Speech Synthesis Method for Man-Machine Dialogue System and Psychological Evaluation of Synthetic Speech

Hideki KASUYA, Chang-Sheng YANG, Wen DING, Jin-Lin Lu, and Takamitsu MATSUSHITA
Department of Electrical and Electronic Engineering, Utsunomiya University
2553 Ishii-Machi, Utsunomiya 321, JAPAN
e-mail: kasuya@utsunomiya-u.ac.jp

1. Introduction

Required for speech synthesis in a man-machine dialogue system are flexibility in controlling 1) prosodic features, 2) segmental features and 3) extralinguistic (personal) features and 4) natural sounding voice quality. In order to meet all these requirements, a speech synthesis method that allows to independently control voice source and vocal tract characteristics is indispensable. Although a waveform concatenation method as PSOLA (pitch synchronous overlap add) is now appreciated well because of its ability of generating natural sounding speech, the method lacks by nature the first three requirements mentioned above. A formant-type speech synthesis method which obviously meets the first three requirements has a long history of elaborated research and development and has realized many products, but still suffers from insufficient synthetic voice quality, postponing achievement of the fourth requirement. It is believed, however, that this is not because of its theoretical limitation as a synthesis scheme but because of an insufficient set of rules that has been used to generate synthesis parameters. In particular, acoustic manifestation of a consonant is quite complex and thus control rules for the synthesis must be highly sophisticated. Any approximation to the acoustics causes degradation of perceived quality of the consonant.

We have developed a novel formant-type speech synthesizer based on concatenation of CV templates, in which a set of formant and voice source parameters is stored as a template of each of vowel-consonant-vowel segments of Japanese, instead of being generated by rules. This is a compromise between scientific and technological approaches. Advantage of this method is that structure of formant/antiformant filters becomes quite simple, consisting only of a set of cascaded filters. In this report, presented are a novel ARX (an auto-regressive with exogenous input model) analysis method to automatically extract formant/antiformant and voice source parameters from an utterance, an algorithm to create the CV template, a synthesis method, applications to the study of rhythm control in speech synthesis for man-machine dialogue.

2. ARX Speech Analysis Method

2.1 ARX Speech Modeling

Speech production process is modeled as a time-variant IIR (infinite impulse response) system with an equation error described as the following equation:

$$\sum_{i=0}^{p} a_i(n)s(n-i) = \sum_{j=0}^{q} b_j(n)u(n-j) + \varepsilon(n) \tag{1}$$

where s(n) and u(n) denote an observed speech signal and an unknown input glottal waveform at time n, respectively. In the above equation, ai(n) and bj(n) are time-varying coefficients. p and q are model orders, and e(n) is an equation error associated with the model. Furthermore, u(n) is the differentiated voicing source signal based on RK (Rosenberg-Klatt) model in which the radiation characteristics of the lips are included. The RK waveform model includes four parameters: pitch period T0, amplitude AV, open quotient OQ, and spectral tilt parameter TL. Equation (1) represents an auto-regressive with exogenous input (ARX) model, because the input signal u(n) of the equation is not white, whereas e(n) is assumed to be white. This is in contrast with ARMA (auto-regressive moving average) model in which the input signal is assumed as white. However, both models can be viewed as a pole-zero model. By performing the z-transform onto eq.(1), one gets the following equation,

$$S(z) = \frac{B(z)}{A(z)}U(z) + \frac{1}{A(z)}E(z) \tag{2}$$

where S(z), U(z) and E(z) are the z-transform of speech signal s(n), voicing source signal u(n), and equation error e(n), respectively. Figure 1 shows an ARX model consisting of an IIR filter and an AR filter. The vocal tract transfer function of voiced sounds is represented by B(z)/A(z), whereas the speech production process of unvoiced sounds could be approximated by an AR model with a transfer function 1/A(z).

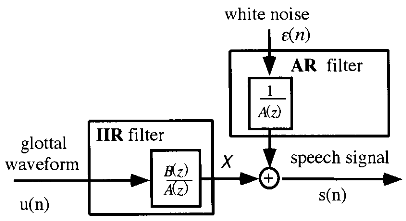

Fig. 1 ARX model decomposed into an IIR filter and AR filter.

