Selecting Optimal Non-uniform Units for Hierarchical Unit Selection

Jun Xu[1], Dezhi Huang[2], Yuan Dong[1], 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
Xujun00@mails.tsinghua.edu.en, dezhi.huang@orange-ftgroup.com,
clh-dcs@tsinghua.edu.en

Abstract

For concatenative speech synthesis based on non-uniform unit selection, the key to improve the synthetic quality is the careful designing of measuring criteria respect to the units adopted. With our previous hierarchical non-uniform unit selection framework [1], two measurements for selecting optimal non-uniform units during searching at different layers are proposed in this paper, including inter-syllable pitch control and spectra distance by phonetic context. These measures are used as components of our cost function, especially for boundaries in front of syllables starting with voiceless consonants. Experiment shows it outperforms our previous system.

1. Introduction

Concatenative speech synthesis systems are able to generate high quality synthetic speech [2]. Such systems often carry a large speech corpus providing quantities of wave segments, and a unit selection technique is adopted for choosing suitable segments from those instances. Recent studies have focused on the degradation of speech quality brought by audible distortions during concatenations [3] [4]. A solution is adopting non-uniform unit selection, which makes use of any longer segments that already exist in the corpus.

As a result of utilizing non-uniform units, the prosodic and acoustic features at segment boundaries are well kept and the perceptual discontinuities are reduced. However, the criteria for choosing best units should be carefully designed respect to the unit type and length. Wave unit whose FO contour most closely matches that of the target was preferred in Takano's Japanese systems [5]. And Lee suggested a perceptual cost function based on acoustic features and phonetic environment to match the longest possible candidates [6].

In many speech synthesis systems, pitch contour is one of the most important measures while searching for a candidate [7], which plays an important role in both accordance measures with target and continuity measures with neighbors [8]. Since Mandarin Chinese is a tonal language, we proposed an inter-syllable pitch control strategy classified by prosodic context when searching for syllables and prosody words.

Another important measure is the spectral continuities at unit boundaries, which has been carefully studied recently. The common approach is adopting statistical distance for spectra coefficients, sometimes together with their time derivatives [9] [10]. An alternative method is classifying the degradation of concatenations by decision tree for search path pruning [11]. And the distance could be altered by a probabilistic concatenation model [12]. To measure how the spectra of a vowel are affected by its following phonetic environment, especially before a voiceless consonant, a spectral distance based on phonetic context is also introduced in this paper.

2. Hierarchical unit selection

2.1. Framework

Our unit selection framework managed to split the search procedure into several layers [4]. And prosody structure is used as the non-uniform hierarchy in our implementation of Mandarin Chinese speech synthesis system for its close relation to the human perception and personal individuality [13]. 3 types of non-uniform units: syllable, prosody word and prosody phrase are therefore used as our synthetic units, as shown in figure 1.

In the hierarchical selection framework, different layers are relatively independent. Each layer performs an independent searching for a typical unit, without considering whether it is good enough for concatenating upper layer unit.

978-1-4244-1724-7/08/$25.00©2008IEEE

Authorized licensed use limited to: IEEE Xplore. Downloaded on December 14, 2008 at 22:42 from IEEE Xplore. Restrictions apply.

To give an example, suppose we are going to search for a prosody phrase U from prosody words sequence u_{l7} u_2, ... u_N. And U might already have some samples in the corpus, which can be used immediately. When evaluating all the instances of U, either directly read from the corpus or concatenated from sub units, we only consider the phonetic, prosodic and acoustic features inside U, but throw away all of information from outside, such as the position of u_k and U in the utterance, or the coarticulations from other prosody phrases.

