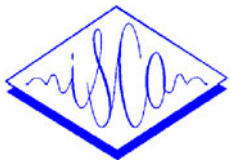
ISCA Archive
http://www.isca-speech.org/archive



7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Considering Readability in Text-to-Speech Recording Script Design

Minghui Dong, Ling Cen, Paul Chan, Haizhou Li

Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore 138632
{mhdong, lcen, ychan, 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. With such methods, sentences with difficult words or incorrect grammar are often selected. It is difficult for speakers to read these sentences correctly and naturally. Also, the selected sentences may not be suitable for child speakers or non-native speakers. In order to address these problems, we propose to consider readability in text selection. The experiment shows that the selected scripts with the proposed method have good unit coverage of the language and good readability.

Index Terms: Text-to-speech, recording scripts, text selection, 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 [1] and the HMM based statistical method [2]. 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 often carefully designed for the largest coverage of the relevant phonetic and prosodic phenomena with the least recording data.

For both of the TTS methods mentioned, in order to cover most of the 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 [3-9] 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. The problem becomes worse if we plan to record children voices or non-native speaker's voices. In such cases, we would like the selected text is simple enough for the speakers to read. 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 define some of the materials and standards used in this work. 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 measures 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, we take the phone to be the basic unit of speech concatenation. In the phonetic aspect, context is one of the important factors in co-articulation in TTS system. As phonetic contexts in a syllable are more coherent than those between syllables, we use the syllable as our selection element. In the prosodic aspect, accent is most important and is easy to determine. Therefore, we further include accent into our consideration and use syllable with accent marks as our unit of 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 [10] 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

The statistical information required is derived from the LDC English Gigaword Corpus [11] 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 part of the 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, $y_1, y_2, ..., y_m$ 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.[12][13]). Here we briefly describe two widely used measures.

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 materials that are easier to read.

Flesch–Kincaid Grade Level (FKGL): FKGL translates FRES scores 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. TKGL grade level is calculated as follows [13]:

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

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

In this work, we will try to design text scripts that are suitable for children or non-native speakers. Therefore, we need to filter out texts with low readability.

3.1. Preprocessing of corpus

The first step of the text processing is to identify the text sentences. Each sentence is usually ended with a period, a question mark or an exclamation mark. However, the period is not only used for marking the ends of sentences. 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. Selecting initial candidate sentence set

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

$$w_l \le w \le w_u, \tag{5}$$

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 read. On the other hand, it is also not efficient to record very short sentences for speech data collection task.

(2) Sentences with words that are not found in the lexicon are excluded. Recording out-of-vocabulary words make it difficult to create phonetic transcription in later stage.

(3) Sentences with less frequently used words are excluded. Less frequently used words are usually found to be more difficult. This ensures the speaker is able to read each word correctly and easily.

(4) Sentences with less frequent word bigrams are excluded. Such sentences are more likely to be ungrammatical sentences, which are not easy to read. It is therefore necessary to exclude such sentences.

(5) Finally, we will also exclude the sentences with high FKGL grade levels, which is an indicator of text difficulty.

Depending on the level of proficiency of the speaker in the language, we may need to filter out the difficult sentences from the initial candidate set. In our work, we calculate the FKGL grade level for each sentence. TKGL grade level for an individual sentence is calculated by

$$L(S) = 0.39 N_w(S) + 11.8 \frac{N_l(S)}{N_w(S)} - 15.59 \qquad (6)$$

where $N_w(S)$ and $N_l(S)$ are number of words and syllables in sentence S respectively.

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 iteration, the sentence that covers most uncovered elements 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, whilst assuming that the more frequent ones will usually be 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

We first conducted an analysis on the 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)[11]. 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 words in 7.15 million documents. The documents are classified into the categories 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 altogether 2,288,791 unique words in the corpus. We sorted the words in descending order of frequency 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 less rapidly as the number of words exceeds 10,000. From our calculations, the most frequent 20,000 unique words cover 95.56% of the word occurrences in corpus.

We also calculated the frequency of word bigrams. In this calculation, we group infrequently used words (defined to be those with frequency ranks more than 20000) into one single category.

