VARIABLE-LENGTH UNIT SELECTION USING LSA-BASED SYNTACTIC STRUCTURE COST

Chung-Hsien Wu, Chi-Chun Hsia, Jiun-Fu Chen, Te-Hsien Liu

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan
{chwu, shiacj, cama, liu}@csie.ncku.edu.tw

ABSTRACT
This paper introduces a variable-length unit selection method based on LSA-based syntactic structure for concatenative speech synthesis. First, a probabilistic context free grammar (PCFG) based parser is used to construct the syntactic structure of the input text sentence. Second, the synthesizer selects the candidate units for each node of the syntactic structure. Latent Semantic Analysis (LSA) is then adopted to estimate the syntactic cost between the target unit and the candidate units in the database. Finally, the concatenation of units with minimum cost is selected using dynamic programming algorithm. Experimental results show that variable-length unit selection based on syntactic structure outperforms the synthesizer without considering syntactic structure. Also, the LSA-based syntactic cost provides better estimation of substitution cost than that calculated only from acoustic features.

1. INTRODUCTION

Recently, unit selection method has become the predominant approach in speech synthesis and captures the attention of most researchers. The process of unit selection in current speech synthesizers is based on some type of dynamic programming that selects units from large database with minimal cost functions. This means that the synthesizer needs to select one possible concatenation of candidate units of an input sentence. For syllable-based synthesizer, with 10 candidate examples of each syllable of input sentence that is N syllables long, the number of alternative combinations is $10^N$. Selection process from large speech corpora was achieved by considering substitution and concatenation cost functions. Beside, researchers have long been trying to improve the naturalness of the synthesized speech by increasing the length of basic units from demi-phones, phones, di-phones, tri-phones, syllables [1], words to non-uniform [2] or variable-length units [3]. Longer synthesis units will reduce the effect of spectral distortion [4].

In order to consistently synthesize natural speech, it is important to use the costs that correspond to perceptual characteristics. Many cost functions based on acoustic features are proposed. Textual features such as the phone type, the location in the sentence or the preceding/succeeding phone type are also used in some synthesizers. Moreover, some researchers study on the combinations of different costs in linear or non-linear methods based on listening test [5] or discriminative training [6].

In the past, not many synthesis unit selection methods considering syntactic structure were proposed [7][8]. In these approaches, searching for all possible combination of non-uniform units is not efficient and necessary. For example in the word sequence “里长城” “里” and “长城” are the candidate units but “里长城” is not a unit for consideration.

A variable-length unit selection scheme based on the syntactic structure and LSA-based syntactic cost function are proposed. Fig. 1 shows the system architecture. After text format processing, the input sentence is parsed into a syntactic structure. Consequently, the structure is vector-quantized into a vector and projected onto the lower dimensional space. The substitution cost for each candidate is estimated by the syntactic cost which calculated from the dis-similarity measure using VSM and joint inside probability. The concatenation cost is calculated from the difference of line spectrum pair frequency (LSF) from the adjacent frames of two successive units. The unit sequence $\hat{u}_{1,N}$ with minimum cost is selected according to the cost function:

$$\hat{u}_{1,N} = \arg \min_{u_{1,N}} \left[ C_S(u_1,u_1') + C_C(u_1,u_2) + C_S(u_2,u_2') \right]$$

where $C_S(\cdot,\cdot)$ and $C_C(\cdot,\cdot)$ represent the substitution and concatenation costs, respectively. $u_{1,N}$ denotes the candidate unit sequence of length N. Finally, the combination with minimum cost is selected using the dynamic programming (DP) algorithm [10] to give concatenative speech output. This study focuses on the estimation of the substitution cost based on LSA.
2. PROBABILISTIC CONTEXT FREE GRAMMAR

Given an input text sentence for speech synthesis, the possible combinations of synthesis units are searched through each node of the corresponding syntactic tree structure. Fig. 2 gives an example for the target sentence “観光旅遊是臺灣地區的主要收入.” This figure illustrates an example of all the possible unit fragments. The possible units would be “觀”, “觀光” or “觀光旅遊” and so on. PCFG is the simplest and most natural probabilistic model for tree structure representation of text sentence. For an input sentence with \( T \) words represented as \( W_{t} = w_{1}, w_{2}, \ldots, w_{t} \), the probability of the derivation from the root of sentence \( N_{0} \) is given by:

\[
P(N_{0} \Rightarrow W_{t} | G) = \text{ Prob}(N_{0} \Rightarrow T | G)
\]

\( \) is the PCFG model.

