Thus one of the key problems of unit selection is to take advantage of any existing long unit for concatenation.

Non-uniform unit selection was introduced in ATR's system [4] in 1988, whichb makes use of all phoneme subsequences in the corpus and concatenate them by pre-defined rules. Synthesis units mixed with N-phone units such as phoneme, diphone and triphone were also used in some systems [5] [6] [8]. To avoid degradation of synthetic speech resulted from concatenating C (consonant) to V (vowel), Tanaka proposed a novel C(V)k unit [9] [10], which denotes a sequenced phonemes starting with a consonant and ending up with k vowels. Another non-uniform unit selection strategy was proposed by Chu [11], which selects the segment of the whole chunk when it exists in the corpus.

The discontinuities in synthetic speech break down the coarticulations of two neighboring units, which cause unnaturalness. However, the coarticulations are not equal at different unit boundaries. For example, syllable is the most commonly used synthesis unit in Mandarin speech synthesis systems, and with a spontaneous prosodic structure, the coarticulation inside a prosodic word is much stronger than that at a prosodic word or prosodic phrase boundary. Therefore it is a good choice to use a whole prosodic word or prosodic phrase as synthesis unit if possible.

Moreover, since the transient part differs at different prosodic boundaries, it is natural to adopt different naturalness measurements at these boundaries. One solution proposed in [12] is clustering these measuring functions by a phonetic decision tree. In order to utilize non-uniform synthesis units, we design a hierarchical unit selection algorithm in our approach, and adopt corresponding measuring criteria with respect to the unit in that layer. We have applied it to our Mandarin speech synthesis system according to the Chinese prosodic structure, and the subjective evaluation approved the effectiveness of the proposed method.

2. Selection schema

The hierarchical selection structure is composed of several layers. Units at each layer could either be directly selected from the corpus or concatenated by one or more units at lower layer. With all the units a selection tree is constructed, of which the root is the utterance to synthesize, and lower layer gradually split the text into smaller prosodic chunks until the basic unit as illustrated in Figure 1.

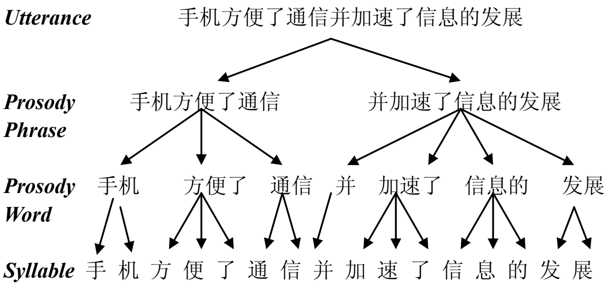


Figure 1: *Example of hierarchical non-uniform units in Mandarin Chinese based on prosodic structure. PPH, PW and SYL are short for prosodic phrase, prosodic word and syllable, respectively. All potential units construct a selection tree.*

Several characteristics of this selection tree can be easily figured out. First of all, a complete unit selection tree should consist of all potential units in the synthetic speech. Secondly, an arbitrary sub-tree also represents a hierarchical unit selection procedure, which makes the selection a recursive procedure, and we can stop at any layer if we have found all the units and they fit well for our result. Thirdly, we can remove any intermediate layer, and when there exists only the basic unit layer, i.e. the bottom layer, we turn to the uniform unit selection algorithm adopted in our early system.

The hierarchical unit selection algorithm can be briefly described as a top-down procedure starting at the root of the tree referred to Figure 1. For each unit, we look for it directly in the corpus, and also try to generate it by concatenating its sub units.

The unit sequences are evaluated by naturalness measuring functions consisting of a target cost function and a concatenation cost function, which are similar to the cost functions defined in [1]. The target cost measures if the candidates are close enough to our target, while the concatenation cost measures how well the F0, spectral shapes and other acoustic features match at both sides of the boundary. The cost functions could be different according to the unit type.

