Readability Consideration in Speech Synthesis Recording Script Selection

Minghui Dong, Ling Cen, Paul Chan, Haizhou Li

Institute for Infocomm Research, A-Star, 1 Fusionopolis Way, Singapore 138632
{mhdong, lcen, yzchan, hli}@i2r.a-star.edu.sg

Abstract

Designing text scripts that cover enough phonetic units and prosodic phenomena is very important when recording speech database for corpus based speech synthesis. When designing recording scripts for speech synthesis databases, a lot of effort is often placed on how to achieve maximal coverage of phonetic units in minimal speech recording. However, when we try to select sentences that have optimal coverage of the speech phenomena, some sentences with difficult words or incorrect grammar are often selected. It is difficult for speakers to read these sentences correctly and naturally at the same time. In order to address the problem in building speech database, we propose a selection process to create easy-to-read text scripts for recording in this paper. In this work, we will consider how to build a candidate set that is easy to read so that the speaker can utter it in the most natural way. We will calculate the statistics of the English text by analyzing the English Gigaword corpus, and filer out the sentences containing infrequent words and bigrams. The experiment shows that the selected scripts have good unit coverage of the language and good readability.

Keywords

Text-to-speech, recording scripts, text selection, TTS corpus design, text readability.

1 Introduction

In the past decade, corpus based text-to-speech (TTS) methods have been working well in generating high quality speech. Two commonly adopted methods are the unit-selection based method (Hunt et al 1996) and the HMM based statistical method (Yoshimura et al 1999). In unit selection methods, a corpus is needed to train prosody models and to create the unit selection database. In HMM-based methods, we also need a corpus to train speech synthesis models. As it is expensive and time consuming to record TTS databases, the recording script for building the TTS database is usually carefully designed to have the largest coverage of the concerned phonetic and prosodic phenomena with the least recording data.

For both of the TTS methods mentioned, in order to cover possible pronunciation occurrences in the TTS systems, it is expected to have a corpus that covers sufficient phonetic elements, such as phones, diphones, triphones, syllables, words, etc. To cover prosodic variations, it is expected that the corpus should cover sufficient elements with different accents, stresses, prosodic breaks, etc. As such, text selection methods are often designed to cover the phonetic and prosodic variations as much as possible. Greedy algorithms are often

designed to select the text sentence that covers most frequent elements first. A lot of research (Santen 1997; Black et al 2007: François 2001; Kominek et al 2003; Bozkurt et al 2003; Lambert et al 2004; Isogai et al 2005) has been done on this topic.

The aforementioned methods succeed in many TTS corpus designs. However, most algorithms emphasize on the coverage of the text, while neglecting its readability. In our experience in TTS database recording, there is a high chance of encountering ungrammatical or anomalous text in automatically selected recording scripts. When reading such sentences, the speakers often make mistakes, and the recordings often sound unnatural. The meaningless sentences also do not help in prosody training. In this work, we will focus on the readability of the selected text in addition to the consideration of coverage of speech phenomena.

In the next section, we will first make some definitions. The methods used will be described in section 3, the experiments carried out will be explained in section 4 and the conclusions will be made in the final section.

2 Definitions

In our work, we will draft a recording script for English Text-to-speech corpus. In this section, we will define some of the terms and conventions used in building the text script set.

2.1. Phonetic and prosodic consideration

Since the text scripts designed are expected to be phonetically and prosodically balanced, we first need to determine how to include phonetic and prosodic elements in our text selection. In our work, the basic unit is the phone. On the phonetic aspect, phonetic context is one of the important factors for co-articulation in TTS system. As phonetic contexts in a syllable are more coherent than that between syllables, we use the syllable as our selection element. On the prosodic aspect, accent is the most important aspect and is easy to determine. Therefore, we further include accent into our consideration. Thus, we use syllable with accent marks as our unit for text selection. For a syllable, the accented version and the unaccented version are considered two different elements in our selection process.

