A Corpus-Based Prosodic Modeling Method for Mandarin and Min-Nan Text-to-Speech Conversions

Sin-Horng Chen
Department of Communication Engineering
Chiao Tung University, Hsinchu
schen@cc.nctu.edu.tw

Abstract

This talk gives an introduction to a recurrent neural network (RNN) based prosody synthesis method for both Mandarin and Min-Nan text-to-speech (TTS) conversions. The method uses a four-layer RNN to model the dependency of output prosodic information and input linguistic information. Main advantages of the method are the capability of learning many human’s prosody pronunciation rules automatically and the relatively short time of system development. Two variations of the baseline RNN prosody synthesis method are also discussed. One uses an additional fuzzy-neural network to infer some fuzzy rules of affections from high-level linguistic features for assisting in the RNN prosody generation. The other uses additional statistical models of prosodic parameter to remove some affecting factors of linguistic features for reducing the load of the RNN.

1. Introduction

A text-to-speech (TTS) system is to convert a text into a clear and natural speech. It first analyzes input text to find some linguistic information. It then generates proper prosodic information from the linguistic information. Meanwhile it finds corresponding primitive waveform template string from an acoustic inventory. Lastly, it adjusts the prosody of the waveform template string to match with the prosodic information and produces the output synthetic speech. In a TTS system, prosodic information generation is the most important part to determine the naturalness of the synthetic speech. Many methods of prosodic information generation have been proposed in the past. They can be categorized into two main approaches: rule based and corpus based. A rule-based approach first assign basic values of prosodic parameters and then applies some rules to adjust these values according to the linguistic information extracted from the input text. Disadvantages of the approach include the difficulty to infer enough rules and long developing time. So a high-performance rule-based TTS system can not be obtained without long-term devotion. A corpus-based approach is based on the idea of learning from observations of a large speech corpus and making prediction about new observations. It first employs a model containing free parameters to describe the dependency of output prosodic information and input linguistic information. The model can be a statistic model, such as linear or nonlinear regressive model, or a neural network, like multi-layer perceptrons (MLP) or recurrent neural network (RNN). It then learns (or estimate) free parameters from training utterances. The parameter learning can be an optimization process using a criterion such as maximum likelihood (ML) or minimum mean-squared error (MMSE). After being properly trained, the model can be used to synthesis (or predict) prosodic information from given linguistic information.

In this talk a corpus-based prosodic modeling method for Mandarin and Min-Nan text-to-speech (TTS) conversions is introduced [7]. It takes the NN-based approach to model the dependency of output prosodic information and input linguistic information. The model is realized using a four-layer RNN with input linguistic information.
containing both word- and syllable-level features
and output prosodic information containing syllable
pitch contour, log-energy level and duration
parameters. The model has been proved to be
effective on exploring the relationship between the
prosodic phrase structure of Mandarin/Min-Nan
speech and the linguistic features of the
responding text. Two variations of the RNN-
based implementation have also been studied for
additional fuzzy-neural network to infer some fuzzy
rules in the pre-processing stage for providing high-
level linguistic cues. The other [10] uses additional
statistical models to normalize/denormalize output
prosodic parameters for providing consistent output
targets as well as for reducing the load of the RNN.

The presentation is organized as follows. Section 2 describes the background of Mandarin and
Min-Nan languages. Section 3 presents the baseline
method of RNN-based prosody synthesis for both
Mandarin and Min-Nan TTS. Sections 4 and 5
discuss, respectively, the two variations of the RNN
prosody synthesis method. Some concluding
remains are given in the last section.

2. Background of Mandarin and Min-
Nan Languages

A. Mandarin

Mandarin is the main spoken language of
Chinese. It is a tonal and syllabic language. There
exist more than 80,000 words, each composed of
one to several characters. There are more than
10,000 commonly-used characters, each pronounced
as a mono-syllable. The total number of
phonologically-allowed mono-syllables is only 1345.
All mono-syllables have a very regular, hierarchical
phonetic structure as shown in Table 1. A mono-
syllable is composed of a base-syllable and a tone.
There are in total 411 base-syllables. A base-syllable
can be further decomposed into two parts: an
optional initial and a final. The initial part contains a
single consonant if it exists. The final part consists
of an optional medial, a vowel nucleus, and an
optional nasal ending (or coda). These 411 base-
syllables are formed by all legal combinations of 21
initials and 39 finals. These 39 finals are, in turn,
formed by the combinations of 3 medials, 9 vowel
nuclei, and 5 nasal endings. There are only five
lexical tones, namely, Tone 1 (or high-level tone),
Tone 2 (or high-rising tone), Tone 3 (or low-dipping
tone), Tone 4 (or high-falling tone), and Tone 5 (or
neutral tone). The first four tones have standard
pitch contour patterns, while the last one is usually
pronounced more arbitrarily.