Assume we have a candidate unit sequence u_k with N units, the target cost and concatenation cost is c_k^t and c_k^c, respectively. Then the total cost of the sequence is

$$C = \frac{1}{N}\left[\sum_{k=1}^{N}(c_k^t) + \sum_{k=1}^{N-1}(c_k^c)\right] \tag{1}$$

More than one sequence is kept at each layer as candidates for upper nodes, and each of them is treated as a single unit sample in upper layer and the total cost is its corresponding target cost.

Thus the tree completion comes in a bottom-up order. When the best sequence is determined at tree root, a backward index would help decode which unit to be read from the corpus and concatenate for final speech.

When a unit contains only one sub unit, for example a prosodic word made up of a single syllable, the target costs at both layers are the same.

To search for a unit U in the selection tree, we first try directly reading M_1 samples from corpus. Suppose U could also be constructed from the sub unit sequence u_1, u_2, ... u_k, we select each of them with multiple instances at the lower layer. After that a viterbi decoding is performed with the target costs of all u_k and the concatenation costs between neighboring samples, we get M_2 samples of U by concatenation. Merge the results by their total costs with a descending order, and we get best $M(= M_1 + M_2)$ samples of units U as in Figure 2. Note here M_1 and M_2 are limited for performance issue.

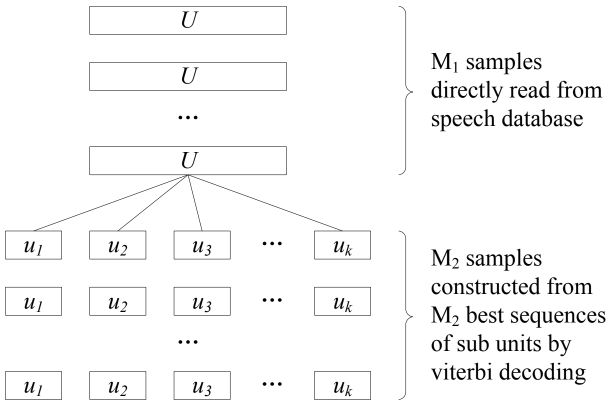

U
U
...
U
M1 samples directly read from speech database
u1 u2 u3 ... uk
u1 u2 u3 ... uk
...
u1 u2 u3 ... uk
M2 samples constructed from M2 best sequences of sub units by viterbi decoding

Figure 2: *Selecting U by both directly reading from speech database or concatenating from sub units.*

3. Implementation in Mandarin speech synthesis

3.1. Prosodic structure

In Mandarin Chinese, the prosodic structure is often simplified to 3 layers [13] (from down to top): prosodic word, prosodic phrase and intonation phrase. Smaller and lower-level units are contained in larger and higher-level units to form a prosodic hierarchy.

Corresponding with our non-uniform unit selection procedure, a 3-layer structure is used: tonal syllable, prosodic word and prosodic phrase, which is already illustrated in figure 1. The intonation phrase is not used in current implementation because most of the utterances in our corpus contain only one intonation phrase.

At a specific layer, target cost and concatenation cost are calculated for everyunit. But actually, the high level target costs are dependent on the lower level ones in current edition. First of all, target costs are calculated for tonal syllables, which are at the leaf of the selection tree. Then the target costs of upper level units, namely prosodic words and prosodic phrases, will sum up all its direct child nodes' target costs according to Formula 1. This makes an assumption that the concatenation cost is 0 when units are neighbors at lower levels. When two units are not next to each other in the corpus, we have to obtain their concatenation cost to judge how well they can be concatenated. The detail will be described later.

3.2. Target cost

Tonal syllables are the synthesis units adopted in our previous system, and they are commonly used in many Mandarin speech synthesis systems. Target cost function at tonal syllable layer is trained from a 2-stage preprocess. Firstly all syllables with the same tonal pinyin are clustered by CART according to contextual information derived from the recorded text in the corpus. That is to make sure the samples within a syllable cluster have small spectral variations.

