Speech Database Compacted for an Embedded Mandarin TTS System

Qing Guo, Bin Wang
Internet Application Laboratory
Fujitsu Research and Develop Center China
Beijing, P.R.C
guoqing@cn.fujitsu.com

Nobuyuki Katae
Audio-Visual Systems Laboratory
Fujitsu Laboratories Ltd.
Akashi, Japan
katae@jp.fujitsu.com

Abstract—In recent years, the unit selection based concatenative speech synthesis system that uses large speech database has become popular because it can produce high quality synthesized speech. However, using such a large speech database is not practical for many applications such as those ported on embedded devices with the storage requirement and the computational complexity involved in searching it. In this paper, it proposed the context based pruning algorithm and waveform adjustment effect based pruning algorithm to compact the speech database. At last, it presents experimental results and discussion.

Keywords—text-to-speech; segment pruning; prosody structure; unit selection

I. INTRODUCTION

A Text-To-Speech (TTS) synthesizer is a computer-based system that can read text aloud automatically, regardless whether the text is introduced by computer input stream or a scanned input submitted to an optical character recognition (OCR) engine. TTS synthesis can be used in many areas, such as telecommunication services, language education, vocal monitoring, multimedia, and it can also be used as an aid to handicapped people. Impressive progress has been made over the last couple of decades in Mandarin TTS research. A lot of high quality Mandarin TTS systems have become mature in commercial use.

Concatenative speech synthesis that uses large speech database has become popular in recent years due to their improved sensitivity to unit context over their simpler predecessors. These systems usually make use of large speech databases and employ sophisticated search algorithms to determine the optimal unit sequence used to synthesize each sentence.

During synthesis, the input text is converted into a phonetic string with predicted prosodic targets. The synthesis system then searches for speech segments that are close to the context of the target phonetic string and its predicted prosodic targets. At last, a pitch and duration modification algorithm, such as PSOLA, is applied to pre-stored units to guarantee that the prosodic features of synthetic speech meet the predicted target values and the transformed waveforms are concatenated together. Some Mandarin TTS systems adopt two-module TTS structure[1][2]. These systems bypass the prosody module because there is no pitch or duration modification will be applied to the selected units before concatenation.

Since comprehensive linguistic and acoustic phenomena coverage is essential for a high quality concatenative speech synthesis system that the greedy algorithm[3] is often adopted for speech corpus designing to cover more phonetic context and prosodic context phenomena in a recording corpus of a given size. Generally speaking, it’s a latest trend that the size of the speech databases recorded in Mandarin TTS systems becomes bigger and bigger.

The size of these speech databases may be as large as several hundred megabytes or even several gigabytes. However, using such a large speech database is not practical for many applications such as those applications ported on embedded devices with the storage requirement and the computational complexity involved in searching it. Thereafter, it is necessary to compact the speech database by some off-line segment pruning algorithms to determine which subset of the database can enable the minimum degradation in synthesis quality to be performed for a given runtime system size.

To improve the efficiency of the speech synthesis systems and maintain the quality of the synthesized speech at the same time, some researchers use some segment pre-selection algorithms to specify a small set of candidate segments from the entire speech database for a target segment. In fact, these segment pruning algorithms and segment pre-selection algorithms often use the same techniques.

Some previous systems reported in the publications have used simple heuristics to select a single version of each unit[4][5], while others have used more complex procedures to select multiple versions[6-10].

Black A.W.[6] adopted a clustering algorithm to cluster units within a unit type based on questions concerning prosodic and phonetic context with an acoustic measure, which uses an acoustic vector comprising of Mel frequency cepstrum coefficients, F0, power and delta cepstrum, F0, power. Donovan R.E.[8] defined a usefulness measurement for each segment. Only those most useful segments are saved for the synthesizer at runtime. Hamza W.[9] proposed a data-driven algorithm to reduce the database size used in concatenative synthesis. The algorithm retains speech segments based on statistics collected by synthesizing a large number of sentences.
using the full speech database. Ling Z.H.[10] introduced an
unit pre-selection method by classification and regression tree
(CART) for Mandarin synthesis system to improve its
efficiency.

Generally speaking, segment pruning algorithm or segment
pre-selection algorithm has two problems need solving. One is
how to measure the difference between two segments in the
clustering process. The other is how to select a representative
segment for each cluster.

This paper is organized as follows. Section 2 introduces the
FUJITSU Mandarin TTS system briefly, which is a state-of-
the-art unit selection based concatenative speech synthesis
system[11]. Section 3 describes two segment pruning
algorithms that were used to compact the speech database for
embedded TTS system. Section 4 provides the experimental
results of listening test. Finally, it makes a conclusion in
Section 5.

