1. INTRODUCTION

Mandarin is a syllable-rhythmic language. There exist around 1,300 syllables. They form a regular CV structure (also known as initial plus final) except for a few with a void consonant (or zero initial). As a tone language, each syllable carries one of five types of lexical tone, denoted by N(eutral), H(igh), R(ise), L(ow), F(all), respectively, which are based on their pitch movement. A few more words focus on the Mandarin prosody. Within the complex of prosodic features, the smallest distinctive configuration is tone. Tone is primary in that larger configurations, including phrasal tune and sentence intonation, are specific modifications of a string of one or more tones; the larger configurations can be predicted given the string and modifications, whereas they cannot be predicted from the modifications alone.

Development of text-to-speech (TTS) conversion systems has been mostly based on the concatenation of waveform units selected from a speech corpus, e.g., [1]. It is crucial to design such a speech corpus that self-contains all speech units and most of their variations relevant to speech naturalness [2]. There are contradictory requirements on the corpus size. A small corpus is desired from a large raw text corpus for synthesis of Mandarin speech, taking account of the varied spectral characteristics, including the within-syllable co-articulation and that between syllables, and the prosody, including tone, phrasal tune and sentence intonation. Based on the quantitative analysis results in previous work [3], the prosody in this paper is approximately predicted from lexical tones and their contexts with effects of the phrase construction and sentence type.

The remainder of this paper is organized as follows. Section 2 describes the synthesis units. Section 3 outlines the multi-tier algorithm. Section 4 presents the experimental result and its evaluation. A summary is given in Section 5.

2. DEFINING SYNTHESIS UNITS

2.1. The syllable as basic unit

The syllable is treated as the basic unit. Syllable-sized units allow one to perfectly model the within-syllable co-articulation, e.g., [1]. Yet, it is possible to use smaller or larger units available in a speech corpus for synthesis. Accordingly, modeling the phonetic and tone contexts of a basic unit is one of the key points for synthesizing natural sounds.

2.2. Phonetic and tone contexts of the syllable

Proper phonetic and tone contexts of a syllable can embed the syllable into specific spectral and prosodic environments in recording. The phonetic and tone contexts for the syllable can be independently modeled by considering the phoneme and tone of the left-neighboring syllable with those of the right-neighboring syllable. For convenience, let $P(n_l, n_r)$ and $T(n_l, n_r)$ denote a set of instances of phonetic and tone contexts, respectively, where $n_l$ stands for the number of corresponding categories of the left neighbor and $n_r$ for that of the categories of the right neighbor.

When focusing on the full phonetic contexts, $n_l$ takes 14 and $n_r$ takes 27. That is, there are $(14 \times 27) = 378$ possible instances in $P(14,27)$ for each syllable. Although not all of these instances occur in real text, some of them are still nuances in spectral characteristics. Based on the knowledge of acoustic phonetics, we further reduce both $n_l$ and $n_r$ to 7 to focus on major phonetic contexts, denoted by $P(7,7)$. Tables 1 and 2 list the final and initial classes, respectively.

Tone context focuses on two contextual tone phenomena: carry-over and anticipatory effects. There are $(5 \times 5) = 25$ full tone con-
texts. By using the findings obtained from quantitatively analyzing the P0 contours covering all kinds of tone combinations in polysyllabic words [3], either $n_1$ or $n_2$, i.e., the number of major tone categories of the left- and right-neighboring syllables, can take 2 or 3 to give the major tone contexts. Therefore, there exist 6 to 9 major tone contexts for the syllable depending on its tone type, simply denoted by $T(3,3)$. Table 3 tabulates these tone categories, each of which is enclosed in a pair of brackets.

### 2.3. Syllable position in the phrase construction

The prosodic effects at the phrase level, such as normal stress and phrasal tune, on the syllable are captured by embedding the syllable at the specific position in the phrase. The syllable position in the phrase is predicted from the phrase construction of the sentence. Each sentence is first automatically segmented into lexical words with POS (part-of-speech) tagging and then parsed for predicting its phrase construction. The parser employs probabilistic-context-free-grammar-based rules, which are directly extracted from the Penn Chinese Treebank [4]. The syllable position in the phrase takes the 11 values as follows.