The features for CART clustering are

- Position in prosodic word
- Position in prosodic phrase
- Position in utterance
- Type of previous tone
- Type of next tone
- Previous phonetic context
- Next phonetic context
- Phonetic type of previous syllable
- Phonetic type of next syllable
- Preceding final class
- Following initial class
- Whether current syllable is retroflex

Three values are used in position features: head, middle and tail. The tone type has 5 values: 4 ordinary tones plus a neutral tone. The previous phonetic context indicates the identities of finals, totally 40, and the next phonetic context contains both the initials and non-initial finals, which has totally 61 values. The previous and next phonetic contexts are also clustered into 6 and 10 categories, respectively. The preceding final class has 4 values and the following initial class has 9 values. Retroflex is a typical characteristic of Mandarin pinyin, and this feature takes a binary value of 0 or 1.

The distance measure for decision tree building uses the first, middle and last frames of the syllable. A feature vector composed of F0, energy, duration and a 13 dimensional MFCC spectra are extracted at these 3 points. After normalization among all syllable samples, Euclidean distance is calculated among these feature vectors.

We have also tried to introduce several frames at the previous and next syllables for calculating a contextual distance but gained no better performance while taking too much preprocess time. This could be explained. When using syllables as synthesis units in our speech synthesis systems, the syllable boundaries are not that stable and have much consistency as those diphone based systems for many western languages. So considering frames near the boundary would not bring much benefit in our environment.

After clustering by CART, a feature vector V_k composed of the 3-point feature vectors of unit u_k mentioned before is prepared to train a Gaussian probability distribution function. If V_0 denotes the feature vectors of the leaf center, and Σ is the covariance matrix for all vectors in the leaf node, we can get the probability how close u_k is to our target as in Equation 2.

$$\begin{aligned} p_k^t &= P(V_k) = \mathcal{N}(V_k|V_0,\Sigma) \\ &= \frac{1}{(2\pi)^{\frac{15}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{(V_k-V_0)^T\Sigma^{-1}(V_k-V_0)}{2}\right] \end{aligned} \tag{2}$$

With this probability, we map it to target cost c_k^t with a predefined constant C_0.

$$c_k^t = C_0(1-p_k^t) \tag{3}$$

As stated earlier, we didn't design a special target cost for upper levels, so when target cost is identified for each tonal syllable, the target costs for prosodic words and prosodic phrases can be quickly calculated and cached.

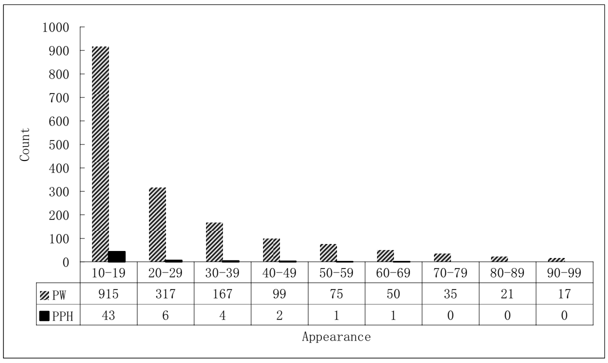
3.3. Concatenation cost

When two units are next to each other in the corpus, the concatenation cost is directly assigned 0; otherwise, we have 3 common concatenation cost functions

- Pitch differences, measured as Euclidean distance on F0 and its delta value at the boundaries.
- Spectral continuities by Euclidean distance on MFCC spectra, proved to be useful for Mandarin Chinese in [14].
- Phonetic context distance. A discrete function to evaluate the mismatch of two initials or finals.

The first two functions are easy to understand. The phonetic context distance requires some explanation. Suppose we have two adjacent unit pairs, u_k, u_{k+1} and v_k, v_{k+1} in the corpus, and we want to concatenate u_k to v_{k+1} in our synthetic speech, the phonetic context distance between them would be the distance of the finals of u_k and v_k, added up the distance of the initials of u_{k+1} and v_{k+1}.