Table 1. Sample lexicon items

Word	Pronunciation
abandoning	(@)0 (b-a-n)1 (d-@-n)0 (i-ng)0
adjustment	(@)0 (jh-uh-s-t)1 (m-@-n-t)0
education	(e)1 (jh-uw)0 (k-ei)1 (sh-n!)0
irrelevant	(i)0 (r-e)1 (l-@)0 (v-n!-t)0
student	(s-t-y-uu)1 (d-n!-t)0

The Unilex lexicon (Fitt et al 1999) provided by University of Edinburgh, UK is used in our work, from which, we have generated the Received Pronunciation. This is a UK English lexicon consisting of 119,356 word items. The lexicon includes the inflection forms of most words. We converted each word item into syllable sequences with its accentuation status marked on each syllable. The lexicon is organized as shown in Table 1. In the table, each

word is decomposed into syllables, where each pair of brackets marks a syllable and 1 or 0 indicates whether or not it is accented.

2.2. Statistics of the language

To have a statistical knowledge of the English language, we used the LDC English Gigaword Corpus (Parker et al 2009) as our reference corpus, which is also used as the source of script selection later. We calculate the following statistics of the language:

- Word frequency: The number of occurrences of a word in the corpus. From word frequency, the syllable frequency can be calculated with the help of the lexicon.
- Word bigram frequency: The number of occurrences of a word bigram in the corpus. The word bigram frequency is used as an index to judge the readability for candidate text set selection.

2.3. Measurements of the selected text

In this part, we will describe the measurements that we used to measure the selected text. The measurements are basically used to measure the coverage of the selected text on the language as well as its readability. The following measurements are used in our work:

Token Coverage Rate (TCR): Token coverage rate is the indication of how many unique basic elements have been covered by the text sentences. Suppose X is the text set that we have selected, Y is the corpus, the token coverage rate is defined as follows:

$$T(X) = U(X) / U(Y) \tag{1}$$

where U(x) is the number of unique tokens in the text set x.

Corpus Coverage Rate (CCR): Corpus coverage rate measures how much of the occurrences of elements in the text corpus is covered by the selected text. Suppose $x_1, x_2, ..., x_m$ are the unique tokens in the text set that we have selected, $y_1, y_2, ..., y_n$ are the unique tokens in the corpus, the corpus coverage rate is defined as follows:

$$C(X) = \sum_{i=1}^{m} f(x_i) / \sum_{i=1}^{n} f(y_i) \tag{2}$$

where $f(x)$ is the frequency of token x in the corpus.

TCR measures the coverage of unique units, while CCR measures the coverage of the units in the language. Bigger corpus coverage means better coverage of the language.

Research on text readability has established a relationship between readability and text properties (e.g. words per sentence, average number of syllables per word, etc) by multiple correlation analysis of human graded text (e.g. Flesch 1948 and Kincaid et al 1975). We will use two widely used measures in our work.

Flesch Reading Ease Score (FRES): FRES score measure the readability of text, and is calculated as the follows [12]:

$$E(X) = 206.835 - 1.015 \frac{N_w(X)}{N_S(X)} - 84.6 \frac{N_l(X)}{N_w(X)} \tag{3}$$

where $N_s(X), N_w(X)$ and $N_l(X)$ are number of sentences, words and syllables in the text X respectively. In the Flesch reading ease test, higher scores indicate material that is easier to read.

Flesch–Kincaid Grade Level (FKGL): FKGL translates FRES score to a U.S. grade level, making it easier to judge the readability level of texts. It can also mean the number of

years of education generally required to understand this text. FKGL grade level is calculated as follows (Kincaid 1975):

$$L(X) = 0.39\frac{N_w(X)}{N_S(X)} - 11.8\frac{N_l(X)}{N_w(X)} - 15.59 \quad (4)$$

where $N_s(X)$, $N_w(X)$ and $N_l(X)$ are number of sentences, words and syllables in the text X respectively.