We also sorted the bigrams in descending order of frequency, and calculated the accumulative coverage percentage. The result is as shown in Figure 2. From the figure, we can see that most frequent unique 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 unique syllable items cover about 98.16% of all the syllable occurrences in the corpus.

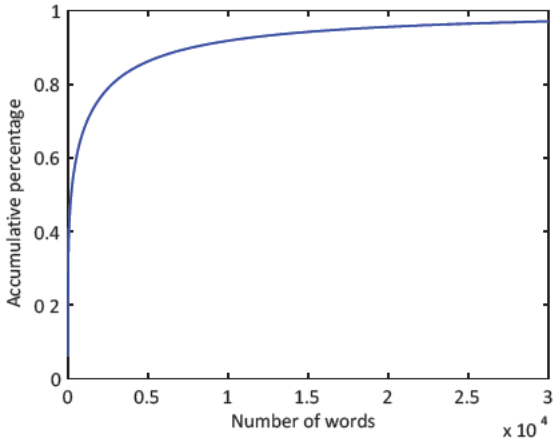


Figure 1. Accumulative percentage of words in the corpus.

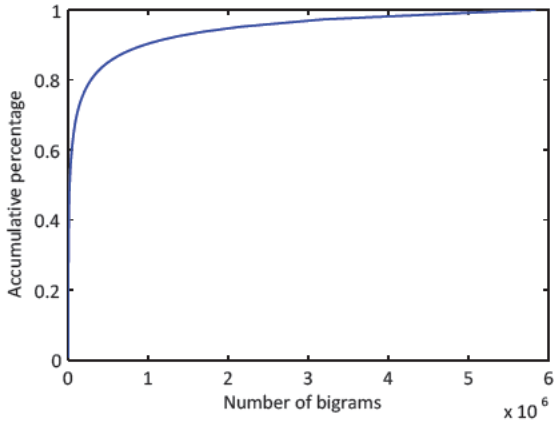


Figure 2. Accumulative percentage of word bigrams in the corpus.

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 while selecting those containing most frequent 10,000 words and the first 500,000 bigrams. We selected 619,888 sentences altogether as our initial candidate set.

We randomly sampled the several hundred sentences, asked three non-native English readers to randomly inspect the selected sentences, and found none of them to be grammatically incorrect. They have also found no unknown words in the text. On the contrary, the original texts contained numerous incomplete sentences, foreign names, names of places, email addresses, internet links, DNA sequences, computer commands, uncommon symbols, etc. Thus, it has been shown that filtering the sentences with infrequent words and infrequent bigrams help to generate text with higher readability for non-native English readers.

4.4. Selecting the recording script

As we wish to select sentences that are easy to read for children or non-native English speakers, we first filter out candidate sentences that are not suitable for students below a particular grade. We have listed in the appendix sample candidate sentences for each grade.

Here, we try to select sentence sets that are suitable for grades 3, 5 and 8. We calculated the TKGL grade level of the sentences. The sentences whose levels are higher than the target were filtered out. In each case, we selected 2000 sentences using the same selection method (MFF method).

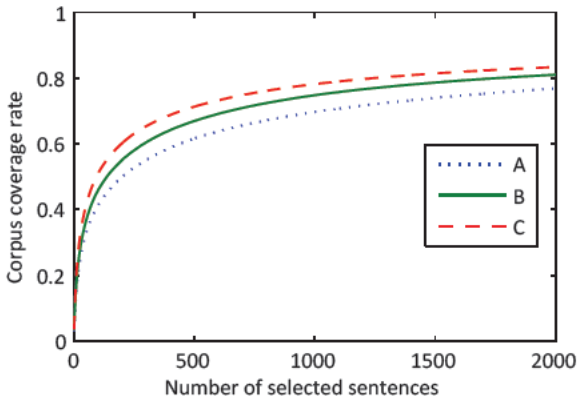


Figure 3. Corpus coverage rate (CCR) of selected text scripts for students of grades 3, 5 and 8 (curves A, B and C respectively)

We calculate the change of CCR with the increase of selected sentences. The result is as shown in Figure 3. In the figure, curves A, B and C indicate the CCR rates for text selection for grades 3, 5 and 8 respectively. From the figure, we can see that we can achieve fairly high CCR rates (76.8%, 81.0%, and 83.3% respectively) with 2000 selected sentences. This shows that, by filtering the sentence set based on the readability grade level, we can select easy-to-read text scripts that cover the language well.