<table>
<thead>
<tr>
<th>Category</th>
<th>Initial</th>
<th>Category</th>
<th>Initial</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>b, d, p, t</td>
<td>T1</td>
<td>z, c, s, f</td>
</tr>
<tr>
<td>I2</td>
<td>g, k, h</td>
<td>T6</td>
<td>zh, ch, sh, r</td>
</tr>
<tr>
<td>I3</td>
<td>j, q, x</td>
<td>T7</td>
<td>m, n, l</td>
</tr>
<tr>
<td>I4</td>
<td>zero initial</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Seven initial classes

where the term word indicates a chunk that a string of one or more lexical words are grouped together by POS-based dependency analysis. It is noted that the syllable-position prediction faces complicated problems, such as robust parsing. To suppress effect of potential prediction errors, the 11 types of syllable position (denoted by P11) are then classified into 6 classes (P6), namely, \{IP\}, \{FsP\}, \{FIP\}, \{IWF\}, \{FWF\}, \{MWmf\}, \{MWmb\}, \{MWb\}. Compared with P11, P6 overlooks the word boundaries and just scans the syllable positions in the phrase in a uniform way.

### 2.4. Separating statements from questions

The effect of sentence intonation on the syllable is treated by separating the statement category (abbreviated by “sta”) from the question category (“que”). This is based on the speech fact that statement intonation and question intonation produce quite different effects on the syllable [3]. In the present experiment, the syllables belonging to the clause ending with “?” are treated as the question category. The remainder are all treated as the statement category.

### 2.5. Multi-tier transcription of the syllable

The syllable is transcribed into unit candidates in a multi-tier way. Each tier of transcriptions focuses on individual environments as follows.

- **Tier 1**: the $n$ syllable-sized units (denoted by set $Sn|y$)
- **Tier 2**: unit $\subset Sn|y$ with major tone contexts ($SnT(3,3)|y$)
- **Tier 3**: unit $\subset Sn|y$ with major phonetic and tone contexts in major positions ($SnT(3,3)P6|y$ and $SnP(7,7)P6|y$)
- **Tier 4**: unit $\subset Sn|y$ under the full tone and full phonetic contexts ($SnT(5,5)|y$ and $SnP(14,27)|y$)
- **Tier 5**: unit $\subset Sn|y$ at all of the 11 positions ($SnP11|y$)

Table 3 tabulates the syllable-sized units (denoted by set $Sn|y$), $y \in \{\text{sta, que}\}$, and its relative frequency of occurrence in a given raw text corpus, respectively, where $i = 1, \ldots, N_X$ (the number of elements in $X$). By definition $\sum_{i=1}^{N_X} p(\mu^X_i|y) = 1$. The coverage of $X$ involved in the sentence set $\Omega$ is defined by $P(X) = \sum_{i=1}^{N_X} p(\mu^X_i|y) f(\mu^X_i|y)$. (1)

where function $f(\mu^X_i|y) = 1$ if $\mu^X_i|y \in \Omega$; otherwise 0.

With $P(X)$ we can then calculate a measure of coverage jointly incorporating all of the 5-tier unit candidates (henceforth, joint coverage) in the priority ranked in the tier order as the criterion for governing the sentence selection. A multi-tier algorithm, adapted from [2], is illustrated as follows for this purpose. It is aimed in selecting a sentence set $\Omega$ with given set size from a given raw text corpus, and making it possible to maximize the joint coverage of predicted unit candidates involved in the $\Omega$.

3. **ALGORITHM OUTLINE**

Let $\mu^X_i|y$ and $p(\mu^X_i|y)$ be the $i$th unit in set $X \in \{Sn|y, SnT(3,3)|y, SnT(3,3)P6|y, SnP(7,7)P6|y, SnT(5,5)|y, SnP(14,27)|y, SnP11|y\}$, $y \in \{\text{sta, que}\}$, and its relative frequency of occurrence in a given raw text corpus, respectively, where $i = 1, \ldots, N_X$ (the number of elements in $X$). By definition $\sum_{i=1}^{N_X} p(\mu^X_i|y) = 1$. The coverage of $X$ involved in the sentence set $\Omega$ is defined by $P(X) = \sum_{i=1}^{N_X} p(\mu^X_i|y) f(\mu^X_i|y)$.

$$\text{where } f(\mu^X_i|y) = 1 \text{ if } \mu^X_i|y \in \Omega; \text{ otherwise 0.}$$
Initialization

Initialize $\Omega$ by either taking an existing sentence set or randomly extracting sentences from the corpus.

