ABSTRACT

A new method for speech synthesis by concatenating waveforms selected from a dictionary is described. An adult male recorded a two-hour speech with acoustic phonetic labels. This data was used to construct the dictionary. The dictionary contains 35,000 waveforms which are identified by their duration, average pitch, pitch contour and average energy. The number of the phonetic labels is thirty-five. In the speech synthesis phase, given a phoneme string and prosody information, the optimum waveforms are selected by matching their attributes with the given phonetic and prosodic information. The matching score is defined as a function of phonetic coincidence and prosodic attribute differences. Selected waveforms are then concatenated to produce speech. The speech has high intelligibility and naturalness.

1. INTRODUCTION

Starting with NTT's ANSER system, which was first introduced in 1981[1], quite a variety of Japanese systems and devices that make use of synthetic speech have appeared on the market. Today, one's chances of coming into contact with synthetic sound are greater than ever and will surely continue to increase in the years ahead. Nevertheless, unconstrained voice output based on synthetic speech by rule, still leaves something to be desired in terms of quality and has failed to satisfy the high expectations of users.

In synthesis by rule, the conventional approach has been to use a parametric coding technique such as linear prediction coding (LPC) for voice synthesis. In most published reports, phonemes (CV, VCV, and others) and their corresponding coarticulation envelopes are taken as concatenation units.[2]-[5] With this method, however, there is mismatching between vocal tract spectra and sources spectra because the pitch at the time of synthesis is different from the pitch at the time of analysis. This results in abnormal amplitudes and declines in spectral Q values, all of which cause a deterioration in quality. One other problem with the LPC method is that it introduces pulses and noise in the excited source, and thus incorporates those sources of distortion as well.

A number of attempts have been made to generate clear synthetic sound from waveforms by circumventing these problems.[6]-[7] However, in methods that use waveforms, the sounds of the speech elements themselves have always been clear; the major problem has been in controlling the pitch and duration because the number of speech elements has been too small.

This paper will describe an alternative approach to obtaining high-quality synthetic sound. In this speech synthesis method, waveforms are derived from a large volume of naturally uttered individual words and sentences. Waveforms are then selected that most closely resemble the input text. Since the waveform dictionary consists of a large set of original waveforms based on many kinds of utterances, it is expected to result in a synthetic sound that is superior in terms of clarity and naturalness.

2. METHOD OUTLINE

A schematic overview of the new method is shown in Fig.1. This paper will be primarily concerned with those portions of Fig.1 pertaining to synthesis by matching (indicated by shading).

First, textual analysis is carried out to convert Kanji to Kana (i.e., Japanese ideograms to Japanese phonetic symbols), establish articulation units and to determine accent types. As a result of this analysis, it is possible to obtain phonetic symbol sequences and prosody information, such as accent types of compound names and the number of moras in breath groups. From this prosody information, various rules and tables are applied to obtain the pitch contours, time duration patterns, and amplitude patterns for each mora related to the sounds to be synthesized.

Next, waveforms that most closely match to the foregoing speech characteristics are selected from the waveform dictionary. This process is detailed in section 5. If it is necessary, further processing is applied to the selected waveforms to make them fit the conditions even more closely. Processing of pitch was fundamentally avoided and only duration and...
amplitude were adjusted, because of the difficulties of waveform pitch control.

In cases where no waveforms could be found in the dictionary that matched the conditions, waveforms were generated using an existing technique such as LPC. The waveforms thus obtained were finally concatenated to yield the desired continuous synthetic voice output.

