Corpus Design for Malay Corpus-based Speech Synthesis System

Tian-Swee Tan and Sh-Hussain
Faculty of Biomedical Engineering and Health Science, P11, Center for Biomedical Engineering, University Technology Malaysia, 81310 UTM Skudai, Johor DT, Malaysia

Abstract: Problem statement: Speech corpus is one of the major components in corpus-based synthesis. The quality and coverage in speech corpus will affect the quality of synthesis speech sound.

Approach: This study proposes a corpus design for Malay corpus-based speech synthesis system. This includes the study of design criteria in corpus-based speech synthesis, Malay corpus based database design and the concatenation engine in Malay corpus-based synthesis system. A set of 10 millions digital text corpuses for Malay language has been collected from Malay internet news. This text corpus had been analyzed using word frequency count to find out all high frequency words to be used for designing the sentences for speech corpus.

Results: Altogether 381 sentences for speech corpus had been designed using 70% of high frequency words from 10 million text corpus. It consists of 16826 phoneme units and the total storage size is 37.6Mb. All the phone units are phonetically transcribed to preserve the phonetic context of its origin that will be used for phonetic context unit. This speech corpus had been labeled at phoneme level and used for variable length continuous phoneme based concatenation.

Conclusion/Recommendation: This study has proposed a platform for designing speech corpus especially for Malay Text to Speech which can be further enhanced to support more coverage and higher naturalness of synthetic speech.

Key words: Text to speech, unit selection, concatenation, corpus-based speech synthesis, speech synthesis, variable length unit selection

INTRODUCTION

Nowadays, a new concept of speech synthesis: Corpus based speech synthesis (unit selection), has been introduced[1]. It has emerged as a promising methodology to solve the problems with the fixed-size unit inventory synthesis, e.g., diphone synthesis[2]. Corpus-based approaches to speech synthesis have been advocated to overcome the limitations of concatenative synthesis from a fixed acoustic unit inventory[1]. The frequency of unit concatenations in e.g., diphone synthesis has been argued to contribute to the perceived lack of naturalness of synthesis speech.

Definition and concept of corpus-based TTS: Corpus-based TTS system creates the output speech by selecting and concatenating units (e.g., Speech sounds or words) from a large (can be up to several hours long) speech database to select the units to be concatenated[3-5].

The idea of corpus-based or unit selection synthesis is that the corpus is searched for maximally long phonetic strings to match the sounds to be synthesized[6]. According to Nagy[1], as the length of the elements used in the synthesized speech increases, the number of concatenation points decreases, resulting in higher perceived quality. If the database offers sufficient prosodic and allophonic coverage, it is then even possible to generate natural sounding prosody without resort to signal manipulation.

Rutten[6] also states that with the arrival of corpus-based synthesis, segmental voice quality is further improved because this new synthesis technique aims at reducing speech modeling and manipulation in favor of speech unit selection. If the database offers sufficient prosodic and allophonic coverage, it is then even possible to generate natural sounding prosody without resort to signal manipulation. However, signal processing is still needed for segment concatenation and system flexibility (e.g., rate and pitch level changes).
This technique is capable to search for maximally long phonetic strings to match the sounds to be synthesized\cite{4}. It uses a large inventory to select the units to be concatenated\cite{2}.

**Advantages:** As compared to diphone or triphone synthesis, corpus-based speech tends to elicit considerably higher ratings of naturalness in auditory tests\cite{4}. This is because the number of real concatenation points become much smaller\cite{1,3}.

Moreover, the database of traditional diphone and triphone concatenation TTS systems is recorded with monotonous prosody whereas the units from a large speech corpus retain their natural and varied prosody. Thus, it becomes possible to Concatenate larger chunks of natural speech, providing superior quality over diphone and triphone concatenation\cite{1-3,5}.

As the corpus in its entirety provides the acoustic basis for such synthesis, the development of an optimal corpus represents an essential task of corpus-based synthesis\cite{2,4}.

