TOSHIBA ENGLISH TEXT-TO-SPEECH SYNTHESIZER (TESS)

Chang K. Suh, Takehiko Kagoshima, Masahiro Morita, Shigenobu Seto, and Masami Akamine
(chang@alum.mit.edu; kagoshima, morita, seto, akamine@krl.toshiba.co.jp)

Kansai Research Laboratories, Toshiba Corporation
8-6-26, Motoyama-minami-machi, Higashi-nada-ku, Kobe, 658-0015, Japan

ABSTRACT

Toshiba English Text-to-Speech Synthesizer utilizes several new techniques to produce synthesized speech that is more natural-sounding and intelligible than that created by conventional synthesizers. The closed-loop training method creates synthesis units that most closely resemble the training data and are the least susceptible to prosodic distortion noise by analytically solving an equation that minimizes distortion between target units and training data. The pitch contour model creates a codebook of representative word-based F0 contours by first clustering the training data using word stress and syllable numbers. Within each cluster, the training data is divided into different groups using lexical and phonological attributes of each word. In each group, a representative contour is created using approximate error estimation. The resulting approximate errors are used in offset level prediction for each contour. These techniques have significantly improved the prosodic quality, naturalness and intelligibility of the resulting synthesized speech.

1. INTRODUCTION

Many text-to-speech (TTS) systems today have improved the intelligibility of synthesized speech, but still suffer from unnatural prosody and prosodic modification artifacts. The synthesized speech created by these systems is, though intelligible, quite robotic in intonation, and lacks human voice qualities. These problems are mainly attributable to two aspects of TTS systems: creating and concatenating synthesis units, and generating pitch contours for a sentence.

To create synthesis units, many systems select an optimal unit from training data using a context-oriented clustering method (COC) [1] or decision-tree-based clustering of context-dependent phonetic units [2]. In these methods, distortion within each cluster is minimized, but the synthesis units tend to suffer from prosodic modification that results in distortion and noise in synthesized speech.

To generate pitch contours, many systems use linguistic rules to define prosodic parameters dynamically [3], use stochastic learning techniques such as decision trees [4], or select an optimal pattern from a table based on phonological, grammatical and contextual properties [2]. Even though these methods have significantly improved the quality of prosodic patterns, the resulting synthesized speech still tends to sound robotic and lacks resemblance to the prosodic qualities of the original speaker.

Toshiba English Text-to-Speech Synthesizer (TESS) aims to remedy these problems by employing two new techniques: the closed-loop training (CLT) method for creating synthesis units and a codebook-based pitch contour model for generating F0 contours.

The CLT method [5] improves the prosodic quality of synthesized speech by minimizing distortion due to prosodic modification. Its most noticeable difference from conventional methods is that, instead of selecting a synthesis unit from training data, it analytically creates an optimal unit by solving an equation, which is a summation of distortion between the target unit and the training data. By differentiating the equation and setting it to zero, the method creates an optimal unit from scratch. The method allows the synthesized speech to retain the original speaker’s voice qualities with non-robotic naturalness and minimal prosodic distortion.

The codebook-based pitch contour model [6] differs from other conventional methods in several key areas. First, the model does not generate pitch patterns dynamically, but rather selects a contour from a codebook of representative pitch contours derived from corpus using Quantification Method Type I (QMTI) [7], a “sums-of-products” method. Second, it trains and predicts a pitch contour and its offset level on the frequency axis independently from each other. The resulting
2. SYSTEM OVERVIEW

TESS is a complete text-to-speech system that uses a diphone-based synthesizer, which generates synthesized speech using PSOLA and LPC. It consists of several modules, as shown in Figure 1.

The prosody generation module consists of F0 contour model, duration prediction, phrase break prediction and pitch accent prediction. The common theme among these modules is that they are all totally data- and speaker-driven. All these modules can be divided into offline training and online prediction. The offline training uses C4.5 decision trees [9] and QMTI to create prediction models from the corpus, and in turn, the prediction models are used to generate prosody during the synthesis phase.

The corpus currently contains about 1,200 sentences with around 14,000 words in approximately 100 minutes of speech narrated in neutral-declarative tone. The system employs 44 phonemes, and about 1,200 diphones are created from the corpus at the moment. The corpus was labeled with an in-house HMM-based automatic labeling tool, and F0 information was automatically obtained.

