An Empirical Study on Break Prediction of Prosody for Hakka’s TTS Systems

Fong-Long Huang, Ming-Shing Yu, Jun-Yi Wu, Shu-Tsai Gue

Abstract
This paper aims at the prosody generation for Hakka’s language based on the data mining approaches, and implement the TTS system on Internet. Our system is composed of the following four components: 1) Text analysis, 2) Prosody prediction, 3) Selection of speech unit, and 4) Speech generation module. More than 2427 monosyllabic speech units and 2234 word speech units of Hakka and several silences with various durations have been recorded as basic units for speech synthesis.

We focus on adding breaks to speeches, with emphasis on predicting the types of break. There are three kinds of breaks: major break, minor break and no-break between words. We train a break model and predict break based on the data mining approaches - Bayesian network (BN) and CART classifier. The best precision rate for testing achieves 79.88% based on the CART.

Keywords: Hakka Language, Text-to-speech (TTS) system, Break Prediction, Data Mining.

1. Introduction
Speech processing (SP) studies the domain of speech signals and the processing methods of these digital signals. It is always combined into natural language processing (NLP). The technology development is widespread day after day; the information system with speech service already became important tendency. Speech Processing may divide into two broad domains: Speech Recognition and speech synthesis. The former is to recognize the speech signal with respect to the text output and the latter is to synthesize the speech with frequent prosody for the text or articles inputs.

The speech synthesis system is also called the Text-to-Speech (TTS) system; its application is very widespread and popular, like the man-machine interface design, language studies, language translation, directory assistance and e-book, and so on. TTS system may be applied for multilingual translation and online e-learning in web. User may hear digital content and learn the language with the natural and prosody speech [11].

1.1. Previous Works on Hakka’s TTS
In the past, most researches focus on Mandarin speech and language processing, such as Pu-Tong-Hua. There are also several dialects in Taiwan, such as the well-known Taiwanese (MinNan) and Hakka languages.

The thesis [1] implemented a TTS system of Hakka language, in which there are more than 36000 items in Hakka’s lexicon and 671 basic speech units without tone for synthesis. The recurrent neural network (RNN) is employed to generate the prosody features. The speeches are modulated and synthesized by PSOLA (Pitch Synchronous OverLap-Add) while two special stop tones (higher and lower stops) in Hakka’s language are neglected. PSOLA will lead to the unclear speeches because of nasal effectiveness.

The system is not implemented on Web, which is necessary for language e-learning. [2] also implement an e-learning system on web, while they didn’t study the prosody generation for Hakkas.
1.2. Features of Chinese and Hakka

Chinese and Hakka are both tonal languages with monosyllable. There are 408 speeches without tone. Such a speech has usually five tones (tone 1, 2, 3, 4 and soft tone). Totally there are 1408 basic speeches with tone. Hakka language has 6 tones for so-called four-county accent and 7 tones for sea-land accent. The pitch contours different tones for Chinese and Hakka are shown in Fig. 1.

Two kinds of stops are also displayed in the figure for Hakka speeches: higher and lower stop. Table 1 presents the pitch contour, tone shapes and tonal relationship between Chinese and Hakka with 4 county accents. The tone shapes referred to the pinyin standard of MOE [4]. All the stops tones in Hakka language are the speeches which pinyin transliteration ends always with one of three character b, d and g, such as lib(立), ded(得) and bag(伯) in Hakka language.

![Figure. 1 Pitch contour of tone for Chinese and Hakka.](image)

<table>
<thead>
<tr>
<th>tone types</th>
<th>lower stops</th>
<th>fair stops</th>
<th>high stops</th>
<th>higher stops</th>
<th>lower stop</th>
<th>higher stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2, 4, 6</td>
<td>0, 4, 6</td>
</tr>
<tr>
<td>Hakka</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Principal Modules of TTS

Our text-to-Speech (TTS) system for Hakka’s language are composed of following modules:
A) Text Analysis
B) Prosody prediction
C) Selection of speech unit
D) Speech generation

The main modules of TTS can be expressed in Fig. 2. Three databases are generated in TTS: Language models for polyphones in Hakka’s language, Hakka’s lexicons and Hakka’s Speech database. All the speech units are recorded by a people with nature Hakka tone and frequent talking ability. More than 2427 monosyllabic speech units and 2234 word speech units of Hakka and several silences with
various durations have been recorded as basic units for speech synthesis.