The key idea of corpus-based synthesis, or unit selection, is to use an entire speech corpus as the acoustic inventory and to select at run time from this corpus the longest available strings of phonetic segments that match a sequence of target speech sounds in the utterance to be synthesized, thereby minimizing the number of concatenations and reducing the need for signal processing. In an ideal world, the target utterance would be found in its entirety in the speech database and simply played back by the system without any concatenations and without any signal processing applied, effectively rendering natural speech.

**MATERIALS AND METHODS**

**Method and development:** According to Fung and Meng\cite{7}, the development steps can be summarized as below:

- Corpus development for recording
- Waveform segmentation
- Unit selection for concatenated synthesis

According to Chou\cite{5}, a large speech corpus (on the order of 10 h, or even much more, of speech) produced by a single speaker is collected. The corpus is designed so that almost all linguistic and prosodic features for the target language (either for general domain or specific domain) have been included. Parallel analysis of all prosodic and linguistic features of the speech signals as well as the corresponding texts can lead to a much better prosodic model. There can be many repetitions of a given voice unit in the corpus, but in different context with different prosodic features.

In a different approach, longer elements (such as phrases, words or syllables) are also labeled in the speech database and can be selected directly (without the implicit selection mechanism of a cost function)\cite{1,9}.

**Speech corpus design criteria:** The criteria in designing speech corpus are size, corpus coverage, domain and quality as shown in Fig. 1. This section will discuss on the design criteria and conclude with the CBMTTS speech corpus’ design specification.

**Size of corpus:** The recent developed corpus-based speech synthesizers tend to rely on large-scale database, several vary from hours to more than 10 h speech corpora to provide sufficiently natural output speech\cite{8}. But the increase of corpus size will affect the unit selection time and slow down the process of synthesis\cite{4}.

Consequently, the size of corpus is the first issue to be tackled before other issues. This meant a comparison between the minimum and maximum sizes. A maximum size would mean a greater probability of the corpus containing the biggest possible units to match the text to be synthesized- from sound strings to words or even phrases\cite{4}. Unfortunately, big databases have been found complicated to maintain and even more complicated to annotate\cite{9}. Moreover, segmentation and tagging of corpus units is a cumbersome and time-consuming process - it has been found that one-minute corpus takes 1000 min to mark up. This is why some people decided to make the corpus as small as possible, yet containing as much relevant material as possible\cite{4}.

The design of a well utilized speech corpus requires determining an optimal set of elements for storage in the database. Optimality in this case means finding equilibrium between a large number of elements demanded by quality requirements and a minimal element number constrained by performance considerations\cite{1}.

![Fig. 1: Criteria in designing speech corpus](image-url)
Corpus database coverage: According to Ni [10], as anyone who has built a unit selection synthesizer, knows the quality of synthetic speech is highly dependent on the unit coverage of a speech corpus.

The corpus database design shall be “phonetically rich corpus”. In other word, it is to include as many as possible phonetic combinations, including intra-syllabic and inter-syllabic structures, in a corpus of acceptable size [5]. The corpus words should also include at least one instance of all possible units [1]. For instance, for Estonian speech which has diphones as smallest units, the first database for Estonian speech synthesis consisted of ca 1700 diphones.

To ensure this, some projects split the texts into two parts, designed along different lines. The first part provides coverage of frequent words and phrases which determined from the statistical properties of the target domain and allow selecting the longest possible elements for concatenation. The other part ensures coverage of diphones for the diphone based synthesis [1].

From the Estonian text analysis, it is concluded that the 10 most frequent word forms covered 31% of the input corpus. As little as 500 words ensure 92% coverage, while with 2300 words reach 99%. A corpus from an unrestricted domain requires approximately 70,000 word forms to reach 90% coverage [1].