Since coarticulations are much stronger in smaller units, only the tonal syllable layer has a very complex concatenation cost function. And in Mandarin Chinese, the pinyin of a syllable are made up of initials and finals, while some of them contain only finals. To differ the transient part from vowel to vowel (including voiced consonant) and others, two different distance measures are designed for syllable boundaries.



	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
▨PW	915	317	167	99	75	50	35	21	17
■PPH	43	6	4	2	1	1	0	0	0

Figure 3: *Unique prosodic word and prosodic phrase count with appearances from 10 to 100.*

The boundaries of the syllables from vowel to vowel or voiced consonant have clear F0 contour and spectrum shape, thus both pitch differences and spectral continuities are computed. F0 and MFCC from the frames at both sides of the boundary are extracted for the distance measures. The concatenation cost for the other syllable boundaries and prosodic word boundaries only contains the phonetic context distance. There is an additional distance measure for prosodic word layer. The average F0 should be dropping from the previous word to the next one according to prosody prediction result. For prosodic phrase layer, the concatenation cost is always set to 0 since there is often an obvious pause at the boundaries.

4. Experiments

4.1. Speech corpus

The speech corpus we used in our system comes from the TH-CoSS Mandarin corpus of Tsinghua University, which is designed to create, test and evaluate Mandarin speech synthesis systems. The most recent edition contains sub-databases of 5000 sentences of Mandarin corpus from reading material and 1000 sentences for testing. All sentences are tagged with pinyin and pitch contour, and are divided into syllables, prosodic words and prosodic phrases according to Mandarin's prosodic hierarchy. This work is automatically done with manually correction.

Not all the prosodic word and prosodic phrase are necessary to be indexed for selecting. Actually, most of the prosodic words appear only once in the corpus, still less does a prosodic phrase. We made a statistic analysis of the corpus, and a partial result is shown in Figure 3. The prosody units composed of a single sub unit were excluded.

The corpus produced totally over 30,000 unique prosodic words and 40,000 unique prosodic phrases. About 60% of prosodic words and 95% of prosodic phrases only have a single instance. And finally we decided to index the prosodic words and prosodic phrases who appear at least 10 times in the corpus. That doesn't bring lots of extra disk cost.

4.2. Experiment results

All utterances for test are randomly selected from People's Daily 1998. They have 32 syllables in average.

The first experiment is designed to find how many prosodic units will be directly selected. We used 1000 utterances and the non-unique count produced by the utterance and the directly selected counts are recorded. Here prosodic units who have only one sub unit were skipped. As expected, the result in Table 1 shows about 30% PWs are directly selected but the PPHs are difficult to be matched as a whole.

Table 1: *Statistical result of PW & PPH counts*

prosodic unit	produced count	selected count	percentage
PW	5379	1646	30.6%
PPH	2122	13	0.6%

The second experiment is a subjective preference test to compare the results from current system to previous one, based on basic cost functions. Twenty utterances are synthesized by both engines for comparing. Five listeners took part in the experiment to give a preference score and they didn't know from which engine the speeches were generated in advance.

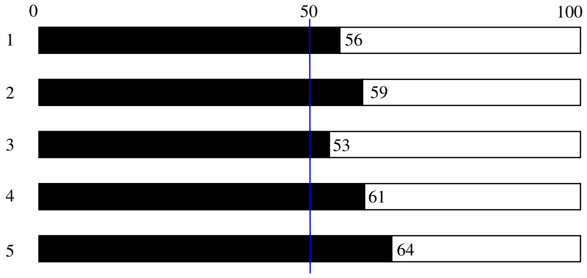


Figure 4: *Preference test result from 5 subjects. The dark bar is the percentage that listeners prefer the result from current system.*

The new selection framework did gain better performance. The major reason for that is the usage of longer units are improved in current system, which avoid the discontinuities of acoustic and spectral features. The hierarchical selection methods searches for high-level prosodic units, and the unit-related cost functions also generate a smooth transient part for different boundaries.

