TAGCO: A Tagalog Speech Corpus

Quennie Joy B. Mesa

Abstract — Speech corpus design is one of the key issues in building high quality text-to-speech synthesis system. The understandability and naturalness of output speech depends on the quality of the speech database being used. This paper presents the design, development and content construction of a Tagalog speech corpus. The corpus contains 156.53 minutes recording of read speech and 99.34 minutes of spontaneous speech. Speech recordings are accompanied by word level transcription and word-to-phoneme lexicon. The corpus is designed for the development of an HMM-based speech synthesis system, and may also be used as basic infrastructure building in speech processing and applied speech technology.

Index Terms— Corpus analysis, corpus development, phoneme set, speech annotation, speech corpus, speech database, speech segmentation.

1 INTRODUCTION

Speech corpus is the building block of all spoken technology. The quality and volume of speech data in the corpus defines the performance of any spoken language-based systems [1], [3], [4]. Ideally, in the construct of a data-driven acoustic model, speech training datasets should reflect the tasks for which the synthesizer is built. Using training datasets similar in style and contents to that of the data to be synthesized ensures appropriate balance in terms of acoustic information. Problems can arise when contexts required in synthesis is different from anything seen in training [4].

However, for statistical model-based approaches to speech modelling such as HMMs, it requires minimal training datasets to estimate all the parameters of the model [5], [6] thus, reducing the amount of speech data in a corpus.

In this paper information on the script design, data collection procedure and text data processing used in the development of a Tagalog speech corpus was presented.

2 CORPUS DEVELOPMENT

2.1 Script Design and Phonetic Coverage

2.1.1 Read Text

Text gathered from Tagalog news articles from the internet were used for the construct of the read text corpus. Although the grammar and structure of the Tagalog language used in the news articles was somewhat different from the spoken language, it had informal contents and includes non-grammatical expressions thus approximating the natural attributes of the language [7].

Elicitation of phonemic and prosodic cues that characterizes the Tagalog speech were the criteria used in the selection of sentences [8]. Each news article contains 3 to 52 sentences. Each sentence contained 2 to 13 words and had either of the following themes: (1) storytelling, (2) describing a scene (3) relating a news event and (4) relating an emotional event.

The speaker read a total of 12,356 words in the recording sessions. The text manuscript consists of five hundred seventeen interrogative sentences, 187 exclamation and 1,580 declarative sentences. The data were collected for about eight weeks.

Several limits on the sentence length was observed to ensure readability:

- Maximum word length: 20 characters
- Maximum number of words in a sentence: 13

2.1.2 Spontaneous Speech

The spontaneous speech data consist of unscripted and unprompted dialogues. Speakers are allowed to pursue or converse with whichever topic they feel comfortable about.

Although the majority of the recordings consist of conversations, there are some in which the speakers assume an interviewer-interviewee relationship.

2.2 Speaker Selection

The quality of recorded utterances determines the performance of a speech engine [1], [3], [4]. Recording condition, timbre and speaking style of speaker were some of the attributes that defines the recorded speech quality. Speaker selection is an essential and critical step in the development of any spoken corpora [7].

For this study, the criteria used for speaker selection are the following; (1) native speaker; (2) ability to read-out the whole text script at normal speaking rate and pronunciation, and (3) can converse fluently using the language.

As a result, a female Tagalog native speaker was chosen for the read speech and two female native speakers for the spontaneous speech with the following profile as shown in Table 1;
Table 1. Speaker Profile

<table>
<thead>
<tr>
<th>Speaker Attributes</th>
<th>Read Speech</th>
<th>Spontaneous Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
<td>Female</td>
</tr>
<tr>
<td>Language</td>
<td>Native Pilipino</td>
<td>Native Pilipino</td>
</tr>
<tr>
<td>Dialect</td>
<td>Tagalog</td>
<td>Tagalog</td>
</tr>
<tr>
<td>Age</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Occupation</td>
<td>Teacher</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

2.3 Recording Specification
The recording of utterances was done in a recording studio. Signals are sampled synchronously and stored directly to a hard disc. Speech data was digitized at a sample rate of 44.1 kHz and stored as a mono-bit wav file.

For the read speech, the speaker was given a copy of the speech manuscript and instructed to read the text at normal speaking rate and pronunciation. The speech is organized in 2,284 spoken utterances containing 12,356 words, 11,800 were native words and 556 loan words. For each sentence, the speaker can utter it repeatedly until she is satisfied with her pronunciation; however, all the mispronounced or unsatisfied utterances are backed up at the same time.

For the spontaneous speech, speakers were permitted to speak on any topic they like. Sentences may not be grammatically correct; may not be a complete sentence; with interjections, and words may be repeated.

In the course of a conversation, speakers may interrupt each other; speak with one another, may use different words (loan or colloquial), may use different prosody, different intonations and different accents.

3. TEXT DATA PROCESSING
In this study, a chain of text processing phases was implemented as shown in Fig.1.

3.1 Phase 1: Annotation
For speech data annotation, two kinds of sound editors were used. Preprocessing of sound file prior to actual annotation was done using the Speech Filing System (SFS) [9]. Sixteen-bit spectral was subtracted to clean the speech data.

For sound energy standardization a 16-bit automatic gain was selected. Using the same software, speech data was segmented into sentences and manually saved into separate sound files.

Going further, the clean and preprocessed sound files are annotated into words, syllables and intonation layers. Words are labelled orthographically and the syllables are transcribed as shown in Fig. 2. All annotations were done manually.

3.2 Phase 2: Syllabification
For phase 2, the normalized read and spontaneous speech were syllabified. The syllabification process consists of cutting-up words into syllables [7]. For this study, words in a sentence were not only segmented into syllables, also the syllable’s position in a word was labeled. For the onset syllable (syllable at the beginning of a word), numeric 1 is affixed before the syllable. For the coda syllable (syllable at the end of a word) numeric 2 was used. Middle syllables do not have any numeric label for it is assumed that its’ duration is fixed.

Table 2. Example of a labeled file

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Text File (Orthography)</th>
<th>Pronunciation (Syllable)</th>
<th>Break Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent00031</td>
<td>dahl</td>
<td>1 da.heel2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>dito</td>
<td>1 de.eh.toh2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>muling</td>
<td>1 moo.leeng2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>nabuksan</td>
<td>1 na.book.san2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>ang</td>
<td>1 ang</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>isyu</td>
<td>1 ees.yoo2</td>
<td>1</td>
</tr>
</tbody>
</table>

3.3 Phase 3: Transcription Correction
Despite efforts of minimizing background noise and pronunciation mistakes etc. while recording, it is inevitable for errors to crept in. For this study, errors were identified manually by listening to the recorded spoken speech. Criteria used for speech quality classifications are: (1) good; (2) with channel distortion; (3) with background noise, and (4) useless. Pronunciation mistakes were carefully identified and corresponding changes/revisions were made to ensure that utterances correspond with the transcriptions.

4 CONCLUSION
The utmost objectives of the study described is to generate a speech corpus that will be used for the development of an HMM-based Tagalog speech synthesizer and a corpus that
may be used as basic infrastructure building in Tagalog language speech processing and applied speech technology. The study provides specifics in reference with the materials, methodology and acoustic basis in the generation of a relatively naturally sounding synthetic speech.

ACKNOWLEDGMENT

The author acknowledges the assistance provided to her by Dr. Kyung-Tae Kim, Dr. Jong-Jin Kim and Dr. Y.J Lee and the Isabela State University. This work has been supported by the Hannam University Research Fund.

REFERENCES

[9] https://www.phon.ucl.ac.uk/resource/sfs/