3 Text Selection Methods

3.1. Preprocessing of corpus

The first step of the text processing is to identify the text sentences. Each sentence is normally ended with a period, a question mark or an exclamation mark. However, the period is not only used for marking the end of a sentence. It is also used for abbreviations, such as Mr., Dr., U.N., etc. The sentence identification process needs to exclude such exceptions. In our work, an abbreviation list is created. When a period is detected, the list is first checked to judge whether it is an acronym. It is otherwise considered to be the end of a sentence. Some examples from the list of exceptions are given in Table 2.

Table 2. Sample abbreviations

Dec.	Miss.	Nov.
Del.	Mo.	Oct.
Dept.	Mr.	Okla.
Dr.	Mrs.	Ont.
Drs.	Ms.	Ore.
Etc.	Neb.	Pa.
Feb.	Nev.	Ph.
Fla.	No.	Prof.
Ft.	Nos.	Prop.

3.2. Candidate sentence set selection

When sentences are identified, the next step is to filter the sentences that are unsuitable for use in recording scripts. The following rules are used to filter out the unwanted texts:

(1) The number of words in a sentences is limited as

$$w_l \le w \le w_u, \quad (5)$$

where w is the acceptable number of words in a sentence, w_l and w_u are the lower and upper bounds of w, respectively. Sentences with words more than w_u or less than w_l are excluded. Overly long sentences are normally more difficult to and it is not efficient to record very short sentences.

(2) Sentences with words that are not found in the lexicon are excluded. Recording out-of-vocabulary word may make it difficult to create phonetic transcription in later stage.
(3) Sentences with less frequently used words are excluded. Less frequently used words are normally difficult. This ensures the speaker is able to read each word correctly and easily.
(4) Sentences with less frequent word bigrams are excluded. Sentences with low bigram frequency are more likely to be ungrammatical sentences, which are not easy to read. It is therefore necessary to exclude such sentences.

Among the above rules, (3) and (4) are based on the statistics of the language. The two rules help to improve the readability of the text.

3.3. Recording script selection

There are two generally used greedy algorithms for text selection. The first method is the most frequent first (MFF) selection method. In each round, the sentence that covers most uncovered element in the corpus will be selected. This method tries to generate sentence text with highest corpus coverage rate.

The second is the least frequent first (LFF) selection method. In each round, the sentence that covers the least uncovered elements in the corpus will be selected. This ensures that the least frequently used elements are covered, as the frequent ones are normally been covered when the least frequent ones are covered. This method tries to generate text with highest token coverage rate.

The MFF method is used in our experiment as we wish to achieve the maximal coverage of the language.

4 Experiments

Firstly, we conducted analysis for corpus, and then performed experiments to examine our selection strategy.

4.1. Description of the text corpus

In our work, the text corpus we used is the English Gigaword Corpus Fourth Edition from LDC (Corpus LDC2009T13) (Parker et al 2009). The corpus is a comprehensive archive of newswire text data that has been acquired over several years by the LDC. The content of the corpus comes from six sources:

- Agence France-Presse, English Service
- Associated Press Worldstream, English Service
- Central News Agency of Taiwan, English Service
- Los Angeles Times/Washington Post Newswire Service
- New York Times Newswire Service
- Xinhua News Agency, English Service

The corpus consists of 19.4 GB English text, which includes about 2.97 billion of words in 7.15 million of documents. The documents are classified into four categories, namely, story, multi, advisory and others. Considering the huge size of the text corpus, the statistics from the corpus can be considered a reliable reference to English language.

4.2. Statistics of the language

We have calculated the statistics of the corpus. There are totally 2,288,791 unique words in the corpus. We sorted the word frequency in descending order and calculated the accumulative percentage of word items in the corpus. The accumulative percentage of the first 30,000 words is as shown in Figure 1. From the figure, we can see that the most frequent 10,000 words cover more than 90% of words in the corpus. The percentage increases slowly when the number of words is more than 10,000. From our calculation, the most frequent 20,000 words cover 95.56% of the words in corpus.

We also calculated the frequency of word bigram. In this calculation, we group infrequently used words (frequency rank more than 20000) into one single category.

