SELECTING NON-UNIFORM UNITS FROM A VERY LARGE CORPUS FOR CONCATENATIVE SPEECH SYNTHESIZER

Min Chu, Hu Peng, Hong-yun Yang, Eric Chang
Microsoft Research China
Beijing, P.R.C
Email: minchu@microsoft.com

ABSTRACT

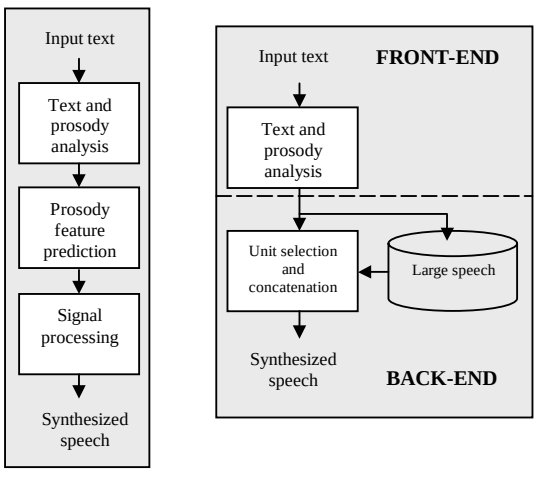
This paper proposes a two-module TTS structure, which bypasses the prosody model that predicts numerical prosodic parameters for synthetic speech. Instead, many instances of each basic unit from a large speech corpus are classified into categories by a CART, in which the expectation of the weighted sum of square regression error of prosodic features is used as splitting criterion. Better prosody is achieved by keeping slender diversity in prosodic features of instances belong to the same class. A multi-tier non-uniform unit selection method is presented. It makes the best decision on unit selection by minimizing the concatenated cost of a whole utterance. Since the largest available and suitable units are selected for concatenating, distortion caused by mismatches at concatenated points is minimized. Very natural and fluent speech is synthesized, according to informal listening test.

1. INTRODUCTION

The conventional concatenative TTS systems normally have three modules, as in Figure 1(a), which are text processing module, prosody prediction module and signal processing module. First, text taken as input is converted into a sequence of phonetic transcriptions (phonemes, diphones, semi-syllables or syllables) with high-level prosodic descriptions, such as the prosodic hierarchies, stress, focus, and breaks, etc. Then, an 'appropriate' set of prosodic contours, such as fundamental frequency, duration and amplitude, is calculated by the prosody module. At last, pitch and duration modification algorithm, such as PSOLA, is applied to pre-stored units to guarantee that the prosodic features of synthetic speech meet the predicted target values. These systems have the advantages of flexibility in controlling of prosody. Yet, they often suffer from significant quality decrease in timbre. Mechanical or reverberant sounds are two typical distortions that can be perceived in the synthesized speech. [1] and [2] present schemes to select a proper unit from multiple instances with varied spectral features to achieve better smoothness between concatenated units. [3] puts forward a unit selection method that takes variations of both spectral and prosodic features into account to reduce the extent of signal processing that is required to correct the prosodic characteristics of selected instances. However, they still claim that even the best selection will not in general exactly match the desired utterance and further signal processing will still be required to modify the selected units. Above unit selection schemes are applied on speech databases with the size changing from 10 minutes to 3 hours. However, how many pre-stored instances are enough for concatenative synthesis, in which no signal processing is needed, and how to obtain the best representative instances for each unit are still issues that haven't been studied well.

This paper investigates the above issues based on an ultimate assumption that we have a very large speech corpus that contains enough prosodic and spectral varieties for all synthetic units. This assumption is valid under a constraint that the whole corpus retains the same speaking style, which is referred as the "relax reading style", the same speech rate and the same timbre. Since no pitch or duration modification will be applied to the selected units before concatenation, a two-module TTS structure (see Figure 1(b)) is adopted in our approach. It bypasses the prosody module that generates numerical prosodic features in most of the conventional TTS systems. Though, the text and prosody analysis module is also very important, this paper is only limited to techniques used in the back-end.