Many high-performance Mandarin TTS
systems have been developed in the past few years
[1-3]. Among them, [1] used the rule-based
approach to synthesize prosodic information, while
[2,3] adopted the corpus-based one. [2] is our
system. The system consists of four main functional
blocks: text analyzer, RNN prosody synthesizer,
acoustic inventory, and PSOLA speech synthesizer.
It tokenizes the input text into a word sequence in
text analyzer, takes all 411 base-syllables as the
basic synthesis units and stored in acoustic
inventory, adopts the RNN-based approach to
generate prosodic information (to be discussed in
Section 3), and uses the PSOLA synthesis method to
generate the output synthetic speech.

B. Min-Nan speech

Min-Nan speech is a spoken dialect widely
used in Fu-Jan and Taiwan. Just like Mandarin, Min-
Nan speech is also monosyllabic and tonal. Each
character is pronounced as a syllable carrying a
lexical tone. There are only 877 base-syllables and 8
tones including a degenerated one which is not used
by modern Taiwanese. These 877 base-syllables also
have the same initial-final structure like Mandarin
base-syllables. There are 18 initials and 82 finals. Although Min-Nan speech has similar linguistic characteristics as Mandarin speech, it is colloquial and does not have a standard written form. There exist two popular written forms in Taiwan. One is the Romanization form which uses Roman alphabets to spell each base-syllable and uses a number to specify its tone. The other is a hybrid one which uses Chinese characters to represent ordinary words and represents some extraordinary syllables in the Romanization form. Unfortunately, the system to represent words in Chinese characters is still not standardized. This makes the text analysis very difficult for Min-Nan language.

In the past, there are very few studies in Min-Nan TTS. Only a few preliminary studies were found in Taiwan in recent years [4-6,10]. [6,10] is our system. It is designed and implemented using an approach similar to our Mandarin TTS system development.

3. The Baseline RNN Prosody Synthesis Method

Prosodic information to be synthesized includes eight parameters representing syllable-initial duration, syllable-final duration, syllable log-energy level, syllable pitch mean and shape, and inter-syllable pause duration. Fig. 1 shows a block diagram of the method. It uses an RNN to automatically learn the relation between input linguistic features and output prosodic parameters. The RNN is a four-layer network with one input layer, two hidden layers, and one output layer. It can be functionally decomposed into two parts. The first part consists of a portion of the input layer and the first hidden layer with all outputs being fed back as inputs of itself. It operates in a word-synchronous mode using word-level input linguistic features such as POSs, word lengths, and punctuation marks extracted from the context of the current word. The second part of the RNN consists of the other part of input layer, the second hidden layer and the output layer. It operates in a syllable-synchronous mode using syllable-level input linguistic features such as tones, initial types, final types, and syllable’s location extracted from the context of the current syllable. The RNN has been proven to be effective on learning the mapping from input linguistic features to output prosodic parameters. Important contextual information of the input linguistic features are automatically detected and well represented in the hidden layer outputs which in turn are used to determine the output prosodic parameters. The RNN can be trained using a large single-speaker speech data set by the back-propagation through time algorithm [12].

The method has been applied to both of our Mandarin and Min-Nan TTS studies. For Mandarin TTS, a large single, male speaker database was used to test the method. The database contained 452 sentential utterances and 200 paragraphic utterances. Texts of these 452 sentential utterances were well-designed, phonetically balanced short sentences with lengths less than 18 characters. Texts of these 200 paragraphic utterances were news selected from a large news corpus to cover a variety of subjects including business (12.5%), medicine (12.0%), social event (12.0%), sports (10.5%), literature (9.0%), computers (8.0%), food and nutrition (8.0%), etc. All utterances were generated by a male speaker. They were all spoken naturally at a speed of 3.5-4.5 syllables per second. They were digitally recorded in a 20 kHz rate. All speech signals and the associated texts were manually pre-processed in order to extract the acoustic features and the linguistic features required to train and test the RNN. The database was divided into two parts. The one containing 491 utterances (or 28060 syllables) was used for training and the other containing 161 utterances (or 7034 syllables) was used for testing.
Table 2 shows the experimental results.

