Automatic Creation of CV Templates for Formant Type Speech Synthesis Based on HMM-Based Segmentation and Syllable Boundary Detection

Takahiro Ohtsuka, Chang-Sheng Yang
and Hideki Kasuya

Department of Electrical and Electronic Engineering Utsunomiya University
2753 Ishi-machi, Utsunomiya, 321-8585 Japan

Abstract: An automatic method to create CV formant-source templates from continuous speech corpus is proposed for speech synthesis, where the boundaries of the CV templates are decided on the basis of the Mahalanobis distance. The synthetic experiments have proved the method to be useful.

INTRODUCTION

A formant type speech synthesizer has immense advantages in that it allows us to generate speech with various voice quality variations and talker individualities. But it has suffered from unnatural speech quality, not because of theoretical limitations but because of an incomplete set of rules for the synthesis. Any insufficient approximation to the acoustics causes degradation of perceived quality of synthetic speech. We have developed a novel formant type speech synthesizer in Japanese based on concatenation of CV (consonant-vowel) formant-source templates obtained from natural utterances, in which multiple sets of formant and voice source parameter values are used for each of the CV syllables. This paper describes an automatic method to create the CV formant-source templates from speech corpus. The ARX (autoregressive with exogenous input) analysis method [1] is first used to automatically extract formant and voice source parameters and then an HMM-based segmentation [2] is performed to locate the phone segments. The segments are further analyzed to detect the boundary of CV templates. The method is proved to be useful in creating the CV templates by experiments performed on 503 Japanese sentences.

METHOD

An automatic procedure to create the CV template is described in Figure 1. Inputs to the system are a continuous speech signal and the corresponding phonetic transcriptions. An HMM method is applied to segment speech into phone units. Based on the phone segments obtained, boundaries of the CV templates are detected according to the Mahalanobis distance to be described below. In the postprocessing, improper CV templates are removed according to predetermined criteria. The source parameters of the CV templates are finally normalized for speech synthesis.

Acoustic Analysis

Acoustic parameters needed for HMM-based segmentation and for the CV templates are extracted from speech data. The ARX analysis method [1] is used to automatically extract formant and voice source parameters for the CV template. In addition to the first three formant frequencies, the mel scale cepstral parameters are also used for the HMM-based segmentation, because they provide good performance in the phone unit recognition.

Phone segmentation by HMM

Using the phonetic transcriptions and acoustic parameters, phone segments are located purely based on the Viterbi algorithm. The HMM parameters are estimated by the Baum-Welch algorithm with acoustic-phonetic HMMs, which is almost automatically derived from the Roman alphabet of Japanese sentences.

FIGURE 1. Automatic creation of CV template.

Detection of boundaries of CV template

A beginning of the CV template is defined as an instant where the articulation begins to change from a preceding vowel V₁ to the consonant C.
FIGURE 2. Detection of boundary of CV template.

This instant is obtained by the Mahalanobis distance of an acoustic measurement to the average of the vowel $V_1$ (see Figure 2). The Mahalanobis distance is given by

$$d^2(n) = (x(n) - \mu_v)^T \Sigma_v^{-1} (x(n) - \mu_v),$$

where $\Sigma_v$ is a variance-covariance matrix of a vowel $V$, $\mu_v$ is a mean vector of $V$, and $x(n)$ is an observation vector at time $n$. $\Sigma_v$ and $\mu_v$ are obtained directly from the model parameters used in the HMM. A point where $d(n)$ first exceeds a threshold is regarded as the beginning of the following consonant $C$, thereby yielding the initial boundary of the CV segment. The threshold is obtained by multiplying the minimum of $d(n)$ to $V_1$ by a constant. The final boundary of the segment is similarly detected as shown in Figure 2.

Postprocessing

Erroneous phonetic transcriptions create undesirable CV templates that should eventually be removed. The following criteria have been applied for that purpose:

1. Duration of the segment is extremely different from the average of the template duration.
2. The source amplitude value is too small, the fundamental period does not exist, or the Mahalanobis distance is unreasonably large.

Normalization of CV template

The voice source parameters in the template should be normalized so that they can be used in various prosodic contexts in the synthesis. The fundamental period $T(n)$, the amplitude of voicing source $AV(n)$ and the amplitude of unvoiced source $NA(n)$ are all normalized as follows:

$$T'(n) = \frac{T(n)}{T(0) + \frac{T(N-1) - T(0)}{N-1} \times n},$$

$$AV'(n) = \frac{AV(n)}{AV(0) + \frac{AV(N-1) - AV(0)}{N-1} \times n},$$

$$NA'(n) = \frac{NA(n)}{NA(0) + \frac{NA(N-1) - NA(0)}{N-1} \times n},$$

where $N$ is the duration of the CV template, $n = 0, \ldots, N-1$. It is considered that $NA(n)$ is in proportion to $AV(n)$, so it is normalized with regard to $AV(n)$.

EXPERIMENTS

Speech data were phonetically balanced sentences read by a male radio announcer. The sampling frequency was 14.7 kHz. The acoustic parameters were:

1. The first three formant frequencies,
2. 10 mel scale cepstral coefficients,
3. Normalized energy,
4. First order derivatives of the above parameters.

Dimension of the observation vector was 28. The formant frequencies were extracted by the ARX method and re-sampled every 5 ms. The mel scale cepstral parameters were analyzed every 5 ms using a Hamming window of 15 ms. A context independent HMM with three states left-to-right model and single Gaussian density with diagonal covariance matrix was used. The parameters of the HMM were estimated from 200 sentences. The phonetic transcription was almost automatically derived from the Roman alphabet of 503 Japanese sentences.

RESULTS AND DISCUSSION

Phone segmentation by the HMM provided 77.7% correct boundaries for 17,029 phones, given a tolerance of 20 ms. The performance was considered reasonable as the first step to the next boundary detection phase. Perceived quality of speech synthesized by the CV templates thus obtained was natural sounding, validating the proposed CV boundary detection algorithm and the normalization procedure for the voice source parameter.

REFERENCES