The paper is organized as follows. Section 2 discusses the designing, annotating and indexing techniques for the speech corpus that is used in our approach. A 15-hour Mandarin speech corpus has been used in our case. Section 3 presents a multi-tier and non-uniform unit selection scheme. Finally, summary of our major findings and outline for the future works are given in Section 4.



(a) (b)

Figure 1. (a) The Conventional three-module TTS structure; (b) Our two-module TTS structure

2. PREPARATION OF SPEECH CORPUS

The unit selection method presented in this paper is valid under the assumption that we have a very large speech corpus containing enough prosodic and spectral varieties for all synthetic units. The quality of the resulting synthetic speech will depend to a large extent on the variability and availability of representative units. It is crucial to design a corpus that covers all speech units and most of their variations in a feasible size.

2.1 Choosing Proper Synthesis Unit

For a concatenative speech synthesizer, there exit several possible choices for basic synthesis unit, such as phonemes, diphones, demisyllables, syllables, words or phrases. Both smaller units and larger units have advantages and disadvantages. On one hand, it is not difficult to collect a speech corpus that embodies many prosodic and spectral varieties of small units like phonemes. Yet, it is almost impossible to cover many varieties of larger units such as words or phrases in a feasible size. On the other hand, since small units mean much more concatenation points, the synthesized speech tends to suffer more distortions caused by mismatches between concatenated units.

Mandarin is a syllabic based language and it has regular CV structural syllables. Very strong co-articulation can be found between phonemes in a same syllable, yet, co-articulations between phonemes across syllable boundaries are weaker[4]. Tonal syllables are chosen as the basic units in our synthesis approach for Mandarin. There are about 1600 tonal syllables, taking neutral tone and tone sandhi into account. When synthesizing, non-uniform and larger units are used when they are available.

2.2 Text Script Generation

Tonal syllables are the basic synthesis units in our approach. A multi-dimensional *descriptive contextual variation vector* (DCVV), which can be derived directly from text, is used to represent all possible prosodic and spectral variations of each syllable. Six coordinates are used in this paper. They are:

- *Position in phrase* (PinP): position of current syllable in its carrying prosodic phrase. It takes 4 values.
- *Position in word* (PinW): position of current syllable in its carrying prosodic word. It takes 4 values.
- *Left phonetic context* (LeftPh): category of the final of the left-neighbored syllable. Left-neighbored finals are classified into 11 categories.
- *Right phonetic context* (RightPh): category of the initial of the right-neighbored syllable. Right-neighbored initials are classified into 26 categories.
- *Left tone context* (LeftT): category of the tone of the left-neighbored syllable. Left-neighbored tones are classified into two categories.
- *Right tone context* (RightT): category of the tone of the right-neighbored syllable. Right-neighbored tones are classified into two categories.

Among them, PinP, PinW, LeftT and RightT are factors mainly contributing to prosodic variations, and, LeftPh and RightPh are factors reflecting the co-articulation effects between syllables. There are 4*4*11*26*2*2=18304 possible instances for each syllable. However, not all these instances occur in real text.

A text corpus of five-year People's Daily, which contains about 97 million Chinese characters, is used as the raw corpus for statistic. All characters are converted into *syllable-dependent DCVVs* (SDDCVV). A SDDCVV consists of a syllable name and its DCVV. Totally 1521 different tonal syllables and 2.3M different SDDCVVs (referred as occurred SDDCVV) are found in the text corpus. After sorting all occurred SDDCVVs decreasingly with their *frequencies of occurrence* (denoted as OccurFreq), we found that the accumulated total of OccurFreqs of the top 44k SDDCVVs (they are only 1.9% of all occurred SDDCVVs) is larger than 50%. They are the most frequently used contextual varied syllables and are put into a *list of necessary vectors* (LNV) listing all SDDCVVs to be covered by the speech corpus. One more constraint for generating LNV is to contain at least 10 DCVVs for each Mandarin syllable. At last, a LNV of 46k SDDCVVs with their OccurFreqs is generated.