For Min-Nan TTS, another single, male speaker database was used to test the method. The database contained 255 utterances including 130 sentential utterances with length in the range of 5-30 syllables and 125 paragraphic utterances with length in the range of 85-320 syllables. The total number of syllables is 23,633. Pre-processing of the database were also done manually. Table 3 shows the experimental results. By comparing Tables 2 and 3 we find that the baseline method performed equally well for Mandarin and Min-Nan TTSs.

4. The SONFIN-RNN Prosody Synthesis Method

In the baseline RNN prosody synthesis method, the prosody pronunciation rules are completely explored by the RNN and implicitly stored in its weights. The RNN structure is well-designed to separately consider the affections from word-level and syllable-level linguistic features. A variation of the baseline method to consider the affection from word-level linguistic features by an additional fuzzy neural network was proposed in [13]. Fig. 2 displays its block diagram. It is composed of two parts: a self-constructing neural fuzzy inference network (SONFIN) [14] and an RNN. The function of the SONFIN is to infer some fuzzy rules of human’s prosody phonology from input linguistic features. It generates four outputs to be used in the RNN for help synthesizing four sets of prosodic parameters. Each output is a linear combination of affections of several fuzzy rules inferred automatically. Input features used include the location of the current syllable in a phrase (WordInPhrase), the location of the current phrase in a sentence (PhraseInSentence), the length of the current sentence, part-of-speech (POS), and punctuation marks. The function of the RNN is to generate four sets of prosodic parameters by using syllable-level linguistic features and the outputs of the SONFIN. The four sets of prosodic parameters include: (1) four parameters representing the pitch contour [11], (2) initial and final durations, (3) pause duration, and (4) log-energy level. Table 4 shows the experimental results of the SONFIN-RNN method. It can be found from Table 4 that the SONFIN-RNN method is comparable to the baseline method. By analyzing the responses of the SONFIN, some prosody phonologic rules such as, location of accent in a polysyllabic word, can be observed.

5. The Hybrid Statistic/RNN Prosody Synthesis Method

In the baseline method, the dependency of the output synthesized prosodic parameters on the input linguistic features is completely taken care by the RNN. The load is heavy because linguistic features of different levels can interactively affect the pronunciation of these prosodic parameters. The RNN may therefore neglect some secondary affecting factors. Another drawback of the baseline method is that the speaking rate variability resided in the training database is totally ignored. To cure these two drawbacks, a statistical/RNN prosodic information synthesis method was proposed [8,9]. Its block diagram is displayed in Fig. 3. It differs from the baseline method on incorporating an additional statistical model-based normalization/denormalization module. The function of the statistical-based normalization/denormalization module is to normalize the prosodic parameters in the training phase and to denormalize the RNN outputs in the synthesis phase. Its goal is to eliminate some affecting factors for providing more consistent and compact training targets to efficiently train the RNN. Statistical models of the eight output prosodic parameters are used in the normalization and denormalization processes. We divide these eight
parameters into six groups to be modeled separately. In the following, we discuss them in detail.

A. The syllable initial duration model [8]

Syllable initial duration is modeled as normal distribution but considering three major affecting factors resided in the observed data of real speech. They include utterance-level speaking speed, lexical tone, and prosodic state. The model is expressed by

\[ Z_{i}^i \cdot \beta_{n}^{i} \cdot \beta_{p}^{i} \cdot \beta_{l}^{i} = X_{n}^{i}. \]