5. Conclusion

In this work, we proposed a method to select easy–to-read text for TTS database recording. This method can be used to select text scripts that are suitable for children or non-native speakers. Here, the English Gigaword corpus is used as our raw material for analysis and selection. By using word and word bigram frequencies as the candidate set selection criteria, we are able to generate easy–to-read text sentences that are grammatically correct. By filtering the candidate set with readability measure, we are able to generate candidate sets suitable for speakers with different levels of language ability. From the filtered 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 good coverage of the language and better readability.

6. References

[1] A Hunt, A W Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", Proc. of ICASSP 1996.

[2] 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.

[3] J P H van Santen, "Methods for Optimal Text Selection", Proc of Eurospeech 1997.

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

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

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

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

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

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

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

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

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

[13] 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.

Appendix

Sample Candidate Sentences for Grades 1 to 10:

Grade 1

- A bad thing done for a good cause is still a bad thing.
- But I don't think he could do now what he did then.
- He let go of the bat at the end of his swing.
- I know I'm young and I have a lot to learn.
- It is not there yet but it will start with the young.

Grade 2

- I see each of them as a sort of love story.
- And the last few years we would talk to each other some.
- But after a few days, he had had a chance to think.
- He had used some of his ill-gotten gains to buy a farm.
- He tries to set aside a half hour each day to do so.

Grade 3

- A few homes are visible through the trees, but just a few.
- And they won't be there to make threats or tear up their tickets.
- At least five homes have been destroyed by fire in the area.
- But for me, as I suspect for my children, there's camp.
- Do you think there is a sea change under way in the market?

Grade 4

- I wonder how I got to it, but not why I'm doing it.
- And her old friends, on China's team here, are a little surprised.
- And the answer of course is to do something on special teams.
- As far as that goes, I don't really look at it any more than that.
- Bush at Camp David in June as well as the Mideast crisis.

Grade 5

- A man, woman and child were injured in the blast, the sources added.
- After the way this year has gone I am delighted to win again.
- And I understand what's going on out there for a lot of the guys.
- Before long, he was paying for music lessons for all of the girls.
- But everybody in the game did what they were supposed to do.

Grade 6

- A ninth person was treated and released from the hospital, she said.
- After that, he said, it was a case of getting the runs as quickly as possible.
- And I decided to live up north of San Francisco and drive down once a week.
- And these are the best of times for America's government, in at least three ways.
- At least it appears that peace has come to the Democratic Party.

Grade 7

- A large part of coaching and playing is the ability to adjust.
- After a six-week trial, a federal jury ruled against the women.
- And it threatens to weaken the party in a key state for the general election.
- As is often the case in technology, the answer is: It depends.
- Bush on his re-election, the government said Friday in a press release.

Grade 8

- A new management team was ordered to slash the budget in return for more government money.
- Brazil were only able to clinch a place in their final qualifying match.
- But he also said recently that he had no intention of sending troops to Iraq.
- No injuries were reported, and several streets in the downtown were closed, the radio said.
- On Tuesday, the committee was given access to some of the documents.

Grade 9

- A tropical wave will continue to move across the area this afternoon and tonight.
- And no decision in politics is final until there's an announcement," he said.
- Birth rates in Japan and the European Union are well below replacement level.
- He insisted that he was innocent, but resigned last week amid mounting pressure.
- It noted that recovery in the United States also appears to be on track.

Grade 10

- But the Italian and French said Monday that Switzerland had failed to provide any information.
- For now, it remains a part of Serbia-Montenegro, the successor state of Yugoslavia.
- He had an operation and the recovery time has been longer than expected.
- Iraq has the world's second-largest proven oil reserves after Saudi Arabia.
- The biggest problem, according to the players, is their inability to stick to Murray's system.