The system employs a pronunciation dictionary that currently contains about 200,000 entries. The inflections are processed by deriving the pronunciation from the root word found in the dictionary. The use of a pronunciation dictionary limits the number of words the system can synthesize, but allows more intelligible and natural pronunciations in synthesized speech.

3. SYNTHESIS UNITS

In [5], the CLT method was applied to CV/VC-type synthesis units for the Toshiba's Japanese TTS system, and was shown to be effective in generating minimal-distortion synthesis units. TESS provides the first opportunity for the CLT method to be used in generating English synthesis units, and the results have been very favorable. The implementation of the CLT method has necessarily been altered due to linguistic differences between English and Japanese, and may account for discrepancies between the performances of TESS and the Japanese system.

In principle, TESS uses the analytical CLT method introduced in [5]. The process of generating synthesis units is shown in Figure 2. First, the pitch period and duration of synthesis units are modified to match those of the training data. After performing distance calculation and clustering, an analytic equation is generated. The equation is a summation of squared errors between the training data and the synthesis units in the cluster, which can be written as the following:

\[
E_i = \sum_{r_j \in G_i} (r_j - g_{i,j}A_{i,j}u_j)^T (r_j - g_{i,j}A_{i,j}u_j), \tag{1}
\]

where \(G_i\) is the cluster set, \(r_j\) is the training vector, \(u_i\) is the synthesis unit vector, \(g_{i,j}\) is the gain vector that adjusts the signal levels between \(r_j\) and \(u_i\), and \(A_{i,j}\) is a matrix operator that modifies the pitch period and duration of \(u_i\). By taking a partial derivative of Eq. (1) with respect to \(u_i\) and setting it to 0, the optimal synthesis unit can be obtained.

4. F0 CONTOUR PREDICTION

The F0 contour model in TESS recognizes that, at word level, F0 contours can be effectively expressed with a limited number of representative vectors that are selected based on lexical, grammatical and phonological attributes of the training data. The codebook-based F0 control module used in TESS has already been proposed and successfully applied to the Toshiba's Japanese text-to-speech system [6]. The module has been adapted to TESS and undergoing various tweaks, and it shows a remarkable performance at the moment. The predicted pitch contours have a notable similarity to pitch contours of an open data sentence spoken by the same speaker. Such a performance is partly expected by the notion that there exist common word-level patterns in English [8].

In principle, TESS uses the codebook-based pitch contour model that has been proposed in [6], but many basic ideas and implementation details have been reconsidered to accommodate linguistic differences between English and Japanese.

First of all, two languages belong to two different categories of language. The Japanese language is considered a mora-timed language while English is considered a stress-timed language. This fundamental linguistic difference makes such a basic idea as the length of a word change inevitably. In Japanese, the timing and location of falling in intonation is considered important in generating a natural-sounding pitch pattern. To correctly designate the timing of such a falling, the length of a word is expressed in terms of morae, each of...
When the economy slows down, which is a CV pair of Japanese characters. Therefore, the vowels always fall at the latter part of each mora, and a falling in intonation always align with the correct mora.

In English, on the other hand, the location of vowels is not fixed at the same place in every word and is not as easy to anticipate. In a stress-timed language like English, the location of a peak in a pitch pattern needs to align with a vowel to generate natural-sounding intonation. In order to achieve such an alignment, the length of a word in TESS is described in terms of the number of vowels. Because each vowel becomes an anchor point in each unit length, the peak of a pitch pattern always correctly aligns with a stressed vowel, during both the analysis phase and the synthesis phase.

Another important difference in implementation lies in the method of clustering. In the Japanese system, the stress of a word is the sole criterion for clustering the training data because only overall shapes of F0 contours are considered important. However, in TESS, the clustering is based on the word stress and its vowel numbers. The pitch contours in English tend to be more dynamically changing and contain more frequent occurrences of unvoiced consonants, which introduce breaks in pitch contours. The additional criterion enables the system to model the pitch contours in more detail and provides more accuracy when choosing a representative vector during the synthesis phase.