According to Chou [5], the corpus is designed so that almost all linguistic and prosodic features for the target language (either for general domain or specific domain) have been included. Parallel analysis of all prosodic and linguistic features of the speech signals as well as the corresponding texts can lead to a much better prosodic model. There can be many repetitions of a given voice unit in the corpus, but in different context with different prosodic features.

Corpus database quality: According to Piits [4], a system with a good selection module and a high-quality speech corpus may yield output speech of extremely high quality, even if the signal processing module is rather simple.

The final quality of a concatenation synthesis system is directly related to the continuity of the spectrum at the concatenation point. A key factor in the final quality of a text-to-speech system based on unit concatenation is the continuity, or smoothness, of the formant trajectories at the concatenation point [11].

For high quality speech synthesis, the recording script should offer variety of comprehensive word [12]. Besides that, it should also cover all the prosodic and acoustic variations of the units. It is not feasible to record large and even large databases provide the support for the complexity and combinatorial of the language; instead it is need to find a way for optimal coverage of the language [2].

Corpus domain: The domain or focus application of design for CBMTTS is very important since the limited domain can reduce the size of corpus and yet preserve the quality of synthetic speech. Several projects has been developed in restricted domain such as in weather forecasts context [1,3] and talking clock. Since the CBMTTS has not yet being specified for any domain application, so the domain specification will not be taken into the consideration for this project.

Malay speech corpus design: Basically, the concatenation engine used for CBMTTS is an engine that using word-based concatenation engine. This engine first uses word corpus to match the input word as in Fig. 2. If the word is not existed in word corpus, then it will form the word from phoneme corpus. To provide the database support for word concatenation or phoneme concatenation, the speech corpus for word and phoneme has to be designed.

Malay text corpus design: Before the creation of speech corpus, text corpus should be designed first. To design the text corpus, few stages need to take into account [12]:

- Select a source text corpus to fit the target domains
- Analyze the source text corpus to obtain the unit statistics
- Select appropriate prompt subjects from the source text
- Inspect and remove unsuitable sentences

Fig. 2: Process of synthesize speech from word-based concatenate engine
For Malay text corpus design as shown in Fig. 3, firstly a set of 10 million words of digitized text will be collected from internet news. Then it will go through the process of analysis of the collected text by doing the word frequency count and filter all the non Malay words.

Finally, a set of high frequency word will be used for designing sentences which are for speech corpus design.

Speech corpus design: As shown in Fig. 4, the process of corpus building starts from text corpus design and then speech corpus design.

For text corpus design, a set of online Malay News consisting of 10 million words has been collected in 3 months. Since all the texts are in html format, a text processing tool has been designed to extract text automatically from html file and do the word frequencies count[13]. The information of 10 millions words is shown in Table 1. The design of a well utilized speech corpus requires determining an optimal set of elements for storage in the database[1]. To help in finding an element set of optimal size and composition along these guidelines, some statistical analyses have been conducted[1].

<table>
<thead>
<tr>
<th>Categories</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berita Harian</td>
<td>3033569</td>
<td>30.25</td>
</tr>
<tr>
<td>Bernama</td>
<td>1218054</td>
<td>12.15</td>
</tr>
<tr>
<td>Cybersing</td>
<td>1211249</td>
<td>12.08</td>
</tr>
<tr>
<td>Malaysia Kini</td>
<td>2968951</td>
<td>29.61</td>
</tr>
<tr>
<td>Utusan Melayu</td>
<td>1595303</td>
<td>15.91</td>
</tr>
<tr>
<td>Total sum</td>
<td>10027126</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 2: 10 millions words resource distribution

<table>
<thead>
<tr>
<th>Categories</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berita Harian</td>
<td>3033569</td>
<td>30.25</td>
</tr>
<tr>
<td>Bernama</td>
<td>1218054</td>
<td>12.15</td>
</tr>
<tr>
<td>Cybersing</td>
<td>1211249</td>
<td>12.08</td>
</tr>
<tr>
<td>Malaysia Kini</td>
<td>2968951</td>
<td>29.61</td>
</tr>
<tr>
<td>Utusan Melayu</td>
<td>1595303</td>
<td>15.91</td>
</tr>
<tr>
<td>Total sum</td>
<td>10027126</td>
<td>100</td>
</tr>
</tbody>
</table>