II. SYNTHESIS SYSTEM OVERVIEW

A. Basic Synthesis Unit

In the Fujitsu Mandarin TTS system, syllables are the basic
synthesis units in the unit selection model. However, in order
to avoid serious degradation of synthesized speech quality,
initials and finals are processed respectively with a PSOLA
algorithm to have the speech prosody and concatenating speech
waveforms modified.

There are about 205 syllable initials and finals, 1,600 tone
syllables in Mandarin. Using syllable initials and finals as basic
units can improve the robustness of a statistic decision tree
model. 182 finals and 23 initials have been defined in our
system. In addition, some high frequency retroflex finals are
also been considered.

In our study, syllable initials and finals (with tone) are
defined as the basic synthesis units.

B. Corpus Design

Speech corpus design is critical in building high quality text
for speech synthesis systems. Usually, reading speech is
adopted for it seems to be the easiest way to obtain a recorded
speech corpus with highest control of the content. Generally, some high frequency retroflex finals are
also been considered.

In our study, syllable initials and finals (with tone) are
defined as the basic synthesis units. In addition, some high frequency retroflex finals are
also been considered.

In our study, syllable initials and finals (with tone) are
defined as the basic synthesis units.

C. Prosodic Structure Labeling

The prosody structure is composed of four tiers[12]:
prosodic word (PW), minor phrase (MIP), major phrase (MAP)
and intonation group (IG). Prosodic word is a tone group
bearing one word stress. Minor phrase contains one or more
prosodic words, bears one phrasal stress and the perceived
break between MIPs is longer than that between PWs. Major phrase contains one or more PWs, bears one phrasal stress and the perceived
break between MAPs is longer than that between MIPs. The criterion for prosody structure labeling is listening
perception. Major phrases are often marked by commas with
incomplete pitch resetting while intonation groups are marked
by periods, quotation marks or semicolons with full pitch
resetting.

Additionally, three levels of stress have been defined,
namely the stressed, the normal and the neutralized.

The following is a sample transcription of a certain
sentence in the speech corpus. “/”, “|”, “|” and “@” represent
PW, MIP, MAP and IG in the transcription respectively. A
syllable marked with “_H” means that it is a stressed syllable,
and a syllable marked with “_L” means that it is a neutralized
one.

III. THE PRUNING ALGORITHMS

The performance of a pruning algorithm depends on the
infrastructure of a TTS system itself. Generally, a TTS system
contains three modules; they are the text analysis module, the
prosody generation module and the back-end module. The
back-end module is responsible for unit selection and speech
synthesis. To ensure that the unit selection algorithm can find
the most appropriate candidate, a speech database must be
diverse enough. The diversity of the speech database means
that the syllables in the speech database have many samples with
different prosody parameter and acoustic parameter.

To implement an embedded TTS system used for those
applications ported on embedded devices, two pruning
algorithms are designed to compact the entire speech database
in this paper. These two pruning algorithms, context based
pruning algorithm and waveform adjustment effect based
pruning algorithm, are described in this section respectively.

A. Context based Pruning Algorithm

A pruning algorithm should have two effects. One is that it
can remove spurious atypical units which may have been
caused by mislabeling or poor articulation in the original
recording. The other one is that it can remove those units which are
so common that there is no significant distinction between
candidates.
As for the second type of pruning, removing those overly common units, three factors should be considered. Firstly, the diversity of prosody among segment samples should be maintained. Secondly, the diversity of spectrum among segment samples should be maintained. Thirdly, the diversity of context should be covered as much as possible.

Figure 1 gives the framework of the first pruning algorithm implemented in our Mandarin TTS system.

In step 1, two clustering algorithms are implemented, which adopt MFCC parameter and pitch contour to measure the difference between two segments in the clustering process respectively. As we know, the most important indicator of prosody information is pitch contour in Mandarin. The variety of pitch range and pitch pattern is the basis of the continuity of synthesized speech because that too much pitch conversions resulting in bad synthesis voice quality. Therefore, we used pitch contour based clustering algorithm in the listening test, which will be mentioned in the next section. Hence, the syllables which have similar pitch contours are put together in a same cluster. As a result, syllables with variety of pitch contours will be reserved in the compacted speech database.

In addition, considering that the diversity in prosody and acoustic of a syllable is caused by its different context in the speech, a context coverage strategy is used to ensure that those syllables with high frequency context types are reserved in the step 1. The context type of a syllable in the speech database is defined by six factors as following: The category of the previous syllable; The category of the following syllable; The tone of the previous syllable; The tone of the following syllable; Syllable position in the prosodic word that the syllable belongs to; Prosodic word location in the prosodic phrase that the syllable belongs to.