Iteration

(1) Select the next sentence $t$ from the raw text corpus.
(2) Select a sentence $s$ from $\Omega$ temporarily, and evaluate the stepped joint coverage $C_i^s$, $i = 1, \ldots, 5$, before and after the exchange, as defined in eqs. (2) to (6), where $x \in \{a, b\}$.

$$C_i^s = w_x P(Sn|sta) + w_y P(Sn|que)$$

(2)

$$C_i^s = w_x P(Sn(T,3,3)|sta) + w_y P(Sn(T,3,3)|que)$$

(3)

$$C_i^s = w_x(\alpha P(Sn(T,3,3)|P6|sta) + (1 - \alpha) P(Sn(7,7)|P6|sta)) + w_y(\alpha P(Sn(T,3,3)|P6|que) + (1 - \alpha) P(Sn(7,7)|P6|que))$$

(4)

$$C_i^s = w_x(\alpha P(Sn(T,5,5)|sta) + (1 - \alpha) P(Sn(14,27)|sta)) + w_y(\alpha P(Sn(T,5,5)|que) + (1 - \alpha) P(Sn(14,27)|que))$$

(5)

where $\alpha$ is a coefficient for balancing between prosodic and phonetic contexts: $w_x$ and $w_y$ are weight coefficients with the constrained condition $w_x + w_y = 1$.

(4) The evaluation is carried out in the following order. If any one is satisfied, fix the exchange. Otherwise discard $t$.

(a) $C_1^s > C_0^a$
(b) $C_0^a$ but $C_2^s > C_0^d$
(c) $C_0^a$ but $C_2^s > C_0^d$
(d) $C_0^a$ but $C_2^s > C_0^d$
(e) $C_0^a$ but $C_2^s > C_0^d$
(f) $C_0^a$ but $C_2^s > C_0^d$

The number of characters in $\Omega$ decreases.

Termination

Terminate the algorithm if a pre-defined maximal count of exchange trials is reached.

4. EXPERIMENTAL EVALUATION

4.1. Selecting a raw text corpus

As described in the multi-tier algorithm, the coverage requirement of a speech corpus is determined statistically from the analysis of a representative text corpus, when given the task domains. We are interested in two general tasks: traveling dialog and news-reading.

An experiment was run on the phrasebook (PB), a dialog text corpus (around 146k sentences) collected at ATR, and three news text corpora, PD91, PD92 and PD93, collected from People’s Dailies from 1991 to 1993 (around 210k sentences for each in this experiment), respectively. The analysis focused on the distribution of syllables and their phonetic and tone contexts. Figure 1, for example, shows the syllable distribution for each corpus. A few distributions of symmetrized Kullback-Leibler divergence, defined in eq. (7), are also superimposed on this figure.

$$D(\text{corpus}, \text{corpus}) = - \sum_{i=1}^{N} \left( p_{i1} \ln \frac{p_{i1}}{p_{i2}} + p_{i2} \ln \frac{p_{i2}}{p_{i1}} \right) / 2.$$  

(7)

Generally, the basic statistics is rather similar among PD91, PD92 and PD93, but considerably different between the news corpus and PB. Based on this experimental result, we took PD93 (PD henceforth) and PB based on this experiment result, we took PD93 (PD henceforth) and PB to set up a raw text corpus which includes around 580k sentences. Table 4 lists the count of unit candidates involved in the text corpus, where #($X$) indicates the number of elements in $X$. In this table, PB is divided into two sub-corpora. One focuses on the hotel reservation dialog (hotel-dialog, or HD), and the other concerns traveling dialog in general (travel-dialog, or TD).