Meanwhile, it is easy to find out the improvement is not significant yet. This is not a surprising result because so far we didn't apply a specific model for high-level prosodic units. A better result can be expected after introducing specific prosody prediction at higher layer.

5. Conclusion and future work

In this paper, we presented a hierarchical unit selection strategy, which reflects the internal structure of speeches and is fit for human perception. With consideration to the transient part at different unit boundaries, the selection method also enables adopting suitable naturalness evaluation functions related to the units at different layers. A set of non-uniform units driven by prosodic structure was proposed in this framework and implemented in our Mandarin Chinese speech synthesis system. In systems for other languages, the non-uniform units may vary corresponding to the prosodic hierarchy they have.

Future work should be focused on the prosodic model corresponding to the structure we used, in order to obtain a better target model for prosodic word and prosodic phrase layers in our selection hierarchy. Further experiments are also scheduled to find correlations between human perception and measuring functions.

6. Acknowledgements

The work in this paper is supported by China National Natural Science Foundation (No. 60433030, No. 60418012) and the Special Funds for Major State Basic Research Program of China (973 Program) (No. 2006CB303101). And we would like to thank the reviewers for their kindly comment.

7. References

[1] Andrew J. Hunt and Alan W. Black, “Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database”, in Proc. of ICASSP 1996, vol.1, pp 373-376, Atlanta, Georgia, 1996.

[2] Robert A. J. Clark, Simon King, “Joint Prosodic and Segmental Unit Selection Speech Synthesis”, in Proc. of Interspeech 2006, Pennsylvania, USA, September 17-21 2006.

[3] Geert Coorman, “Segment Connection Networks for Corpus-based Speech Synthesis”, in Proc. of Interspeech 2006, Pennsylvania, USA, September 17-21 2006.

[4] Yoshinori Sagisaka, “Speech Synthesis by Rule using an Optimal Selection of Non-uniform Synthesis Units”, in Proc. of ICASSP, pp 449-452, 1988

[5] A. P. Breen, “Non-uniform unit selection and the similarity metric within BT's Laureate TTS system”, 3th ESCA Workshop on Speech Synthesis, 1998

[6] Minkyu Lee, “A text-to-speech platform for variable length optimal unit searching using perceptual cost functions”, 4th ISCA Workshop on Speech Synthesis, 2001

[8] Tomoki Toda, “Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit”, in Proc. of ICASSP 2002, Orlando, Florida, May 13-17, 2002.

[9] Kimihito Tanaka, “A Japanese text-to-speech system based on multi-form units with consideration of frequency distribution in Japanese”, in Proc. of Eurospeech 1999

[10] Satoshi Takano, “A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction”, pp 3-10, Vol9, No.1, IEEE Trans. on Speech and Audio Processing, Jan. 2001.

[11] Min Chu, Hu Peng, Hongyun Yang and Eric Chang, “Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer”, in Proc. of ICASSP 2001, Salt Lake City, 2001.

[12] Shinsuke Sakai and Tatsuya Kawahara, “Decision Tree-based Training of Probabilistic Concatenation Models for Corpus-based Speech Synthesis”, in Proc. of Interspeech 2006, Pennsylvania, USA, September 17-21 2006.

[13] Xiaonan Zhang, Jun Xu and Lianhong Cai, “Prosodic Boundary Prediction based on Maximum Entropy Model with Error-Driven Modification”, in Proc. of ISCSLP 2006, Singapore, Dec 2006.

[14] Jun Xu and Lianhong Cai, “Spectral Continuity Measures at Mandarin Syllable Boundaries”, in Proc. of ISCSLP 2006 (Companion Volume), Singapore, Dec 2006.