PROSODY MODIFICATION ON MIXED-LANGUAGE SPEECH SYNTHESIS

Yi Zhang, Jianhua Tao

Institute of Automation, Chinese Academy of Sciences

ABSTRACT

This paper proposes a method to generate natural prosody parameters in Chinese and English mixed-language speech synthesis system which is based on separate Chinese, English, and bilingual corpora. On the other hand, a mixed-language speech synthesis system based on separate Chinese, English, and a small bilingual corpus can hardly generate natural prosody synthesis system based on separate Chinese, English, and bilingual corpora. However, most research focuses on modifying monolingual prosody parameters to adapt for mixed-lingual environment. Experiments have proved that more natural mixed-lingual prosody can be generated with our model.

Index Terms— Chinese and English mixed-language speech synthesis, prosodic assimilation, prosody modification

1. INTRODUCTION

In the past few years, the quality of speech synthesis systems has been clearly improved since the use of large corpus-based concatenative speech synthesis methods. However, most speech synthesis systems can only handle a single language, which is not adequate for many applications dealing with multiple languages. For example, there are many Chinese sentences with English expressions or proper names embedded in. Therefore, mixed-language speech synthesis systems are in great demand.

There are many research aspects to multilingual Text-to-Speech(TTS) systems. In [1], methods for language identification have been described. In [2] focus on text analysis, while [3] concentrate on multilingual prosody modeling. Besides, some multilingual TTS systems have been developed. For example, TTS system designed by AT&T Bell Labs[4], festival system[5], ACTOR TTS system developed at Loquendo S.P.A.[6], Microsoft Mulan[7]. Our work focuses on prosody modification in Chinese-English mixed-language TTS systems.

As known, it’s laborious and time-consuming to build a bilingual corpus. On the other hand, a mixed-language speech synthesis system based on separate Chinese, English, and a small bilingual corpus can hardly generate natural prosody because of the great prosody differences between languages. English is a non-tonal language, and its pitch contour is stable, while Chinese is a tonal language with five tones including neutral tone, whose pitch contour is fluctuant.

In bilingual sentences, Chinese contexts will influence English words, which will change the original English prosody. In order to establish a mixed-language speech synthesis system based on separate Chinese and English corpora, we focus on the influences of Chinese contexts on embedded English words, and propose a prosody modification model to generate natural mixed-lingual prosody.

This paper is organized as follows: Section 2 analyzes how English words are influenced by Chinese contexts. Section 3 presents a prosody modification model for pitch and duration. Experimental results are reported in Section 4. We conclude and discuss future work in Section 5.

2. INFLUENCE OF CHINESE CONTEXTS ON ENGLISH WORDS

The situation we usually meet is: some English words are embedded in Chinese texts, especially in IT-related articles, emails and modern teaching systems. For example: "我们经常使用微软WORD编辑文档"(We usually edit documents with Microsoft WORD). This needs a mixed-language speech synthesis with a pre-selected primary language according to the multilingual speech synthesis classification in [8], in which the main language of a document is identified as primary language of the synthesizer. In our system, primary language and secondary language are Chinese and English respectively. Prosody of English words can be influenced by Chinese contexts and assimilated to that of Chinese, which makes its pitch range wider and duration longer. These can be used to help adjust English prosody in order to generate natural mixed-lingual prosody.

2.1. Corpus Preparation

It’s necessary to build a bilingual corpus in order to observe the influences of Chinese on English words. We designed bilingual corpus according to statistical result of English word’s length in Festival lexicon, which has 105894 words totally. We find that 99.37% of these words have no more
than 5 syllables. Therefore we pay most attention on the English words with no more than 5 syllables.

We randomly select 50 English words with 1 to 5 syllables uniformly distributed, and design 500 sentences with these words embedded in as our corpus. Among these, half are Chinese-English bilingual sentences, and the rest are English sentences for comparison. The corpus is recorded in silent circumstances by a female Chinese native speaker.