Text corpus collection and analysis: The 10 millions words from internet news have been collected for few weeks. There are altogether 10,027,126 collected texts as shown in Table 1. The duration for collecting the text is 3 months and altogether 115738 Malay words in the database. The detail of the texts source is listed in Table 2. From Table 2, it can be seen that the text source from Berita Harian is the highest and contributes 30.25 percent. It is followed by Malaysia Kini which contributes 29.61%.

But the text is in “.html” file, so a tool for extracting text from “.html” files has been designed to extract the text and put it into a file and in sentences format as shown in Fig. 5. Then, the sentences are break into words and the words will go through the process of word frequency count.

The result of top 10 highest frequency words is listed in Table 3. It can be seen that the top highest frequency words have been contributing 10.58 percent of the overall 10 millions texts.

Word selection: The CBMTTS utilizes word based concatenation engine if the target word not support in word-corpus, then the word will be constructed through phoneme based unit selection. To provide highest coverage of words for word based concatenation, word selection is the most important issue to be tackled here.
There are many ways of selecting the words for sentences design. One of the ways is using words selected from frequency dictionary for example Estonian TTS system using frequency dictionary Estonian, which is based on texts from media and fiction⁵. But there are no available frequency word dictionary for Malay language, thus, the frequency count from 10 million words has been used for this purpose.

The design of text corpus will first focus on the word coverage by selecting top highest frequency words that can cover up to 70% of the existing words. To select those words, the word needs to be grouped into few categories such as in Table 4. Then among the groups identify those words that fulfill the requirement of 70% coverage. From Table 4, it can be seen that the word in categories above 1000 have covered 70% of word occurrence. So, those words have been selected to design the carrier sentences. Figure 6 shows the coverage graph for 10 millions words. From this graph, it can be seen that altogether 1451 high frequency words covered 70% of overall words in 10 million collected texts corpus. The word coverage in different word frequency is listed in Table 5. From this table, the word with initial “d” contribute the highest portion of existing word followed by “m” and “s”.

### Table 3: 10 highest frequency words

<table>
<thead>
<tr>
<th>Word</th>
<th>Counting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang</td>
<td>282200.00</td>
</tr>
<tr>
<td>Dan</td>
<td>253097.00</td>
</tr>
<tr>
<td>Untuk</td>
<td>91374.00</td>
</tr>
<tr>
<td>Tidak</td>
<td>84443.00</td>
</tr>
<tr>
<td>Pada</td>
<td>60179.00</td>
</tr>
<tr>
<td>Akan</td>
<td>58222.00</td>
</tr>
<tr>
<td>Saya</td>
<td>55550.00</td>
</tr>
<tr>
<td>Kepada</td>
<td>55497.00</td>
</tr>
<tr>
<td>Mereka</td>
<td>55175.00</td>
</tr>
<tr>
<td>Ke</td>
<td>42310.00</td>
</tr>
</tbody>
</table>

Table 4: Word coverage according to word frequency groups

<table>
<thead>
<tr>
<th>Occurrence</th>
<th>Total words</th>
<th>Total occurrence</th>
<th>% coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;5000</td>
<td>281</td>
<td>4527844</td>
<td>45.16</td>
</tr>
<tr>
<td>2501-5000</td>
<td>319</td>
<td>1115981</td>
<td>11.13</td>
</tr>
<tr>
<td>2001-2500</td>
<td>147</td>
<td>327573</td>
<td>3.27</td>
</tr>
<tr>
<td>1501-2000</td>
<td>278</td>
<td>480475</td>
<td>4.79</td>
</tr>
<tr>
<td>1001-1500</td>
<td>426</td>
<td>521977</td>
<td>5.21</td>
</tr>
<tr>
<td>Total</td>
<td>1451</td>
<td>6973850</td>
<td>69.55</td>
</tr>
</tbody>
</table>