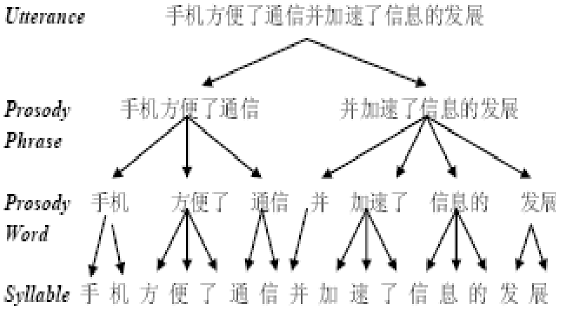


Fig. 1. Example of hierarchical non-uniform units in Mandarin Chinese based on prosodic structure. All potential units construct a selection tree.

2.2. Cost functions

Cost function based measurements are applied in the search procedure, which normally has two aspects. The target cost measures if the candidates are close enough to our target, while the concatenation cost measures how well the FO, spectral shapes and other acoustic features match at both sides of the boundary. We assume the target cost and concatenation cost for u_k are c_k^t and c_k^c respectively. And since u_k might also be concatenated from their sub units, we add a third cost component c_k^u to indicate the total cost of searching at lower layer. Then the cost for the concatenated U is

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

Note when u_k is directly read from corpus, c_k^u is 0. With these cost, we perform the viterbi search to find out the best N results of U.

In our previous system, target cost for syllable layer is modeled by a GMM based on CART. At prosody word layer, target cost is calculated as the total costs of its subsequences. When a prosody phrase is selected as a whole, the target cost is always set to 0. For concatenation cost, 3 types of measure functions are used, including pitch differences and spectral distances at the boundaries, and a predefined phonetic context distance measuring how an initial affects its previous final. However, the first two distances are only effective at boundaries with clear pitch contour and spectra shape. For syllables starting with voiceless consonant, only the phonetic context distance could be applied. In order to overcome this problem, we'll describe our new implementation of cost function components on pitch and spectra measures in next two sections.

3. Adjacent pitch control

When searching for certain unit given the target pitch, we prefer specifying a relative value or range according to its context, rather than assigning an exact value. Thus we introduced an adjacent pitch control strategy here for both target cost and concatenation cost.

Let F_{uk-1} and F_{uk} be the pitch of two neighboring candidates u_{k-1} and u_k, respectively, and $R_k = F_{uk} / F_{uk-1}$. We model the adjacent pitch according to phonetic context feature X by a single GMM.

$$p(R_k \mid X) = \mathrm{N}(R_k \mid R_0, \Sigma)$$
$$= \frac{1}{(2\pi)^{\frac{N}{2}} \mid \Sigma \mid^{\frac{1}{2}}} \exp\left[-\frac{(R_k - R_0)^T \Sigma^{-1} (R_k - R_0)}{2}\right] \qquad (2)$$

The factors of feature X used are listed in Table 1.

Table 1. Prosodic and phonetic context features for adjacent pitch control modeling. T means item used as target cost component, C means item used as concatenation cost component, B denotes both, and X means none. PPH and UTTER are short for prosody phrase and utterance, respectively.

factors	average	range	boundary
initial of u_k	B	T	C
tone of u_{k-1} & u_k	B	T	C
boundary type	B	T	C
pos. in PPH/UTTER	T	T	X

In above table, initial of u_k is a boolean value indicating if u_k starts with voiceless consonant. For target cost, pitch average and pitch range are used. And for concatenation cost, boundary pitch value is used. The position factors is regarded less important in boundary pitch controlling.

With the GMMs trained from the corpus, we set a tolerant parameter to determine whether two adjacent syllables are allowed to be concatenated.

Authorized licensed use limited to: IEEE Xplore. Downloaded on December 14, 2008 at 22:42 from IEEE Xplore. Restrictions apply.