The weighted greedy algorithm[5] is used in selecting a subset corpus from the raw text corpus to cover all DCVVs in LNV. The reciprocal of the OccurFreq attached to each SDDCVV is used as the weight to minimize the size of the selected corpus. Totally 12016 sentences are selected, which contains 177k Chinese characters and 119k different SDDCVVs (5.15% of all occurred SDDCVVs). The additional 73k SDDCVVs raises the accumulated total of OccurFreqs of the covered SDDCVVs to 63.93%. That is to say that we will have about 64% chance to get a syllable from the speech corpus, whose SDDCVV is exactly the same as the one required by synthetic speech. At the remaining 36% of the time, the exact SDDCVV required has not been recorded. A good unit selection algorithm is needed for finding out a 'best' unit with the most similar SDDCVV. The distribution of sentence length (in characters) in the generated text script is given in Figure 2.

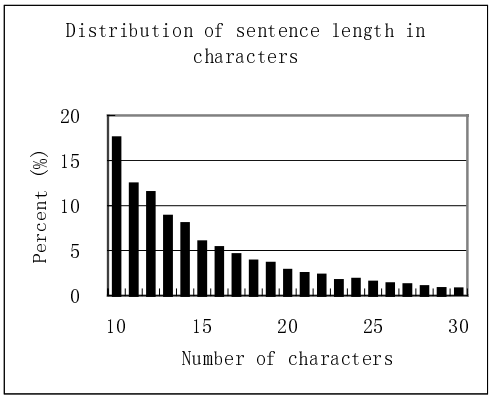


Figure 2. Distribution of sentence length in characters in the generated script for our Mandarin speech corpus.

2.3 Speech Collection And Annotation

All 12016 sentences are recorded in a professional studio by a professional speaker, who is asked to read these sentences in "relax reading style", which is between "formal reading style" and "free talk style", in moderate speed. If there are hesitations or mistakes in a sentence, the sentence will be read again until correct. Pinyin transcriptions for all sentences are generated automatically and potential error points are verified manually.

Syllable boundaries are labeled by force alignment tools in HTK and are verified manually.

All sentences are annotated with *perceptual prosodic boundary index* (PPBI), whose value ranges from 1 to 4 and the values correspond to prosodic word, accentual phrase, intermediate phrase and intonation phrases in prosodic hierarchies.

Since Chinese has no visual cue for word boundaries, PPBI is annotated between syllables (or Chinese characters). A 1 is assigned between two syllables where a smallest perceptible prosodic boundary is heard and a 2 is assigned when a small and weak break is perceived. If a distinct but not very long break is perceived, a 3 is assigned and a 4 is assigned for a long distinct break. The PPBI is annotated by well-trained annotators. The consistency of an annotator over time and comparability between different annotators and the relationship between values of PPBI and the constituents of prosodic hierarchy are interesting research topics we are working on. Constrained by space, we will provide details of this work in a later paper.

2.4 Indexing The Speech Corpus By Prosody-Dependent Decision Trees

Though we do not have a module for predicting numerical prosodic features in our TTS approach, prosodic variation is still the most important factor that should be handled carefully, when selecting a unit from many different candidates. A method of indexing all instances of each unit in the large speech corpus by a *prosody-dependent decision tree* (PDDT) is proposed. Automatic *classification and regression tree* (CART) is used to generate the indexing trees automatically. CART is chosen for its reputation of handling missing data problem and its robustness to outliers and mislabeled data samples.

All syllables in the 15-hour speech corpus are converted to and represented by their SDDCVVs. Simple questions about prosodic related coordinates in SDDCVV, such as PinP, PinW, LeftT and RightT, are listed in the question set. Composite-questions are constructed automatically during the growing phase of a tree to avoid the over-fragmented problem[6].