We also sorted the bigram frequency in descending order, and calculated the accumulative coverage percentage. The result is as shown in Figure 2. From the figure, we can see that most frequent 1,000,000 word pairs cover about 90% of the bigram occurrences.

By referring to the lexicon, we have calculated the frequency of syllables in the corpus. Totally, there are 17385 syllables (with accent marks) in the corpus. Among them, the most frequent 4000 syllables cover about 98.16% of all the syllable occurrences in the corpus.

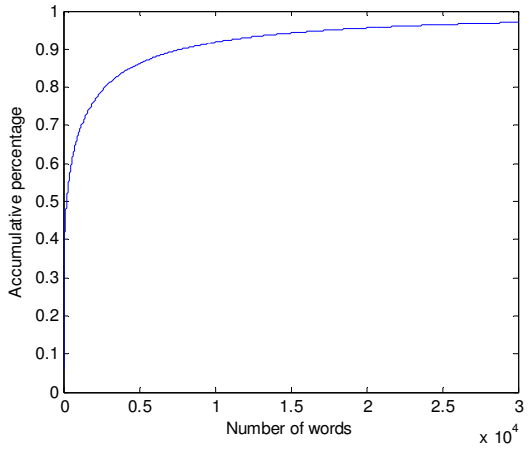

Figure 1. Accumulative percentage of words in the corpus

4.3. Candidate set selection

By applying the rules defined in 3.2, we have selected a collection of text sentences from the corpus. We have excluded the sentences that have less than 8 or more than 16 words during selection. We only select sentences containing most frequent 20,000 words and the first 1,000,000 bigrams. Finally, 1,204,791 sentences have been selected.

We have inspected 400 sentences manually and found them all grammatically correct. We asked 4 non-native English speakers to review the 400 text sentences, and none of them identified unknown words in the text. In contrast, the original texts are full of incomplete sentences, foreign names, difficult words like place names in addresses, email addresses, internet links, DNA sequences, computer commands, symbols, etc. Sometimes, almost 30% of them are not grammatically correct sentences. Thus, it has been shown that filtering the sentences with infrequent words and infrequent bigrams help to generate text with higher readability, even for non-English readers.

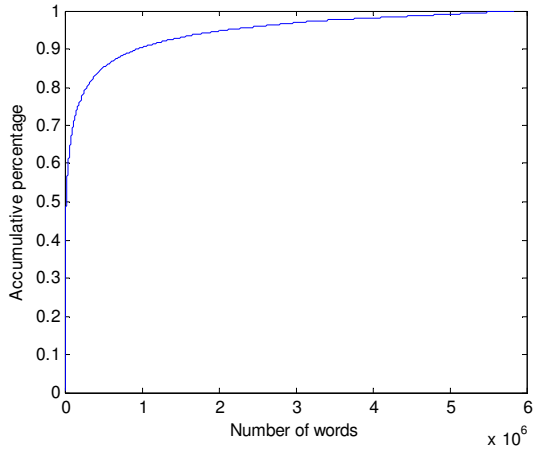

Figure 2. Accumulative percentage of word bigrams in the corpus.

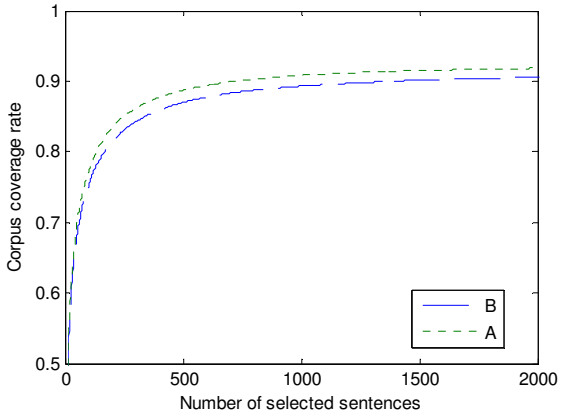

Figure 3. Corpus coverage rate (CCR) when selecting texts from different sets (A is unfiltered set, B is filtered set).