4. Spectral distance based on phonetic context

4.1. Distance measures

When a syllable starts with voiceless consonant, the boundarybefore it would have no clear pitch contour and spectra shape, so we have to use other criteria to judge whether it is a good concatenation. The previous phonetic context distance was a solution for solving this problem, which could be described as following. We classify the initials and finals into several categories, among which a distance value is defined between any two categories. Assume we have "$A\psi$↖←B" and "C›←D" in the corpus. When computing the concatenation cost from $A\psi$to D, the phonetic context distance is calculated as the final distance between $A\psi$ and $C\psi$summed up with the initial distance between $B\psi$ and D. In order to perform a mathematical distance other than expert opinion, we introduce a reference boundary to calculate spectra distance. Let L^A and F^D ψbe the feature vectors from last few frames of final of $A\psi$and the first few frames of initial of D, respectively, then

1. Looking for two neighboring instances, such as "$A^/ - D^/$", in which $A^/$ ←has same finals as A, and $D^/$ ←has same initials as D.

2. If we managed to make $D^/$ ←have very similar phonetic environment as candidate D, we have $F^{D^/}$ ←=· F^D ψin ideal.

3. As "$A^/ - D^/$" are neighbors in the corpus, we may treat $L^{A^/} \approx F^{D^/}$.

4. The distance from $A\psi$to $D\psi$is then calculated as

$$\begin{aligned} D(A,D) &= D(L^A, F^D) \\ &= D(L^A, F^{D^/}) \qquad (3) \\ &\approx D(L^A, L^{A^/}) \end{aligned}$$

4.2. CART clustering

As described above, we treat the transient part of "$A^/ - D^/$" as a reference boundary from L^A $\psi$to F^D, which is regarded as the best concatenation from A's final to D's initial when $D^/$ ←has similar phonetic environment as D, we used CART to classify all L^A ψ by context information of boundary "$A\psi$↖←D" including

- Entities of A's and D's initial
- Boundary type of "A-D"
- Boundary position
- Tone entity of A and D
- A's and D's initial class
- Whether A is retroflex

The distance measure for decision tree building uses the last 3 frames of A. A feature vector composed of F0, energy, and 13-dimensional MFCC spectra are extracted at these frames with normalization among all samples. Mahalanobis distance is then computed, as in Figure 2.

The distance calculation at synthesis time is similar. The actual phonetic context information of "$A\psi$↖D" is used to search for a reference boundary "$A^/ - D^/$" in decision tree. Then we have

$$\begin{aligned} D(A,D) &= D(L^A, L^{A^/}) \\ &= D_{M_a}(L^A, L^{A^/}) = \sum_{i=1}^{N} \left[\frac{L_i^A - L_i^{A^/}}{\sigma_i} \right]^2 \qquad (4) \end{aligned}$$

where σ_{i} is standard deviation of the i^{th} feature of L^A and $L^{A'}$, and N is the dimension of the feature we used near the boundary. In our configuration, N = 45.

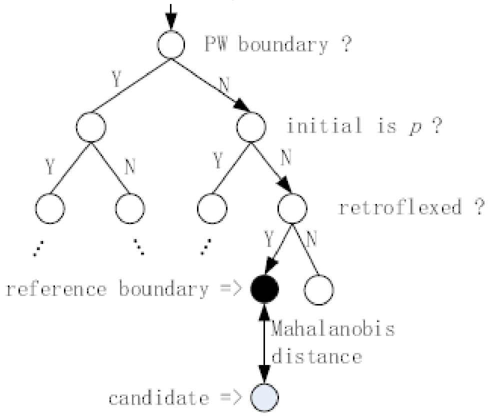


Fig. 2. Searching for reference boundary $A^/ - D^/$ in phonetic context tree, and then compute Mahalanobis distance between L^A and $L^{A^/}$.

5. Experiment

5.1. Corpus

The speech corpus we used in our system comes from the TH-CoSS Mandarin corpus of Tsinghua

Authorized licensed use limited to: IEEE Xplore. Downloaded on December 14, 2008 at 22:42 from IEEE Xplore. Restrictions apply.

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 manual correction.

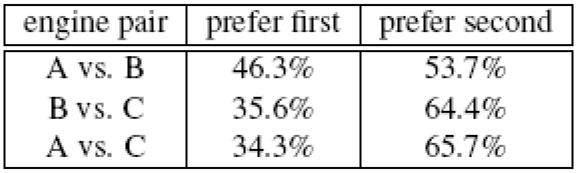
5.2. Experiment and result

A subjective preference test compared the results among 3 Mandarin speech synthesis engines with different selection strategies.