The splitting criterion for CART is greatest reduction of Expected Square Error (ESE), which is defined as the expectation of the weighted sum of the square regression error of prosodic features. Three prosodic features are used in our case, which are average fundamental frequency (F_0Av) and duration (DurAv) of a syllable and dynamic range of F_0 (F_0Range) in a syllable. Suppose F_a, F_b and F_c are actual F_0Av, DurAv and F_0Range for training data X, the ESE for node t is given by equation (1):

$$ESE(t) = E(W_a E_a + W_b E_b + W_c E_c) \quad (1)$$

where E_a, E_b and E_c are square error for F_a, F_b and F_c respectively, and is defined by equation (2). W_a, W_b and W_c are weights for the three square errors.

$$E_j = \left|F_j - P_j(X)\right|^2, j = a, b, c \quad (2)$$

$P_a(X)$, $P_b(X)$ and $P_c(X)$ in equation (2) are the regression values of F_a, F_b and F_c.

A question q is picked for splitting when it maximizes the reduction of ESE defined as:

$$\Delta WESE(t) = ESE(t)P(t) - (ESE(l)P(l) + ESE(r)P(r)) \quad (3)$$

where l and r are leaves of node t, and $P(t)$, $P(l)$ and $P(r)$ are the percentages of data samples belonging to the three nodes.

For each basic unit, instances with significant variations in prosody are classified into different leaf nodes of its indexing tree. While many minor variations in prosody, which may be caused by different phonetic context, different tonal context, or any other unknown factors, are still kept in instances on the same leaf node. These minor differences bring in slender diversity in prosody of synthetic speech. And they are very helpful for getting rid of monotonous prosody.

Currently, all instances are preserved in leaves of indexing trees and they all take part in the unit selection. Various pruning techniques are being studied. They will prune these trees to make the speech database scaleable and well balanced between the resource requirements and perceived naturalness.

3. MULTI-TIER NON-UNIFORM UNIT SELECTION ALGORITHM

Before unit selecting phase starting, input text will be converted into a string of SDDCVV (referred as target). In our approach, an intonation phrase is processed in one unit selection loop. The task for selection algorithm is to find out the ideal instance with a same SDDCVV as the one required by a target unit efficiently when such an instance exits in the source corpus, or, pick up a 'best' instance that differs from the required one in only insignificant aspects, when an ideal instance is missing from the source corpus. A multi-tier non-uniform unit selection method, shown in Figure 3, is presented below. It takes a string of target SDDCVVs as input and shrinks the pool of candidate units step by step before coming to the final decision.

Tier1: a leaf node is selected for each SDDCVV in the target string by answering questions in its indexing tree. The instances on all the selected leaves form a lattice or unit for next tier selection. Since all instances on a selected leaf node have the same or similar prosodic environment, such as PinP, PinW, LeftT and RightT, as the target SDDCVV, they are believed to have similar prosodic features to those required by the target unit.

Tier2: The major differences between instances on the same leaf node are their phonetic context. They are pruned by keeping only the top N candidates with the smallest *contextual distance* (denoted as D_c). D_c is defined as the weighted sum of the distances between each coordinates in source and target DCVVs, as in equation (4):

$$D_c = \sum_{i=i}^{I} W_{ci} D_i \quad (4)$$

where, D_i is the distance between the ith coordinate in source and target DCVV and is given by a pre-defined distance matrix. I is the dimension of DCVV. In our case, $I = 6$. W_{ci} is the weight for normalizing and could be hand-tuned with subjective listening test.

D_c models the contextual mismatching effects between the selected unit and the desired target.

The unit selection phase could stop by selecting a candidate with the smallest D_c for target SDDCVV. However, in order to minimize the chances of mismatches between neighbored units at

concatenation points, a third step is required to find out the largest available segment from the source corpus. Only top N candidates with the smallest D_c are kept and they form a pruned lattice of units, which is used in next selection step.

Tire3: with Viterbi decoding process, the best path is determined by minimizing the concatenated cost (denoted as C_c) of the synthesized utterance. C_c is the weighted sum of contextual distance and smoothness cost of a synthesized utterance:

$$C_c = W_c \sum_{i=1}^{I} D_c(i) + W_s \sum_{i=1}^{I-1} C_s(i) \qquad (5)$$

where $D_c(i)$ is the contextual distance defined in formula (4) for the *i*th unit, and $C_s(i)$, the *smoothness cost*, is defined to model the mismatches between the unit *i* and its succeeding neighbor. In our case, if two concatenated units are continuous segments in source corpus, 0 is assigned to C_s. Otherwise, 1 is assigned. So, if a larger chunk, such as a word or a phrase in a text string to be synthesized, exists in the source corpus, and if the two are in similar contextual environment, the segment of the whole chuck will be selected for concatenation.