where \( Z_{n}^{i} \) is the observed initial duration of the syllable \( S_{n} = (j_{n}, t_{n}, p_{n}, l_{n}) \) with base-syllable \( j_{n} \), lexical tone \( t_{n} \), prosodic state \( p_{n} \), and in utterance \( l_{n} \); \( \beta_{n}^{i}, \beta_{p}^{i}, \text{and} \beta_{l}^{i} \) are the companding (compressing-expanding) factors of initial duration due to the three affecting factors of \( t_{n}, p_{n}, \text{and} l_{n} \), respectively; and \( X_{n}^{i} \) is the normalized initial duration to be modeled as a normal distribution with mean \( \mu_{j_{n}} \) and variance \( \nu_{j_{n}} \). The prosodic state \( p_{n} \) is conceptually defined as the state of a syllable in a prosodic phrase and is labeled automatically by a vector quantization classifier using seven acoustic features extracted from the vicinity of the current syllable. They include normalized final duration, pitch mean, pitch differences with the two nearest neighboring syllables, log-energy mean, and log-energy differences with the two nearest neighboring syllables. Two special states were assigned, respectively, to the beginning and ending syllables of an utterance. The total number of prosodic states was empirically set as 14. An iterative procedure derived based on the ML (maximum likelihood) criterion is employed to sequentially estimate all model parameters and these three companding factors from observed data.

B. The syllable final duration model [8]

Syllable final duration is modeled completely in the same way as syllable initial duration. So we omit the discussions here.

C. The inter-syllable pause duration model

Inter-syllable pause duration is modeled as a normal distribution conditioned on two broad-classes of entering-tone and the prosodic state of the preceding syllable and four broad-classes of the initial of the succeeding syllable. An entering-tone represents an optional unreleased stop of /p/, /h/, /k/, or /t/ resided at the end of a syllable. Due to its special pronunciation style, it will affect the length of the pause following the syllable. Prosodic state is used to represent the structure of prosodic phrase. It surely will affect the length of the pause following the syllable. A well-known example is that a relatively long pause always exists at the end of a prosodic phrase. Initial of the next syllable is also considered because there always exists a pause if the next syllable has a stop initial. On the contrary, it is easy to happen that no pauses exist when the next syllable has a liquid, nasal, or null initial.

D. The pitch level model

To model the pitch contour of a syllable, we first represent it using four orthogonal transform coefficients with one representing the level and the other three representing the shape [11]. We then model the coefficient representing pitch level using a normal distribution conditioned on the tone \( t_{n} \) and the prosodic state \( p_{n} \) of the current syllable. It is well-known that pitch level is seriously affected by the prosodic phase structure. Declination effect for declarative sentential utterance is a typical example. The pitch level of a syllable is also affected by its tone type. Tone 3 and tone 5 usually have lower pitch (frequency) level than the other 5
E. The pitch shape model

Each of the other three orthogonally-transformed coefficients representing the shape of syllable pitch contour is modeled as a normal distribution conditioned on the tones of the preceding, current and succeeding syllables. The pitch shape of a syllable is mainly determined by its tone. It is also affected by the tones of the two nearest neighboring syllables. Tone coarticulation is very common in both continuous Mandarin and Min-Nan speech.

F. The log-energy level model

Log-energy level of syllable is simply normalized to the maximum log-energy level of syllable in the current utterance. This is to eliminate the effect of volume variation.

Using these statistical models, all prosodic parameters are normalized and taken as output targets of the RNN in the training phase. For initial duration, final duration and four pitch parameters, simple normalization operations to firstly subtract out their mean values and then be divided by their standard deviations are performed. For pause duration, the same normalization operation is performed when its value is smaller than 1500 sampling points (or 75 ms). When pause duration is greater than 1500 sampling points, we quantize its value into three levels representing long, medium long, and very long pauses. For log-energy level, no further normalization is performed. All these normalized prosodic parameters are taken as output targets and used to train the RNN. It is noted that Schemes 3 is impractical and taken as an optimistic reference.

Due to the fact that the exact prosodic state of a syllable is not known explicitly, three schemes of the method were tested. Scheme 1 used degenerated versions of these statistical models by suppressing the effect of prosodic state. Information of prosodic state is therefore not used in both the training and testing. Scheme 2 used estimated prosodic states obtained by an RNN classifier using linguistic features extracted from the input text as inputs and the 12 prosodic states as outputs. Scheme 3 used real prosodic states obtained by the same VQ classifier used in the training phrase. It is noted that Schemes 3 is impractical and taken as an optimistic reference.

Experimental results of the method for Min-Nan TTS is shown in Table 3. It can be found from Table 3 that Scheme 1 is slightly better than the baseline method. Scheme 2 is not good and Scheme 3 is the best. It can also be found from Table 3 that the statistical models of final duration and pause duration are better than those of all other 6 parameters.