### Table 5: Word frequency count according to word occurrence categories

<table>
<thead>
<tr>
<th>Word starting with</th>
<th>&gt;5000</th>
<th>2501-5000</th>
<th>2001-2500</th>
<th>1501-2000</th>
<th>1001-1500</th>
<th>Total</th>
<th>% of occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>314552</td>
<td>63421</td>
<td>42621</td>
<td>23857</td>
<td>24799</td>
<td>451250</td>
<td>6.47</td>
</tr>
<tr>
<td>b</td>
<td>344697</td>
<td>107245</td>
<td>31825</td>
<td>47646</td>
<td>46535</td>
<td>577948</td>
<td>8.29</td>
</tr>
<tr>
<td>c</td>
<td>5272</td>
<td>15422</td>
<td>10958</td>
<td>5496</td>
<td>6172</td>
<td>43320</td>
<td>0.62</td>
</tr>
<tr>
<td>d</td>
<td>793327</td>
<td>53176</td>
<td>13203</td>
<td>27675</td>
<td>30912</td>
<td>918293</td>
<td>13.17</td>
</tr>
<tr>
<td>e</td>
<td>12886</td>
<td>3749</td>
<td>4741</td>
<td>5098</td>
<td>8613</td>
<td>35087</td>
<td>0.50</td>
</tr>
<tr>
<td>f</td>
<td>7430</td>
<td>0</td>
<td>4023</td>
<td>0</td>
<td>8817</td>
<td>20270</td>
<td>0.29</td>
</tr>
<tr>
<td>g</td>
<td>0</td>
<td>10566</td>
<td>4507</td>
<td>6728</td>
<td>11738</td>
<td>33539</td>
<td>0.48</td>
</tr>
<tr>
<td>h</td>
<td>58892</td>
<td>32796</td>
<td>13179</td>
<td>8631</td>
<td>10689</td>
<td>124187</td>
<td>1.78</td>
</tr>
<tr>
<td>i</td>
<td>299063</td>
<td>19927</td>
<td>8698</td>
<td>8808</td>
<td>9564</td>
<td>346860</td>
<td>4.97</td>
</tr>
<tr>
<td>j</td>
<td>92426</td>
<td>44151</td>
<td>4527</td>
<td>9751</td>
<td>10980</td>
<td>161835</td>
<td>2.32</td>
</tr>
<tr>
<td>k</td>
<td>368520</td>
<td>141174</td>
<td>4571</td>
<td>51065</td>
<td>59050</td>
<td>637520</td>
<td>9.14</td>
</tr>
<tr>
<td>l</td>
<td>134874</td>
<td>26879</td>
<td>4571</td>
<td>14098</td>
<td>18045</td>
<td>198467</td>
<td>2.85</td>
</tr>
<tr>
<td>m</td>
<td>437135</td>
<td>157995</td>
<td>37842</td>
<td>74997</td>
<td>82837</td>
<td>790806</td>
<td>11.34</td>
</tr>
<tr>
<td>n</td>
<td>61217</td>
<td>12601</td>
<td>4470</td>
<td>10258</td>
<td>5211</td>
<td>93757</td>
<td>1.34</td>
</tr>
<tr>
<td>o</td>
<td>59908</td>
<td>0</td>
<td>2137</td>
<td>0</td>
<td>3976</td>
<td>66021</td>
<td>0.95</td>
</tr>
<tr>
<td>p</td>
<td>305175</td>
<td>154448</td>
<td>55810</td>
<td>60090</td>
<td>67866</td>
<td>643389</td>
<td>9.23</td>
</tr>
<tr>
<td>q</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>r</td>
<td>43143</td>
<td>16258</td>
<td>8889</td>
<td>6263</td>
<td>13653</td>
<td>88206</td>
<td>1.26</td>
</tr>
<tr>
<td>s</td>
<td>457194</td>
<td>157702</td>
<td>37543</td>
<td>68069</td>
<td>46997</td>
<td>767505</td>
<td>11.01</td>
</tr>
<tr>
<td>t</td>
<td>313906</td>
<td>81007</td>
<td>26644</td>
<td>31047</td>
<td>35117</td>
<td>487721</td>
<td>2.12</td>
</tr>
<tr>
<td>u</td>
<td>117011</td>
<td>7133</td>
<td>4824</td>
<td>10616</td>
<td>8543</td>
<td>148127</td>
<td>2.12</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>w</td>
<td>18216</td>
<td>7683</td>
<td>4804</td>
<td>10282</td>
<td>5395</td>
<td>46380</td>
<td>0.67</td>
</tr>
<tr>
<td>x</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>y</td>
<td>282200</td>
<td>2648</td>
<td>0</td>
<td>0</td>
<td>2393</td>
<td>287241</td>
<td>4.12</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>0</td>
<td>2046</td>
<td>0</td>
<td>4075</td>
<td>6121</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Total occurrence Words 4527844 1115987 327573 480475 521977 6973850 100.00
Total occurrence words 452788 56432885 5971398 6451873 6973850 29568790