During the analysis phase, the codebook is created using QMTI according to grammatical, phonological, and lexical attributes of words, such as the parts of speech of current, previous and following words, the phoneme structure of the word, its accent pitch, vowel class, and the stress and pseudo-syllable numbers of previous and next words. The effects of different onset classes [10] have been considered as well.

5. OTHER PROSODY GENERATION MODULES

5.1. Phrase Break Prediction

The phrase break prediction is carried out using C4.5 and QMTI in conjunction. The C4.5 decision tree predicts whether a pause needs to be inserted after a particular word, and if so, QMTI determines its length.

Both training models use the same set of word attributes, such as the parts of speech of the current word, previous words and following words, the role they play in a sentence, the distance to the word that current word is related to, and the presence of punctuation marks.

The C4.5 tree is created by using every word in the corpus as training data, which amounts to around 14,000 words. The resulting C4.5 tree, after pruning, currently has the error rate of 8.7% within the training data and an estimated error rate of 12.1% against unknown data. The cross-validation on ten randomly segmented training data showed the error rate of 12.3%. The QMTI is carried out only on the words that are followed by a phrase break.

5.2. Pitch Accent Prediction

The pitch prediction model assigns one of three tags (high, acc and deacc) to each word. The high tag represents a word in a sentence that has the strongest emphasis and highest F0. The high tag tends to be located towards the beginning of a sentence. The acc tag represents words that are read more emphatically than others do. The deacc tag represents words that are read without any emphasis.

The model uses several grammatical, phonological and lexical attributes of each word to construct a C4.5 tree. Currently, it considers attributes such as a word's position within a sentence, its part of speech and those of previous and next words, the distance to its related word, the distance to the next phrase break, etc. The training data has been manually tagged.

The C4.5 tree, after pruning, currently has the error rate of 14.2% within the training data and an estimated error rate of 18.7% against unknown data. The cross-validation on ten randomly segmented training data showed the error rate of 18.8%.

Figure 3. The F0 contour of s007-02.wav: (a) generated by TESS, (b) original.
5.3. Duration Prediction
The duration prediction is carried using QMTI. First, all the phonemes in the training data are clustered separately. Within each cluster, the QMTI is performed based on several grammatical, phonological, and lexical attributes, such as phoneme types and classes of the current, previous and following phonemes, the part of speech, the locations of phrase breaks, etc. Currently, the model uses raw duration without applying any transformation functions.

While some phonemes have a better model than others do, the overall average error rate from the QMTI training is 22 ms.

6. RESULTS
In order to demonstrate the quality of synthesis units and the naturalness, intelligibility and human qualities of synthesized speech of TESS, a few open-data sentences have been synthesized and included on the CD-ROM. The sentences and their corresponding filenames are the following:

- He said he was forced to abandon his own vehicle when water rose past the windows. (s007-01.wav)
- When the economy slows down, the earnings growth will look more attractive to investors. (s007-02.wav)
- The project's key drawback is the potential for severe traffic back-ups on Memorial Drive. (s007-03.wav)

An example of an F0 contour is shown in Figure 3. The figure shows the beginning of the sentence s007-02.wav synthesized by TESS. To compare pitch patterns only, the original duration has been uses to create the figure. The synthesized speech wave file used the predicted duration. As shown in the figure, the predicted F0 contour has a reasonable similarity to the original contour.

7. CONCLUSION
Compared to robotic intonation and voice qualities of synthesized speech created by conventional synthesizers, TESS generates synthesized speech that has more human-like voice qualities and is more natural in terms of its intonation. Moreover, the resulting synthesized speech very closely resembles prosodic characteristics and voice qualities of the original speaker. Such a resemblance has been made possible by the CLT method for creating synthesis units and the codebook-based pitch contour model for generating intonation.

8. FUTURE WORK
In addition to overall quality improvements in synthesis units and the pitch control module, future work will include improving other parts of the system. Some parts of the prosody control modules, such as pitch accent prediction, phrase break prediction and duration control, have been very lightly covered because they are still ongoing research areas. Future work will be concerned with improving the error rates of phrase break prediction and pitch accent prediction modules, and the quality of duration control module.

9. REFERENCES