The variable-length units are selected from the word sequences obtained from Sinica Tree-Bank [9] to parse the syntactic structure of the input sentence. The maximum probability between all possible parse trees is given by:

\[
\text{Max} \left[ \sum_{i} P(N_{0} \Rightarrow W_{t} | G) \right] = \text{ Prob}(N_{0} \Rightarrow T | G)
\]

The probability for the derivation from a non-terminal \( A \rightarrow \alpha_{i} \) with the most similar syntactic structure in the speech database. The variable-length units are selected from the word sequences \( w_{m,n} \) and a unit \( \tilde{w} \) (a word sequence) is calculated as:

\[
P(N_{0} \Rightarrow w_{m,n}, \tilde{w} | G) = \gamma(\tilde{w}, m, n, w_{m,n})
\]

where the delta function is defined as:

\[
\delta(m, d, w_{m,n}) = \begin{cases} 1, & m = d \text{ and } n \\
0, & \text{otherwise}
\end{cases}
\]

2.2. Joint Inside Probability

The variable-length units are selected from the word sequences with the most similar syntactic structure in the speech database. The probability for the derivation from a non-terminal \( N_{j} \) to a word sequence \( w_{m,n} \) and a unit \( \tilde{w} \) (a word sequence) is calculated as:

\[
P(N_{j} \Rightarrow w_{m,n}, \tilde{w} | G) = \gamma(\tilde{w}, m, n, w_{m,n})
\]

where the delta function is defined as:

\[
\delta(m, d, w_{m,n}) = \begin{cases} 1, & m = d \text{ and } n \\
0, & \text{otherwise}
\end{cases}
\]
\[ \delta(m, n, \tilde{w}) = \begin{cases} 1, & \text{if } \tilde{w} \text{ is a substring of } W_{m,n} \\ 0, & \text{otherwise} \end{cases} \]

Consequently, the syntactic parse tree with maximum probability is constructed by the Viterbi algorithm using the equation

\[ \tilde{p}(m, n, \tilde{w} | G) = \max_{w_{m,n}} \left[ \frac{P(N_r \rightarrow N_{i,j} W_{i,j}, \tilde{w})}{\sum_{w_{m,n}} \frac{P(N_r \rightarrow N_{i,j} W_{i,j}, \tilde{w})}{C(N_r \rightarrow N_{i,j} W_{i,j})}} \right] \]

where \( P(Rule\ r : N_r \rightarrow N_{i,j} W_{i,j}, \tilde{w} | G) \) is given by:

\[ P(Rule\ r : N_r \rightarrow N_{i,j} W_{i,j}, \tilde{w} | G) = \sum_{q} C(N_r \rightarrow N_{i,j} W_{i,j}) \]

and \( \varepsilon_r \) is the normalized entropy of the rule. The expression for \( \varepsilon_r \) is defined as:

\[ \varepsilon_r = -\frac{1}{\log Q} \sum_{q} \frac{C(N_r \rightarrow N_{i,j} W_{i,j})}{C(N_r \rightarrow N_{i,j} W_{i,j})} \log \frac{C(N_r \rightarrow N_{i,j} W_{i,j})}{\sum_{q} C(N_r \rightarrow N_{i,j} W_{i,j})} \]

where, \( W_{i,j}^{(q)} \) represents the qth word sequence in the database, and \( Q \) is the total number of sentences. A value of \( \varepsilon_r \) close to 1 indicates that the rule is distributed in many documents throughout the database, while a value of \( \varepsilon_r \) close to zero means that the rule appears only in a few specific sentences. Hence, \( 1 - \varepsilon_r \) represents a global indexing weight for rule \( Rule\ r : N_r \rightarrow N_{i,j} \).

The cost function considers only acoustic features including line spectral pairs and average energy. An additional set of speech sentences spoken by the same speaker is recorded as the target speech. The cost function defined in Eq. (1) and the dynamic programming algorithm is used to select the unit sequence with minimum cost to give the concatenated speech.