6. Concluding Remarks

The corpus-based prosody synthesis method using RNN to model the dependency of output prosodic information and input linguistic information has been proved to be effective and efficient for both Mandarin and Min-Nan TTS. Preliminary studies to further improve the method via incorporating an additional fuzzy-neural inference network to explore fuzzy rules of affections from high-level linguistic features and via combining statistical models of prosodic parameters to remove some affecting factors of linguistic features have also been proved to be feasible. Challenges in future studies along this research direction include:

1. Sophisticatedly incorporate syntactic and semantic features for developing a near perfect
TTS system.
2. Properly model prosody of spontaneous speech for developing a TTS system capable of speaking in a variety of styles.

References
Table 1: The phonetic structure of Mandarin syllables

<table>
<thead>
<tr>
<th>base-syllable</th>
<th>(initial)</th>
<th>(final)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>411</td>
<td>21</td>
</tr>
<tr>
<td>(initial)</td>
<td>21</td>
<td>39</td>
</tr>
<tr>
<td>(consonant)</td>
<td>21</td>
<td>(medial)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vowel nucleus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(nasal ending)</td>
</tr>
</tbody>
</table>

Table 2  Experimental results of the baseline method for Mandarin TTS

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>Inside</th>
<th>outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch contour</td>
<td></td>
<td>0.84 ms/frame</td>
<td>1.06 ms/frame</td>
</tr>
<tr>
<td>Initial duration</td>
<td></td>
<td>17.2 ms</td>
<td>18.5 ms</td>
</tr>
<tr>
<td>Final duration</td>
<td></td>
<td>33.3 ms</td>
<td>36.7 ms</td>
</tr>
<tr>
<td>Pause duration</td>
<td></td>
<td>23.7 ms</td>
<td>54.5 ms</td>
</tr>
<tr>
<td>Energy level</td>
<td></td>
<td>3.39 dB</td>
<td>4.17 dB</td>
</tr>
</tbody>
</table>

Table 3  Experimental results of prosodic information synthesis for Min-Nan TTS

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>Baseline Method</th>
<th>Statistic/RNN Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Scheme 1</td>
<td>Scheme 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inside</td>
<td>outside</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inside</td>
<td>outside</td>
</tr>
<tr>
<td>Pitch period (ms)</td>
<td>1.01</td>
<td>1.02 1.07</td>
<td>1.02 1.06</td>
</tr>
<tr>
<td>Initial duration (ms)</td>
<td>11.5</td>
<td>11.2 12.2</td>
<td>11.2 11.4</td>
</tr>
<tr>
<td>Final duration (ms)</td>
<td>33.9</td>
<td>30.9 36.0</td>
<td>30.9 35.9</td>
</tr>
<tr>
<td>Pause duration (ms)</td>
<td>25.5</td>
<td>14.5 27.3</td>
<td>14.5 15.6</td>
</tr>
<tr>
<td>Energy (dB)</td>
<td>2.32</td>
<td>2.37 3.14</td>
<td>2.37 3.14</td>
</tr>
</tbody>
</table>
Table 4 Experimental results of the SONFIN-RNN method for Mandarin TTS

<table>
<thead>
<tr>
<th></th>
<th>Inside</th>
<th>outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE Pitch contour</td>
<td>0.86 ms/frame</td>
<td>1.06 ms/frame</td>
</tr>
<tr>
<td>RMSE Initial duration</td>
<td>19.8 ms</td>
<td>20.3 ms</td>
</tr>
<tr>
<td>RMSE Final duration</td>
<td>34.4 ms</td>
<td>36.3 ms</td>
</tr>
<tr>
<td>RMSE Pause duration</td>
<td>42.2 ms</td>
<td>44.8 ms</td>
</tr>
<tr>
<td>RMSE Energy level</td>
<td>3.96 dB</td>
<td>4.09 dB</td>
</tr>
</tbody>
</table>

Fig. 1 A block diagram of the baseline RNN prosody synthesis method.
Fig. 2 A block diagram of the SONFIN-RNN prosody synthesis method.

Fig. 3 A block diagram of the hybrid statistical/RNN prosody synthesis method.