### Sentence design: The aim of sentences design is to create a list of text corpus that covers all the selected 1451 words.
RESULTS

Table 6 shows the total units after extract the phoneme units from the carrier sentences. It can be seen that “a” and “e” is two phonemes with highest occurrence in Malay language. They contribute almost 18.36 and 8.6%. The least phoneme is “_z” and “iu”. The phoneme speech units has been extracted and grouped in same phoneme folder instead of keep it in original carrier sentence which to reduce the buffer or memory allocation when the system want to access to certain phoneme from its source. The length of sentences may vary from 1.994s to 6.608s and the phoneme is only between 32 and 254ms. To allocate memory for phoneme is smaller than allocate memory for sentences. Then, extract the phoneme from the carrier sentence. For instances, if we intend to form a sentence when consist of 15 phoneme, it may be needed to allocate 15 times of memory for each origin sentences before it can be extracted from the origin source. This will consume a lot of memory and slow down the process of concatenation.

DISCUSSION

The summary of the corpus database for CBMTTS is shown as in Table 7. From this design there are altogether 381 carrier sentences which consist of 16826 phoneme units. All the phone units are phonetically transcript to preserve the phonetic context of its origin that will be used for phonetic context unit selection. The total storage size is 37.6Mb which is bigger than previous Malay TTS diphone database because it uses real wave and support unit selection which provide more choice for each target unit.