In step 2, a stability pruning strategy is used for furthermore compacting. Using the predicted pitch contour and the predicted duration of a syllable as a reference, a stability measurement can be scored for all the reserved syllables in the step 1. The context coverage strategy can be used in step 2 too. However, the dimension of the context type of a syllable is usually smaller than that used in step 1.

B. Waveform Adjustment Effect based Pruning Algorithm

The second pruning algorithm is a clustering algorithm also based on pitch contour. As we know, a pitch and duration modification algorithm, such as PSOLA, is applied to pre-stored units to guarantee that the prosodic features of synthetic speech meet the predicted target values. Hence, a new representative unit selecting algorithm by considering the effect of waveform adjustment is used in our second pruning algorithm.

Firstly, all samples of a certain syllable are clustered together. To cluster these candidates, it uses their pitch contours to measure the distance between two samples of a certain syllable.

Then, a representative waveform is selected from each cluster. In order to find a good representative unit, a new measurement method is defined to find the most stable unit in the cluster.

Considering that a pitch and duration modification algorithm will be applied to the selected representative unit to satisfy the predicted target pitch and duration, a stable metric of any unit in a cluster is defined as follows:

Let us denote the waveform of the unit A as Waveform-A. Before calculating the distance between the unit and any other unit (herein it refers to as target unit, Waveform-B) in the cluster, Waveform-A is converted into the target pitch contour and duration of the target waveform at first. The converted waveform is denoted as Waveform-A’. Then, calculate the low power spectrum distance between Waveform-A’ and Waveform-B. The sum of all distances between the unit A and other units in the same cluster is used as the stable metric of the unit A. At last, the unit with lowest stable metric score will be selected as the representative unit of each cluster.

IV. EXPERIMENTAL RESULTS

Firstly, in order to evaluate the effectiveness of our two pruning algorithms, we conducted paired comparing, in which synthesized speech using the compacted speech databases by two pruning algorithm were compared.

20 sentences were used in the pruning algorithms evaluation. 30 people were asked to make comment on each pair of waveforms, A and B. There were 5 mutually exclusive choices for answers, namely: A is better; A is a little better; equal; B is a little better; and, B is better.

<table>
<thead>
<tr>
<th>Wave DB</th>
<th>Averaged Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>200M (Method 1) vs. 200M (Method 2)</td>
<td>2.8 vs. 3.2</td>
</tr>
<tr>
<td>50M (Method 1) vs. 50M (Method 2)</td>
<td>2.9 vs. 3.1</td>
</tr>
</tbody>
</table>

Table 1 gives the results of the pruning algorithms evaluation. To evaluate which pruning algorithm performs well in the evaluation, we marked the above five answers as 5, 4, 3, 2 and 1 respectively. Then, the averaged score was used as the criteria to judge which pruning algorithm is better.
Experimental results show that the second pruning algorithm is a little better than the first one. Hence, in the next turn of listening tests, we’ll use the second pruning algorithm to compact the speech database.

Then we conducted another listening test to evaluate the degradation category rating (DCR) of synthesized speech according to different speech database compacting ratios.

Still 20 sentences were used in the DCR evaluation. 30 people were asked to make comment on each four synthesized waveforms using 1.8G wave DB (the full set of our speech database), 200M wave DB, 50M wave DB and 2.6M wave DB comparing with natural speech of these 20 sentences.

The five answering standards were prepared in the DCR evaluation as following:

5: Inaudible (can not percept the difference between the two data);
4: Audible, but not annoying;
3: Slightly annoying;
2: Annoying;
1: Very annoying.

The listeners made comment on each speech datum. The DCR score was the average score of all listeners.

Remarks. The manual transcribed front-end (linguistic processing) results of 20 sentences were used in both two listening tests.

Figure 2 gives the results of the DCR results of 1.8G, 200M, 50M, 2.6M wave DB respectively.

V. CONCLUSIONS

In this paper, it proposes to use context based pruning algorithm and waveform adjustment effect based pruning algorithm to compact the speech database. The first method emphasized on the coverage of high frequency context types during the compacting process. The second one defined a stable measurement to characterize the effect of waveform adjustment during synthesis of a waveform unit. The most stable unit was selected as representative unit from each cluster to expect that the minimum waveform adjustment degradation was obtained during synthesis. Lastly, it carried out two paired comparing listening tests to evaluate the effectiveness of two pruning algorithms and the degradation category rating (DCR) of synthesized speech according to different speech database compacting ratios.

VI. ACKNOWLEDGEMENT

The first pruning algorithm described in this paper is developed on the basis of the research work of a collaborative project with Institute of Automation, Chinese Academy of Sciences.

REFERENCES