The words or characters with more one phonetic pronunciation are usually polyphone, which are frequently found in every language. How to decide the collect phone for polyphone issue can be regarded as category ambiguity. Disambiguating the ambiguity can alleviate the problems in NLP. The paper address the dictionary matching, statistical N-gram language model (LMs) and voting scheme, which includes two methods: preference and winner-take-all scoring, to retrieve Chinese lexical knowledge, employed to process WSD on Chinese polyphonic characters.

Basic, in the training phase the n-gram language models were generated first by collecting large text data from website to disambiguate the sense ambiguity of Chinese polyphones. The features of speech prosody, such as break (stop), intonation, and energy, will be extracted by real speech data. All the information and features are passed into other module to generate the Hakka’s speech with frequent prosody.

There are more than 36000 Words lexicon in Mandarin/Hakka dictionaries. The number for various word lengths is depicted in Table 2. The schema contains many fields, such as Mandarin word, Hakka 4-county word, its pinyin, examples for Mandarin & Hakka, etc.

<table>
<thead>
<tr>
<th>Word length</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-syllable word</td>
<td>1239</td>
</tr>
<tr>
<td>2-syllable word</td>
<td>18940</td>
</tr>
<tr>
<td>3-syllable word</td>
<td>6715</td>
</tr>
<tr>
<td>4-syllable word</td>
<td>4304</td>
</tr>
<tr>
<td>5-syllable word</td>
<td>299</td>
</tr>
<tr>
<td>6-syllable word</td>
<td>88</td>
</tr>
<tr>
<td>7-syllable word</td>
<td>73</td>
</tr>
<tr>
<td>8-syllable word</td>
<td>15</td>
</tr>
</tbody>
</table>

Figure 2. TTS main modules for Hakkas language.
3. Break Prediction for Hakka

The intonation within sentence will be varied with respect word lexicon and semantic cues. The TTS module should predict prosody parameters, such as speech duration, break, pitch, energy and tone. In addition, the synthesized speech must have intelligibility and clarity so that the listeners understand the meaning of Hakka’s speech. In the paper, the concatenation of basic synthesis unit is adopted without the pitch adjustment of speech. There are 16 punctuation marks in sentence, and an appropriate break will be assigned to the each mask during the speech synthesis module.

As described above, it is appropriate that the break is an important feature of prosody [11]. According to the people’s convention, applicable break with various duration will be inserted into the sentence while they are pronounces the speeches. It will always upgrade the intelligibility and clarity synthesized speeches. In the paper, breaks are categorized into 4 types, as shown in Table 3. As shown in the table, the field time means that the silence speech with that time duration will be inserted and combined into the synthesized speech on our TTS system.

The approaches used for prosody prediction can be found in several works; such as statistical methods [5], neural networks [1, 6] and data mining [7, 8]. In the paper, we adopted the data mining approach. Some related works can be referred to [13, 14].

<table>
<thead>
<tr>
<th>Situation</th>
<th>Break type</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>between five punctuation “.,! ? , “</td>
<td>BIG</td>
<td>0.75</td>
</tr>
<tr>
<td>between prosodic phrase (five punctuation)</td>
<td>Major</td>
<td>0.60</td>
</tr>
<tr>
<td>between prosodic word (in prosodic phrase)</td>
<td>Minor</td>
<td>0.30</td>
</tr>
<tr>
<td>between word (in prosodic word)</td>
<td>no</td>
<td>0.0</td>
</tr>
</tbody>
</table>

3.1 Data Mining

Data mining is a useful technology with great potential to extract important information or features in larger database. Data mining software allows users to solve business decision problems. Data mining can be applied into wide variety of domains, such as text & Web mining, class categories, financial mining, knowledge discovery, etc. Currently, the most popular techniques in data mining are as following:

A) Artificial neural networks  
B) Decision trees  
C) Genetic algorithms  
D) Nearest neighbor method  
E) Rule induction

Data mining can be regarded as he Knowledge Discovery in Databases (KDD) in which process is commonly defined with the following phases:

A) Selection  
B) Pre-processing  
C) Transformation  
D) Data Mining  
E) Interpretation/Evaluation.