Table 6: Total units after extract the phoneme units from the carrier sentences

<table>
<thead>
<tr>
<th>pho</th>
<th>Total (%)</th>
<th>pho</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>_a</td>
<td>107</td>
<td>b</td>
<td>253</td>
</tr>
<tr>
<td>_ai</td>
<td>4</td>
<td>c</td>
<td>77</td>
</tr>
<tr>
<td>_au</td>
<td>1</td>
<td>d</td>
<td>313</td>
</tr>
<tr>
<td>_b</td>
<td>256</td>
<td>e</td>
<td>1448</td>
</tr>
<tr>
<td>_c</td>
<td>29</td>
<td>eh</td>
<td>124</td>
</tr>
<tr>
<td>_d</td>
<td>269</td>
<td>f</td>
<td>38</td>
</tr>
<tr>
<td>_e</td>
<td>3</td>
<td>g</td>
<td>169</td>
</tr>
<tr>
<td>_eh</td>
<td>10</td>
<td>h</td>
<td>374</td>
</tr>
<tr>
<td>_f</td>
<td>17</td>
<td>i</td>
<td>970</td>
</tr>
<tr>
<td>_g</td>
<td>30</td>
<td>ia</td>
<td>87</td>
</tr>
<tr>
<td>_h</td>
<td>65</td>
<td>io</td>
<td>3</td>
</tr>
<tr>
<td>_i</td>
<td>74</td>
<td>_i</td>
<td>1</td>
</tr>
<tr>
<td>_j</td>
<td>12</td>
<td>j</td>
<td>164</td>
</tr>
<tr>
<td>_k</td>
<td>49</td>
<td>k</td>
<td>665</td>
</tr>
<tr>
<td>_lh</td>
<td>248</td>
<td>kh</td>
<td>7</td>
</tr>
<tr>
<td>_l</td>
<td>72</td>
<td>l</td>
<td>514</td>
</tr>
<tr>
<td>_m</td>
<td>447</td>
<td>m</td>
<td>492</td>
</tr>
<tr>
<td>_n</td>
<td>33</td>
<td>ng</td>
<td>500</td>
</tr>
<tr>
<td>_ny</td>
<td>2</td>
<td>ny</td>
<td>91</td>
</tr>
<tr>
<td>_o</td>
<td>21</td>
<td>o</td>
<td>206</td>
</tr>
<tr>
<td>_p</td>
<td>320</td>
<td>p</td>
<td>276</td>
</tr>
<tr>
<td>_q</td>
<td>58</td>
<td>q</td>
<td>1</td>
</tr>
<tr>
<td>_s</td>
<td>258</td>
<td>r</td>
<td>838</td>
</tr>
<tr>
<td>_sy</td>
<td>5</td>
<td>s</td>
<td>410</td>
</tr>
<tr>
<td>_t</td>
<td>178</td>
<td>sy</td>
<td>8</td>
</tr>
<tr>
<td>_u</td>
<td>42</td>
<td>t</td>
<td>652</td>
</tr>
<tr>
<td>_v</td>
<td>5</td>
<td>u</td>
<td>696</td>
</tr>
<tr>
<td>_w</td>
<td>18</td>
<td>ua</td>
<td>107</td>
</tr>
<tr>
<td>_y</td>
<td>59</td>
<td>ui</td>
<td>2</td>
</tr>
<tr>
<td>_z</td>
<td>1</td>
<td>v</td>
<td>10</td>
</tr>
<tr>
<td>a</td>
<td>3076</td>
<td>w</td>
<td>72</td>
</tr>
<tr>
<td>ai</td>
<td>97</td>
<td>y</td>
<td>52</td>
</tr>
<tr>
<td>au</td>
<td>26</td>
<td>z</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 7: Summary of corpus database for CBMTTS

<table>
<thead>
<tr>
<th>Specification</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total carrier sentences</td>
<td>381</td>
</tr>
<tr>
<td>Total phoneme</td>
<td>16826</td>
</tr>
<tr>
<td>Total size</td>
<td>37.6Mb</td>
</tr>
<tr>
<td>Storage format</td>
<td>Wave</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>16kHz</td>
</tr>
<tr>
<td>Recording environment</td>
<td>Normal room (control noise)</td>
</tr>
<tr>
<td>Recording equipment</td>
<td>Multi-speech 4500 (high quality speech DAQ and microphone)</td>
</tr>
<tr>
<td>Speaker</td>
<td>Female</td>
</tr>
</tbody>
</table>

CONCLUSION

A set of Malay Speech Corpus has been designed through this project and it is expected to cover around 70% of high frequency words. This corpus may not be the best corpus that supports most of the existing words but it is sufficient to provide the large coverage of words and at the same time keeping the data size at minimum level such as 381 sentences with 37.6 Mb. It can be further enlarged to bigger size and coverage by selecting more words but will be traded off with storage size.
ACKNOWLEDGEMENT

This research project is supported by CBE (Central of Biomedical Engineering) at University Technology Malaysia and funded by Minister of Science and Technology (MOSTI), Malaysia under grant “To Develop a Malay Speech Synthesis System for Standard Platform Compatibility and Speech Compression” Vot 79190.

REFERENCES