3. LENGTH OF A WAVEFORM UNIT

Our first concern was to determine the most appropriate length unit for waveforms in the dictionary. Three possibilities were considered: pitch length, phoneme length, and syllable length. In evaluating pitch length, it would be difficult to take spectral continuity into consideration as unit selection has a major impact on sound quality. For example, Fig. 2 shows the spectral of the second /o/ in /Yokosuka/: (a) shows the continuously voiced sound, and (b) shows the voiced sound concatenated pitch-length waveforms. These waveforms are manually selected in order to create continuously voiced sound from a number of different words. Listening tests confirm that in (b) there is considerable noise in the phoneme, caused by differences in spectral contours, pitch location, and amplitude derived from the waveforms of a number of different words. This demonstrates that continuity processing based on pitch-length waveforms is quite complex. It was also determined that pitch length yielded from five times to ten times more connection points than phonemes and syllables; and since connection points multiply for vowels, this greatly increased the likelihood of noise being generated. Moreover, from a purely pragmatic point of view, the requirement for such detailed analysis to determine the points of waveform separation is not practical for constructing a pitch-length waveform dictionary.

The main objection to syllable length is that every syllable has an allophonic waveform that depends on pitch and duration. This, therefore, requires a much greater memory capacity than if phoneme length were used as a unit.

Adoption of a phoneme-based unit length preserves, at least the spectral within the phoneme. In terms of time length control, it is thought to be comparatively simple to remove pitch length waveforms from phoneme waveforms and reuse them.

For these reasons, phoneme length was adopted as the waveform unit length for the waveform dictionary.

4. CONSTRUCTION OF THE WAVEFORM DICTIONARY

A flowchart illustrating waveform dictionary construction is shown in Fig. 3. The dictionary is based on a speech about two hours long. The speech included isolated words, sentences, and was made by one speaker, a male announcer. The voice data was passed through a low-pass filter with a cut-off frequency of 6 KHz and digitized at a 12 KHz sampling rate. Acoustic phonetic segments with phonetic labels were obtained manually and 16th-order LPC analysis was performed on the data where frames had 20 ms, durations shifted every 5 ms.

Next, waveforms were automatically separated as follows. The start and end points for waveforms were chosen as the zero cross points preceding the local positive peaks that lie between +5ms and -5ms of the phonetic borders. The segmented waveforms were then organized according to the categories such as phonological context, average pitch within the phoneme, pitch contour, duration, and entered in the dictionary. In order to obtain smooth pitch contours, it is desirable to somehow convert the shapes of the pitch contours of the phonemes into data. We used average pitch as a rough approximation, and an approximate slope of the pitch contour within the phoneme derived using the least squares method.

LENTHY SPEECH

AD DATA

WAVEFORM CUTOUT

WAVEFORM INFORMATION

WAVEFORM DICTIONARY

Fig. 3 Waveform Dictionary Construction
5. WAVEFORM SELECTION

Selecting waveforms from the dictionary that most closely correspond to the prosodic information and the input phoneme string is the key importance to the new method.

(1) SELECTION PARAMETERS

The parameters for selection, arranged as headings in the dictionary, are (a) phonological context C, (b) average pitch V, (c) pitch contour F, (d) duration T, and (e) amplitude A. The contours of spectral and pitch, and characteristics such as duration and amplitude will be preserved in waveforms. This introduces two issues that must be addressed: one, a method of qualitatively expressing these features must be devised and two, the various parameters must be appropriately weighted for selection.

(2) SELECTION PARAMETER WEIGHTING

Parameters are weighted in accordance with the fact that pitch information is thought to have a major influence on the naturalness of synthetic sound. In other words, average pitch within phonemes W and pitch contour wt are weighted heavily, whereas time w and amplitude parameter wa can be assigned small values, because of their potential to modify waveform processing.