3.2 Methods to predict Type of Beak

3.2.1 Bayesian Networks

Predicting the types of break is regarded as classification issue. [7] adopted the decision tree to predict the proper categories. In the paper, we modified several detail and two methods, Bayesian
networks [8] and Classification And Regression Trees (CART)[9], are included into the prediction to promote the better performance.

Supposed that $C$ denotes the category, there are $m$ categories: $C_1, C_2, \ldots, C_m$, $m \geq 1$. Vector $X$ denotes data set attributes with $n$ dimension: $X = \{x_1, x_2, \ldots, x_n\}$. In the paper, it is apparent that $C$ and $X$ will be the types of break and features used to predict the categories for prosody. Basically, bayesian networks can be used to predict which category $C_i$ while data set $X$ belongs to:

$$ P(C_i|X) > P(C_j|X) \quad 1 \leq j \leq m, i \neq j $$

where $P(C_i|X)$ is max. probability for $C_i$ with respect to $X$ among $m$ categories.

Furthermore, $P(C_i|X)$ will be derived bases on the bayisian theorem:

$$ P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} $$

A Bayesian network (BN) over a set of variables $X$ is a Bayesian network structure (BNS) which is organized as a directed acyclic graph (DAG) over $X$, and a set of probability tables of BN:

$$ BNP = \{P(x_i|\text{pa}(x_i))\} $$

where $\text{par}(x_i)$ is the set of parents of $x_i$ in BNS.

A BNS represents probability distributions:

$$ P(X) = \prod_{i \in X} p(x_i|\text{par}(x_i)) $$

To employ a Bayesian network as a classifier, one simply calculates argmax$P(C|X)$ using the distribution $P(X)$ represented by the Bayesian network. Since all variables in $X$ are known, we do use Eq. (4) for calculation of all class values of variables.

As shown in Fig. 3, this picture just shows the network structure of the BN, but for each of the nodes a probability distribution for the node given its parents are specified as well. For example, in the Bayesian network above there is a conditional distribution for state $S_2$ and $S_3$ by given value of state $S_1$.

![Figure 3 Network structure of BN with 3 states.](image)

### 3.2.2 Classification and Regression Trees

Technically, Classification and Regression Trees (CART) is based on landmark mathematical theory introduced in 1984 by four world-renowned statisticians at Stanford University and the University of California at Berkeley [15]. CART is one of tree learning methods, in which a decision tree will be used as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such binary tree models are classification trees or regression trees.

The decision tree of CART, like ID3, C4.5, will be generated by greedy algorithm. The top node of tree is generated first and expanded by top-down and recursive methods. The complexity of CART can be written as $O(n^d|D|^\log|D|)$, where $D$ and $n$ denote data set and data number in $D$ respectively.

During the generation of decision tree, the feature selection method (FSM) can be employed to decide the best feature for splitting the data set. Two of most popular indexs in FSM are information gain and gain ratio. $C_i$ denotes the $i^{th}$ category for predition: $C_1, C_2, \ldots, C_m$ (same definition as above). Some other variables are defined as following: $C_{i,D}$ is the sub dataset of $D$ which are in category $C_i$ and
$|C_{i,D}|$ is the number of $C_{i,D}$. The expected information (EI) for a data set can be defined as followed:

$$EI(D) = -\sum_{i=1}^{m} p_i \log p_i,$$  \hspace{1cm} (5)

where $p_i$ denotes the probability for a data set in category $C_i$.

In Equation (5), $EI(D)$ is expressed by BIT for data set $D$. Probability $p_i$ is calculated as followed:

$$p_i = \frac{|C_{i,D}|}{|D|} \hspace{1cm} (6)$$

The Category $C$ can spitted into the several sub category $C_i$ for data set $D$ based on the attribute $A$. Supposed that $D$ will be spitted into $D_1$, $D_2$, $D_3$, ……, $D_v$, and $v$ denote the number of attribute $A$. After the splitting process, we will the following information to justify the splitted tree:

$$EI_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} EI(D_j) \hspace{1cm} (7)$$

where $EI_A(D)$ is the expected information based on the attribute $A$ for data set $D$.

The gain difference $GN$ for an attribute $A$ can furthermore calculated as followed:

$$GN(A) = EI(D) - EI_A(D) \hspace{1cm} (8)$$

Based on the Equation (8), it is apparent that the attribute which causes maximum gain benefit will be selected as dismembered attribute for the decision tree.