4.2. Extracting a sentence set

By use of the proposed algorithm, 9,479 sentences, including statements and questions, were selected out from the raw text corpus. Particularly, 1,680 sentences were extracted from HD, 2,966 from TD, and 4,833 from PD. The selection was relatively sub-corpus-dependent and carried out step by step: firstly the 1,680 sentences, secondly the 2,966 sentences, and lastly the 4,833 sentences. To maximize the coverage, the joint coverage for extracting a subset was calculated upon the ongoing $\Omega$ plus the subset extracted already, if any. In the process of extracting the 2,966 sentences, for instance, $P(X)$ was calculated upon the $\Omega$ plus the 1,680 sentences and the $P(\mu, \nu | y)$ for TD. In addition, weights $w_{x}$ and $w_{y}$ were fixed to make statements and questions tackled equally, while $\alpha$ took 0.5 according to a preliminary experiment. The algorithm scanned the raw text corpus 2 times.

4.3. Results and discussion

Experimental results indicated that the proposed algorithm could select out a sentence set with the maximum coverage of the unit candidates predicted from the text corpus. The joint coverage monotonically increases with the increase in set size. Figure 2 shows examples of $P(Sn(T,3,3)|P6|y)$ as a function of sentence set size. Also, for a given text corpus there exist a set of relevant sentences to yield the optimal coverage. Augmenting the optimal coverage requires many more sentences as the desired coverage increases. Refer to figure 2, for instance, 1,000 (or 2,000) more sentences to be extracted merely improve $P(Sn(T,3,3)|P6|TD(sta))$ by 0.11% (or 0.24%) from the optimal coverage 92.24%.

Table 4. Count of unit candidates in the raw corpus

<table>
<thead>
<tr>
<th>Category</th>
<th>Hotel-dialog (sta) (que)</th>
<th>Travel-dialog (sta) (que)</th>
<th>People’s-daily (sta) (que)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus size</td>
<td>2.23k 1.16k</td>
<td>95.2k 48.1k</td>
<td>440k 7.33k</td>
</tr>
<tr>
<td>#($S_1</td>
<td>y$)</td>
<td>668 496</td>
<td>1115 987</td>
</tr>
<tr>
<td>#($S_3(T,3,3)</td>
<td>P6</td>
<td>y$)</td>
<td>2500 1482</td>
</tr>
<tr>
<td>#($S_7(7,7)</td>
<td>P6</td>
<td>y$)</td>
<td>5774 2771</td>
</tr>
<tr>
<td>#($S_8(14,27)</td>
<td>P6</td>
<td>y$)</td>
<td>8966 3640</td>
</tr>
<tr>
<td>#($S_9(T,5,5)</td>
<td>y$)</td>
<td>4610 2325</td>
<td>22.3k 15.6k</td>
</tr>
<tr>
<td>#($S_{11}</td>
<td>y$)</td>
<td>7874 3264</td>
<td>84.7k 46.0k</td>
</tr>
<tr>
<td>#($S_{11}</td>
<td>y$)</td>
<td>3377 1924</td>
<td>9896 7874</td>
</tr>
</tbody>
</table>
judged by use of the six punctuation , ? ! : ; . There exist almost 1,000 syllables (the total number of syllables is 1,245) occur at the i-th syllable position, i = 1, ..., 12, in statements. In average, there are 3 to 6 instances of the unit with typical phonetic and tone contexts at each of the 12 positions as well. Taking into consideration of the bias syllable distribution, as shown in figure 1, we believe that the sentence set is good enough for producing the varied prosodic and specific characteristics for speech synthesis.

5. SUMMARY

This paper presented a multi-tier algorithm to select a sentence set from a large raw text corpus, taking account of the varied phonetic and prosodic characteristics. Compared with conventional greedy algorithms, the proposed method introduces a multi-tier mechanism in the coverage evaluation for restricting the number of degrees of freedom of the problem to facilitate data-driven selection. A 9,479 sentence set was selected out from a 580k-sentence corpus including dialog and news text. The analysis results indicated that the sentence set covered all of the syllables under such linguistic environments as typical phonemic and tone contexts, major syllable positions within the phrase construction and the sentence type. The linguistic environments in turn control the syllable with specific prosodic and spectral environments in recording. Further work is planned to investigate the relationship between size of the sentence set based speech corpus and quality of synthetic speech.

ACKNOWLEDGEMENT This research was supported in part by the Telecommunications Advancement Organization of Japan.

6. REFERENCES