(3) SELECTION EVALUATION FUNCTION

The input phoneme string is segmented by applying windows of length N (N=7, 6, ... 1) to each phoneme in the string, and the segments are looked up in the waveform dictionary. If candidate waveforms are found, the following evaluation function \( H \) is applied to define the closeness of the match:

\[ H = bn + (1-b)W \]

where,

\[ W = \sum_{n=1}^{N} (Vp\cdot Vs) + \sum_{n=1}^{N} (Fp\cdot Fs) + \sum_{n=1}^{N} (Tt\cdot Ts) + \sum_{n=1}^{N} (Ap\cdot As) \]

Here, \( Vp, Vs, Fp, Fs, Tt, Ts, Ap, As \) are the values extracted from the waveform dictionary. \( n \) is a nonlinear ratio constant, represents the effect of adjacent phoneme influence. The letter \( p \) is the value found in the dictionary, and \( s \) is the goal value obtained by prosody pattern establishment processing. Finally, \( b \) is a coefficient representing the balance between active selection based on the phonetic environment and static selection derived from the prosody. Thus, evaluation function \( H \) yields the difference between the desired prosody pattern and the waveform prosody characteristics. Moreover, since the phonological environment, that is the feature expressing spectral variation, has already been corrected by the window length function, both prosody and phonological parameters are taken into account.

(4) EXAMINATION OF BALANCE COEFFICIENT \( b \)

In order to determine the balance coefficient \( b \), synthetic sounds were produced incorporating different values of \( b \) as a parameter, then subjecting the resultant sounds to comparative tests. The comparative test conditions were as follows:

- [Values of \( b \)] 5 values: 0.1, 0.3, 0.5, 0.7, and 0.9.
- [Values of \( w \)] \( wv=0.45, wt=0.25, wa=0.15 \)
- [Subjects] 5 males.
- [Test procedure] Subjects were instructed to judge between two sentences at a time based on "Which is better quality?"

The object of the tests was to evaluate the effect of \( b \). Each sentence was used 20 times for a total of 120 sentences altogether. Natural voice prosodies were used to enable the subjects to render clear "yes-no" judgments as to which sentence was to be selected. From the results, shown in Fig.4, it is apparent that the highest score was attained when \( b=0.5 \). This is the point, in other words, where the phonemic environment and the prosody conditions are judged by the subjects to be in balance or equal.
6. CONSTRUCTING WAVEFORMS

When all candidate waveforms failed to satisfy the threshold, a technique such as LPC is applied to prepare LPC parameters expressing spectrals corresponding to each phoneme. With the indicated pitch, it is possible to generate an excited waveform using the pulse and residual. The size of the dictionary determines how often waveform selection fails; large dictionaries are obviously better than small ones. Because this method uses computer RAM extensively and the cost of RAM is rapidly decreasing, it is thought that waveform selection failures will be uncommon.

7. EVALUATION OF SYNTHESIZED SOUND

Results of listening tests (23 short sentences synthesized using the described method) show that the generated waveforms possess high clarity and are judged as having excellent quality. There were some concerns, however, about discontinuities in pitch, spectral pattern, and phase because of the changing waveforms used for each phoneme. However, the listening tests indicate that the deterioration is minimal in most cases, and satisfactory synthetic sound is obtained. An example of a concatenated waveform is shown in Fig. 6.

8. CONCLUSION

With the goal of improving the quality of sound synthesized by rule, a new model based on a waveforms dictionary which incorporates phonetic environment, pitch, duration, amplitude, and other information was developed. From an input phoneme string and prosody information, appropriate waveforms are looked up in the dictionary, selected, and concatenated to generate high quality synthetic sounds.

We still need to quantitatively evaluate the quality of the synthesized voice and verify its effectiveness. At the same time, we intend to develop an upgraded version of the method based on further studies in the following areas:

(1) Reconsideration of the categories used in the waveform dictionary and the way these are expressed. An efficient method of expressing phoneme pitch pattern is particularly desired.

(2) Completion of the waveform dictionary.

(3) A more efficient lookup method for the appropriate waveform.

ACKNOWLEDGEMENT

The author wishes to express his gratitude for the guidance and encouragement received from Dr. Ryohei Nakatsu. He also wishes to express his thanks to the members of his group, especially Mr. Kazuo Hakoda and Mr. Shin-ya Nakajima.

REFERENCES