2.2. Influence of Chinese Contexts on pitch of English words

Our work concentrates on the stable part of English words and pays little attention on coarticulation at boundaries between Chinese and English parts. Based on our observation, we find the English pitch contours are influenced by Chinese contexts obviously. Fig. 1 illustrates the contrast of pitch contours of the word "conference". The Y axis is pitch, whose value ranges from -50HZ to 450HZ, with 50HZ as increment. The left figure shows the pitch contour of "conference" in sentence: "After presenting her findings yesterday at a conference", and the right one shows it in sentence "员工代表conference定于明天举行". Table 1 shows fundamental pitch ($F_0$) distribution of English words in bilingual corpus and English corpus.

![Fig. 1. Pitch contours of the word "conference"](image)

Table 1. $F_0$ distribution of English words in bilingual corpus and English corpus. The unit is HZ

<table>
<thead>
<tr>
<th></th>
<th>$\mu_{F_{\text{max}}}$</th>
<th>$\mu_{F_{\text{min}}}$</th>
<th>$\mu_{\text{Avg}}(F_0)$</th>
<th>ASTD($F_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilingual</td>
<td>317.5</td>
<td>181.4</td>
<td>250.8</td>
<td>51.9</td>
</tr>
<tr>
<td>English</td>
<td>271.5</td>
<td>204.6</td>
<td>136.1</td>
<td>29.8</td>
</tr>
<tr>
<td>difference</td>
<td>46</td>
<td>-23.2</td>
<td>66.9</td>
<td>22.1</td>
</tr>
</tbody>
</table>

Table 2. $F_0$ distribution of English words and Chinese words in bilingual corpus. The unit is HZ.

$\text{Avg}(F_0)$ is the average value of $F_0$. We can figure out that all statistical features listed above of English words in bilingual corpus are almost equivalent to those of Chinese words adjacent to them. This means $F_0$ of English words are influenced by Chinese contexts around them and they have similar pitch range and $F_0$ mean value.

2.3. Influence of Chinese Contexts on duration of English words

Another important prosodic feature is duration. Fig. 2 demonstrates how the average duration of English words in English and bilingual corpus changes when the syllable number increases. We can see that the average duration of English words in bilingual corpus is longer than that in English corpus. This reflects the duration assimilation of English words to Chinese contexts.

![Fig. 2. Average duration(ms) of English words in different corpora](image)

Table 2. $F_0$ distribution of English words and Chinese words in bilingual corpus. The unit is HZ.

Another important prosodic feature is duration. Fig. 2 demonstrates how the average duration of English words in English and bilingual corpus changes when the syllable number increases. We can see that the average duration of English words in bilingual corpus is longer than that in English corpus. This reflects the duration assimilation of English words to Chinese contexts.

3. PROSODY MODIFICATION MODEL

When building a mixed-language speech synthesis system based on separate English and Chinese corpora, prosody of English words is much different from mixed-lingual prosody.
according to observation above. It needs to be modified in order to give natural mixed-lingual prosody.

Our system selects phone as English basic units, and uses class and regression tree (CART) to predict pitch and duration of each phone.

3.1. Pitch Modification Model

Since the pitch range of English words is larger in bilingual corpus than that in English corpus, we define a model to change the pitch:

\[ Y_{new} = \lambda \ast \frac{\text{astd}_{Ch}}{\text{astd}_{En}} \ast (Y_p - \mu_p) + \mu_{Ch} \]  

(1)

Where \( Y_{new} \) is the new pitch of a phone after modification, \( \text{astd}_{Ch} \) and \( \text{astd}_{En} \) are the average standard deviation of Chinese and English respectively, \( Y_p \) is the original output pitch predicted by CART, \( \mu_p \) is the mean value of \( Y_p \), \( \mu_{Ch} \) is the mean value of pitch of Chinese contexts in bilingual corpus. \( \lambda \) is an additional modification coefficient which is relevant to the English word length and the speaker.

This pitch modification model allows the pitch range extend \( \lambda \ast \frac{\text{astd}_{Ch}}{\textd_{En}} \) times wider than its original value, and make the mean value of modified pitch close to \( \mu_{Ch} \).