### 4. EXPERIMENTAL RESULTS

To evaluate the proposed method, we collect a database of text corpus with 4,212 sentences and the number of words in the corpus is 68,392. It covers the 1,342 tonal syllables of Mandarin speech and the average appearing time of each tonal syllable is 51.79. The PCFG model is initiated using Eq. (2) based on the Sinica Tree-Bank Corpus in which the syntactic tree structures are manually annotated. Then the EM algorithm [11] is adopted to estimate the probability of each rule based on the corpus. The speech corpus is spoken by a female speaker. The size of speech corpus is 72.21 hours. An HMM-based speech recognizer is used to automatically segment the utterances into syllable segments. Speech signal is sampled at 22.05 kHz and analyzed with a frame size of 10 ms.

In the first experiment, the performance of arbitrary variable-length and syllable-based unit selection schemes were evaluated. Two sets of test sentences were synthesized separately. The cost function considers only acoustic features including line spectral pairs and average energy. An additional set of speech sentences spoken by the same speaker is recorded as the target speech for evaluation. Each set has 50 sentences different from the speech corpus. There are 10 listeners participated in the listening test. The order of the synthesized speech utterances is random for each listener. Table 2 gives the result for the preference listening test. The result shows that longer unit provides more natural sounding speech.

Table 3 shows the preference listening test result for arbitrary vs. syntactic structure variable-length unit selection. The same cost function in the previous experiment is used and the 50 sentences in the experiment for each method are different from the previous experiment. Although the number of possible candidates from arbitrary variable-length unit selection scheme is more than those from syntactic structure, some units are still
insufficient. The result shows that no significant difference between them was obtained.

The effectiveness of syntactic cost is also evaluated. The unit selection scheme in this experiment is based on the syntactic structure. Two sets of 50 test sentences are synthesized with the LSA-based syntactic structure cost and acoustic cost, respectively. 10 listeners in experiment one also conducted this experiment. Table 4 shows the result of the preference listening test. The syntactic cost function outperforms the acoustic cost function by 10.8%.

Fig. 3 illustrates an example of prosodic parameter sequences of the mean pitch period, duration of syllable and average energy for two speech sentences. One is synthesized using proposed synthesizer and the other is spoken by the speaker who provided the speech corpus. The synthesis unit sequence is “

<table>
<thead>
<tr>
<th>Prefer arbitrary variable-length</th>
<th>No difference</th>
<th>Prefer syllable-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.7%</td>
<td>2.4%</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

Table 2. Preference listening test for arbitrary variable-length vs. syllable-based unit selection scheme

<table>
<thead>
<tr>
<th>Prefer syntactic structure</th>
<th>No difference</th>
<th>Prefer arbitrary</th>
</tr>
</thead>
<tbody>
<tr>
<td>35.2%</td>
<td>24.6%</td>
<td>40.2%</td>
</tr>
</tbody>
</table>

Table 3. Preference listening test for arbitrary vs. syntactic structure unit selection scheme

<table>
<thead>
<tr>
<th>Prefer LSA-based syntactic cost</th>
<th>No difference</th>
<th>Prefer acoustic cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>55.7%</td>
<td>9.4%</td>
<td>34.9%</td>
</tr>
</tbody>
</table>

Table 4. Preference listening test for LSA-based syntactic cost and acoustic cost function

This work proposed a variable-length unit selection scheme using LSA-based syntactic cost function. Longer unit will reduce the spectral discontinuity, but insufficient long units reduce the naturalness of synthesized speech. Syntactic structure provides a significant approach to search the candidate units similar to natural speech. The proposed LSA-based syntactic cost function provides an effective estimation of substitution cost. The results show a promising performance on variable-length unit selection for Mandarin speech synthesis.

5. CONCLUSION

This work proposed a variable-length unit selection scheme using LSA-based syntactic cost function. Longer unit will reduce the spectral discontinuity, but insufficient long units reduce the naturalness of synthesized speech. Syntactic structure provides a significant approach to search the candidate units similar to natural speech. The proposed LSA-based syntactic cost function provides an effective estimation of substitution cost. The results show a promising performance on variable-length unit selection for Mandarin speech synthesis.

6. REFERENCES