4.4. Reading script selection

To test whether our method, which filters out sentences containing infrequent words or bigrams, can generate recording scripts with reasonable coverage of text corpus, we did a compassion test for text selection on two different candidate sentence sets. Candidate sets A and B are both selected sentence sets from the Gigaword corpus with lengths 8 to 16 words. The only difference between them is that we did not filter the sentences with infrequent words and bigrams when generating A, while, for sentence set B, we only selected sentences

containing the most frequent 20,000 words and the first 1,000,000 bigrams. By selecting 2000 sentences with the same selection method (MFF method), we then compare the result.

We first calculates TCR for the 2000 sentences generated from the two data sets, and found for data B, the selected scripts can cover 35.7% of unique tokens, while for data A, the coverage is 26.7%. Then we calculate the change of CCR with the increase of selected sentences. The result is as shown in Figure 3. From the figure, we can see that when selecting 2000 sentences, the percentage of coverage of the selected set from the filtered set is very close to that of the unfiltered set (91.8% for A and 90.6% for B). Thus, we can conclude that, although the selection from filtered text will cover less unique units, it does not affect the corpus coverage of the language very much.

Table3. FRES readability scores

Candidate text set	FRES score	FKGL grade level
A	35	12
B	48	10

We also calculated the FRES score and FKGL grade level for the text scripts selected from the both candidate text sets, and the result is as shown in Table 3. From the table, we can see that the text generated from A received a score of 35, which is suitable for grade 12 students, while text from B received a score of 48, which is easier and suitable for grade 10 students. FRES and FKGL are based on number of words and number of syllables in the text. In our work, we use the same measure to control the number of words per sentence for both the sentence sets, and did not intentionally control the number of syllables in the words.

Therefore the use of the filtering methods is able to generate text with good coverage of units in the language with better readability at the same time.

5 Conclusions

In this work, we proposed a method to select easy to read text for TTS database recording. The English Gigaword corpus is used as our raw material for analysis and selection. By using word frequency and word bigram frequency as the candidate set selection criteria, we are able to generate nice text sentences that are grammatically correct and with good readability. From the candidate set, a greedy algorithm is used to select the recording script. The experiment shows that the selected recording scripts with our method have a similar coverage of the language and better readability compared with normal approach.

6 References

A Hunt, A W Black, “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database”, Proc. of ICASSP 1996.

T Yoshimura, K Tokuda, T Masuko, T Kobayashi, T Kitamura, “Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based Speech Synthesis, Proc. of Eurospeech 1999.

J P H van Santen, “Methods for Optimal Text Selection”, Proc of Eurospeech 1997.

A W Black and K A. Lenzo, "Building Synthetic Voices", Carnegie Mellon University, 2007

H François, O Boëffard; "Design of an Optimal Continuous Speech Database for Text-to-Speech Synthesis Considered as a Set Covering Problem", Eurospeech 2001.

J Kominek and A W Black, "CMU Arctic Databases for Speech Synthesis", Carnegie Melon University, 2003.

B Bozkurt, O Ozturk, T Dutoit, "Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection", Proc. of Eurospeech 2003.

T Lambert, A Breen, "A Database Design for a TTS Synthesis System Using Lexical Diphones", Proc of ISCSLP 2004.

M Isogai, H Mizuno, K Mano, "Recording Script Design for Corpus-Based TTS System Based on Coverage of Various Phonetic Elements", ICASSP 2005.

S Fitt and S Isard, "Synthesis of regional English using a keyword lexicon," in Proc. Eurospeech 1999.

R Parker, et al., English Gigaword Fourth Edition, Linguistic Data Consortium, Philadelphia, 2009.

R Flesch, "A new readability yardstick", Journal of Applied Psychology, Vol. 32, pp. 221–233，1948.

J P Kincaid, et al., "Derivation of new readability formulas for Navy enlisted personnel", Research Branch Report 8-75, Millington, TN: Naval Technical Training, US Naval Air Station, Memphis, TN, 1975.