3.2. Duration Modification Model

According to Fig. 2, the ratio of duration of English words with specific number of syllables in bilingual corpus to that in English corpus are the five points shown in Fig. 3. According to their distribution, these points can be fitted by a conic:

\[ \text{ratio} = a \ast \text{sylnum}^2 + b \ast \text{sylnum} + c \]  

(2)

Fig. 3. Distribution of English words duration ratio

The new duration of the current English phone \( \text{dura}_{new} \) is:

\[ \text{dura}_{new} = \text{ratio} \ast \text{dura}_p \]  

(3)

Where \( \text{dura}_p \) is the original predicted duration. This formula can lengthen the duration \( \text{ratio} \) times longer than its original predicted value.

4. EXPERIMENTS AND DISCUSSIONS

300 bilingual sentences are used for testing the prosody modification model. English words in these sentences are target words, whose pitch contours are carefully annotated. Pitch and duration of these words are predicted, and then the modification model is used on predicted results. Mean Square Error (MSE) is used as evaluation criterion.

4.1. Evaluation of prosody modification model

In formula (1), \( \text{astd}_{Ch} \), \( \text{astd}_{En} \), \( \mu \) and \( \lambda \) are 1.74, 250.8 and 0.9 respectively based on our corpus. From Fig. 4 it becomes apparent that after modification, the new pitch contour is much closer to the target contour than the original predicted one, and it goes up and down more quickly.

Fig. 4. A sample of comparison among pitch contours. The line marked (1) is the target pitch contour. The line marked (2) is the new pitch contour after modification. The line marked (3) is the original predicted pitch contour.

To evaluate the performance of pitch modification model, we first align new pitch with target pitch with Dynamic time warping (DTW). We find that after modification the MSE decreased by 41.88% in average. Fig. 5 gives the result of MSE comparison and shows that MSE decreases obviously in most samples. It shows 16 pairs of sentences for clarity.

Fig. 5. Comparison of MSE(HZ) before and after pitch modification. The left bars are MSE between original predicted pitch and target pitch. The right bars are MSE between modified pitch and target pitch.
However, sample 4 and 11 are two exceptions, that’s because the original predicted pitch is bad. The pitch modification model is a linear function, and it can hardly attain good modification result when the original predicted pitch isn’t satisfying in shape.

After generating new pitch contours with pitch modification model, the duration modification model is used to get new duration of English words. Based on our experiments, in formula (2), coefficient $a$, $b$ and $c$ are 0.0379, -0.2421 and 1.51, respectively. After modification, the MSE decreases by 58.92% in average. The MSE decreases obviously after modification, see Fig. 6.

![Fig. 6. The comparison of MSE(ms) before and after duration modification. The left bars are MSE between original predicted duration and target duration. The right bars are MSE between modified duration and target duration.](image)

4.2. Subjective Evaluation

In our subjective evaluation, 50 synthetic bilingual sentences are used. 23 of them have English words without using our model, whose prosody are generated directly by CART model. The rest have English words with prosody modified by pitch and duration modification models. This subjective evaluation focuses on prosodic naturalness, ignoring the speech quality. Fig. 7 shows the average evaluation result of 10 participants, which reveals that the prosody generated by the prosody modification model is much better than that without modification and closer to that of the real speech. This proves our analysis and assumption.

![Fig. 7. Subjective Evaluation Result. The left bar is the average score of synthetic speech without our model, the middle one is the average score of synthetic speech with our model, the right one is the average score of real speech.](image)

5. CONCLUSION AND FUTURE WORK

This paper analyzes the prosodic influences of Chinese contexts on embedded English words and the prosodic assimilation of English words to Chinese. We present a prosody modification model based on the analysis results. This model can modify the pitch range and the duration for generating natural mixed-lingual prosody for Chinese-English sentences.

However, this model also has some shortcomings. Firstly, we can’t get satisfying modified pitch contour unless the original predicted result has similar shape to the target contour. Secondly, the parameters in the model base on specific corpus, which leads to inefficiency.

The pitch modification model just concentrates on the stable part of English words and pays little attention on coarticulation at boundaries between Chinese and English parts, which is also a very important affect on prosody. This gives a guideline for our future work.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No.60575032, No.70611120555) and 863 Programs (No.2006AA01Z138, No.2006AA01Z194).

7. REFERENCES