1. Engine A: Syllable based with cost function.
2. Engine B: Prosodic units based with hierarchical framework.
3. Engine C: Current implementation with updated cost functions.

20 utterances were synthesized by three engines for comparing. Utterances from any 2 of 3 engines' output were paired for comparing, producing 60 pairs in total. 10 listeners took part in the experiment to give their preference score for each pair. The orders of engines and utterances are randomly sorted so all listeners didn't know in advance. The test result is in listed in Table 2.

Table 2. Preference test results for 3 engines.



engine pair	prefer first	prefer second
A vs. B	46.3%	53.7%
B vs. C	35.6%	64.4%
A vs. C	34.3%	65.7%

The result shows although the improvement from engine B with prosodic units to engine A with ordinary cost functions is not significant, but the new engine C, which can be regarded as the engine B added to updated cost functions performs much better than our previous two systems.

6. Acknowledgement

This work is supported by China National Natural Science Foundation (60433030,60418012) and the Special Funds for Major State Basic Research Program of China (973 Program) (No. 2006CB303101).

7. Summary

Based on our previous design of hierarchical unit selection framework based on prosody structure, we proposed two distance measures for selecting optimal prosodic units, including pitch control and spectra distances measures classified by phonetic context. These two measurements were applied at boundaries before syllables starting with voiceless consonants as an addition of our previous costs as well as other boundaries, and the synthesis quality is improved.

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, and more concatenation cost components should be experimented as well as weight tuning.

8. References

[1] Jun Xu, Dezhi Huang, Yongxin Wang, Yuan Dong, Lianhong Cai and Haila Wang, "Hierarchical Non-uniform Unit Selection based on Prosodic Structure", in Proc. of INTERSPEECH 2007, Antwerp, Belgium, August 27-31, 2007.

[2] 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.

[3] Jithendra Vepa and Simon King, "Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis", pp 1763-1771, Vol.14, No.5, IEEE trans, on Audio, Speech and Language Processing, Sept. 2006.

[4] Tomoki Toda, Hisashi Kawai, Minoru Tsuzaki and Kiyohiro Shikano, "An Evaluation of Cost Functions Sensitively Capturing Local Degradation of Naturalness for Segment Selection in Concatenative Speech Synthesis", pp 45-56, Vol.48, No.l, Speech Communication, Jan. 2006.

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

[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.

[7] Robert A. J. Clark, Simon King, "Joint Prosodic and Segmental Unit Selection Speech Synthesis", in Proc. of

Authorized licensed use limited to: IEEE Xplore. Downloaded on December 14, 2008 at 22:42 from IEEE Xplore. Restrictions apply.

INTERSPEECH 2006, Pennsylvania, USA, September 17-21 2006.

[8] Dacheng Lin, Yong Zhao, Frank K. Soong, Min Chu and Jieyu Zhao, "Iterative Unit Selection with Unnatural Prosody Detection", in Proc. of INTERSPEECH 2007, Antwerp, Belgium, August 27-31, 2007.

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

[10] Barry Kirkpatrick, Darragh OBrien, Ronan Scaife and Andrew Errity, "On the Role of Spectral Dynamics in Unit Selection Speech Synthesis", in Proc. of INTERSPEECH 2007, Antwerp, Belgium, August 27-31, 2007.

[11] Nobuyuki Nishizawa and Hisashi Kawai, "A Preselection Method Based on Cost Degradation from the Optimal Sequence for Concatenative Speech Synthesis", in Proc. of INTERSPEECH 2007, Antwerp, Belgium, August 27-31, 2007.

[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.

Authorized licensed use limited to: IEEE Xplore. Downloaded on December 14, 2008 at 22:42 from IEEE Xplore. Restrictions apply.