### 4. Evaluation and Analysis

#### 4.1. Data Sets of Experiments

The text examples of Hakka are collected from MOE and Hakka committee, Taiwan; which are pre-processed. There are 1200 totally well-defined sentences with 10196 various breaks and all the texts are divided into two data sets, 80% and 20% for training and testing phases, respectively.

In our experiments, all the sentences will be inserted one or several breaks into proper location between words by Hakka’s expert with fluently speaking and comprehensive ability.

#### 4.2 Precision of Prediction

In the paper, there are three types of break: major break, minor break and no-break between words. Several important features of Hakka’s lexicon and semantics in sentences are combined to evaluate and promote the predicting performance. The features with respect to the breaks for target word include:

- Part-of-Speech of target and neighbor’s word
- Length of target and neighbor’s word
- Location of words with respect to the target word
- Number of words preceding target word in sentences.
- Number of words after target word in sentences.
- Combination of features above.

#### Table 4 Experiments Features employed for break’s prediction

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. pre1next2,Length,POS</td>
</tr>
<tr>
<td>2. pre2next2,Length,POS</td>
</tr>
<tr>
<td>3. pre2next2,Length,POS,preType</td>
</tr>
<tr>
<td>4. pre2next2,Length,POS,PW</td>
</tr>
</tbody>
</table>
All the experiments are shown in Table 4, the best precision rates (PRs) for testing achieves 79.88% based on the CART and yet the lower PRs still achieves 77.69% by BN. The best PR for BN is 77.04%. In the Table 4, the features of “pre1next2, Length, POS” denote the 1 word preceding and next two words with respect to target, length of target in Chinese characters, part-of-speech of target are employed in the experiment. Figure 4 shows the precision rates for various features on two methods. The selected features for best performance include 2 words preceding and 2 words after target word, target word itself, target word’s length, Part-of-speech and previous word.

Table 5 present the confusion matrix for three types of break, which are expressed as no, minor and major for feature set 4 (pre2next2,Length,POS,PW), CART, in Table 4. The number in the table is the prediction for these breaks. The diagonal elements in the matrix are the collect predictions for each type of break. Table 6 present the confusion matrix for three types of break which are expressed as no, minor and major for feature set 4, BN.

The no-break can be predicted best among the three breaks, while the prediction performance for major break is relatively lower. Therefore, we should take more considerations on the improvement.

**Table 5** confusion matrix of breaks’ prediction for feature set 4, CART.

<table>
<thead>
<tr>
<th>Break</th>
<th>Prediction</th>
<th>No</th>
<th>minor</th>
<th>major</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>1229</td>
<td>196</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>minor</td>
<td>123</td>
<td>458</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>major</td>
<td>7</td>
<td>36</td>
<td>34</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6** confusion matrix of breaks’ prediction for feature set 4, BN.

<table>
<thead>
<tr>
<th>Break</th>
<th>Prediction</th>
<th>No</th>
<th>minor</th>
<th>major</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>1183</td>
<td>227</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>minor</td>
<td>131</td>
<td>445</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>major</td>
<td>8</td>
<td>39</td>
<td>33</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusions

This paper focuses on the break prediction of prosody for Hakka’s TTS system based on the data mining approaches, and implement the TTS system on Internet. Our system is composed of the following four components: Text analysis, Prosody prediction, Prosody prediction, and Speech generation module.
We focus on the issue of inserting breaks to speeches with emphasis on predicting the types of break. In the paper, there are three kinds of breaks: major break, minor break and no-break between words. We train a break prediction model based on the data mining approaches. The best precision rate for testing set achieves 79.88% based on the CART method and the performance of CART is superior to the BN. The selected features for best performance include 2 words preceding and 2 words after target word, target word, word’s length, Part-of-speech and Previous word.

In the future, our research works will be as followings:
1) Expansion for Hakka’s lexicon.
2) Improvement for prosody generation of speeches.
3) Disambiguity for Hakka’s polyphones.

6. Acknowledgments

We appreciate that the paper is partially supported by Council for Hakka Affairs Committee and National Science of Council, Executive Yuan, Taiwan.

7. References

[3] TTS system, CSIE NUU, Taiwan:http://203.64.183.226/public2/word1
[10] [Classification and Regression Trees (C&RT) http://www.statsoft.com/textbook/classification-and-regression-trees/]
[16] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques 2nd, Published by Elsevier.