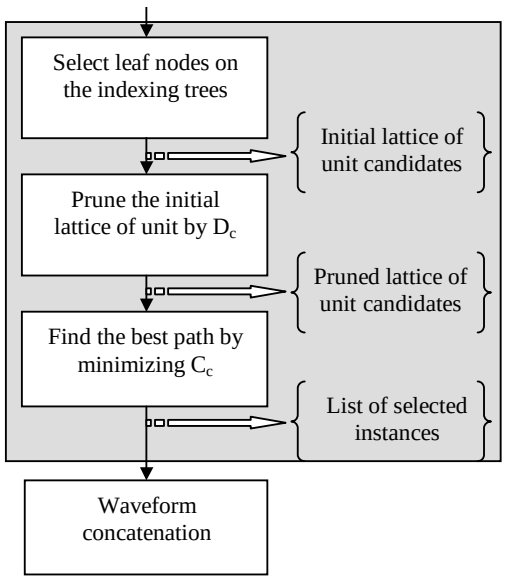


Figure 3. Flowchart of the proposed multi-tier non-uniform unit selection algorithm

Since no signal processing, except for very limited smoothing at the concatenated points, is applied to the selected units when they are concatenated, the synthetic speech sounds very similar to the original speaker and preserves the original speaking style. Due to time constraints, no formal assessment has been carried out to evaluate the quality of the synthesized speech. Feedbacks from internal listening test and external demonstrations confirm that speech synthesized by the proposed method is more natural and fluent than those generated by conventional three-module systems.

4. CONCLUSION

This paper has proposed a two-module TTS structure, which does not have a prosody model to predict numerical target values for prosodic features. Instead, CART is trained to index the instances of each basic unit, in which all instances are differentiated by their prosodic variations. Though, instances having significant variations in prosodic features are classified into different categories, many trivial variations in prosody are still kept in the same categories. Slender diversity in prosody can be perceived in the synthetic speech, which is helpful for generating natural sound speech.

The presented multi-tier non-uniform unit selection method makes final decision of choices under the condition of minimizing the concatenated cost of the synthesized utterance. A simple two-value smoothness cost is defined in our approach for finding out larger units. Since the largest available segments are selected from source corpus, the distortion caused by mismatches at concatenated points is minimized. Though the two-value smoothness cost works well in the current stage, where a very large speech corpus is used, more precise cost function should be defined, when speech database is pruned and not many larger units can be found.

Currently, all segments in our 15-hour speech corpus are indexed and used in the unit selection procedure. We believed that there are many redundancies. Techniques for making the speech database scaleable and well balanced between the resource requirements and perceived naturalness are under studying.

Methods for evaluating the quality of synthesized speech and for finding out the influences of different factors on naturalness are in designing.

5. REFERENCES

[1] Wang, W. J., Campbell, W. N., Iwahashi, N. and Sagisaka, Y., "Tree-based unit selection for English speech synthesis", ICASSP'93, vol.2, 191-194

[2] Hon, H., Acero, A., Huang, S., Liu, J. and Plumpe, M., "Automatic generation of synthesis units for trainable text-to-speech systems", ICASSP'98, vol.1, 293-296

[3] Black, A. and Campbell, N., "Optimizing selection of units from speech database for concatenative synthesis", ICASSP'96, 373-376, 1996

[4] Chu, M., Tang, D., Si, H., Tian, X. and Lu, S., "Research on perception of juncture between syllables in Chinese", Chinese Journal of Acoustics, Vol.17, No.2, 143-152

[5] Sproat, R., editor, Multilingual text-to-speech synthesis: the Bell labs approach, Kluwer Academic Publishers, 1998, 17-20

[6] Huang, X.D., Acero, A., Hon, H. and Meredith, S., Spoken Language Processing(draft), chapter 4