2.2 Joint Estimation of Voice Source and Vocal Tract Parameters

An adaptive algorithm based on Kalman filter is used to estimate the time-varying coefficients of ARX model, while the simulated annealing method is applied to estimate RK voicing source parameters. By solving a polynomial equation consisting of A(z) coefficients, formant frequencies and bandwidths are obtained, whereas the roots of a B(z) polynomial equation yields antiformant values. As there is no theoretical guarantee for the stability of Kalman filter, the estimated formant frequencies and bandwidths are monitored. For each unstable estimate, the negative values of bandwidths are inverted to achieve the stable formant parameters. In the analysis, pitch period T0 is defined as the interval between two successive glottal closure instants, since the instants can be easily obtained from the estimated RK glottal waveform. This definition seems quite reasonable from both production and perception viewpoints. The open quotient OQ is an important parameter to determine spectral characteristics of the glottal waveform. A hybrid method has been devised for the estimation, in which two different optimization criteria are successively applied in the time and frequency domain.

2.3 Results of ARX Analysis

Figure 2 illustrates an example of the ARX analysis for an utterance of /muzukasii reedaio oboeru/ (to learn a difficult example problem) by a male, in which (a) and (b) represent sound spectrograms of original and re-synthesized speech, respectively, and (c) and (d) respectively indicate original and re-synthesized speech waveforms. It can be seen in the figure that acoustic characteristics of the re-synthesized speech signal stands comparison with those of the original. In fact, perceived quality of the re-synthesized speech was quite similar to the original.

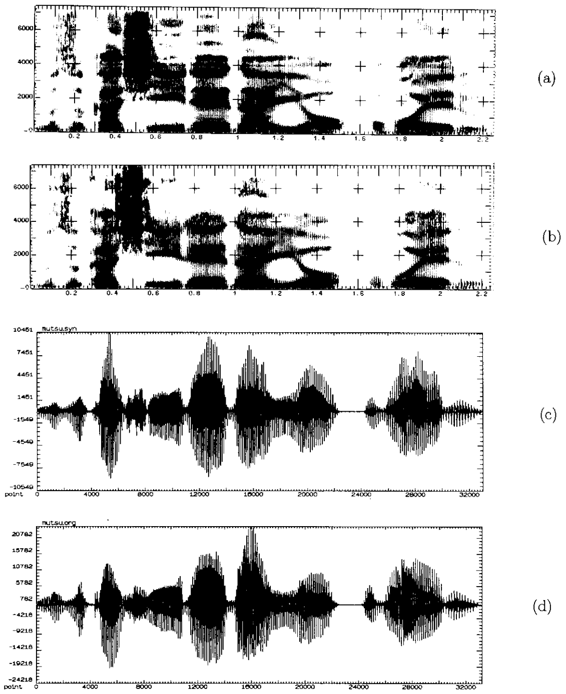

Fig. 2 Results of /muzkasii reedaio oboeru/ uttered by a Japanese male speaker. (a) sound spectrogram of original speech, (b) sound spectrogram of re-synthesized speech, (c) original speech waveform, and (d) synthesized speech waveform.

3. Creation of CV Templates

Since acoustic features of a consonant C spreads not only over a sound interval of CV (consonant-vowel) but over a transitional segment from a preceding vowel V to the consonant, the templates should be prepared for all the combinations of a VCV sequence of

Japanese phonemic system. VCV was read by a speaker in a carrier sentence as /sorewa VCV desu/ (That is VCV), for which the ARX analysis was performed. Articulatory movements were then traced on the formant trajectories, by computing several time-varying functions of formant frequency trajectories. A starting point of a CV template is defined as a time point where a value of the functions abruptly increases, while an end point of the template is regarded as a time point where a value of the functions rapidly decreases, signaling stationary movements of the articulatory organs.

4. Speech Synthesis System

A proposed speech synthesis system based on concatenation of the formant/antiformant and voice source parameters is illustrated in Fig. 3. Inthe template dictionary, formant/antiformant (vocal tract) parameters and normalized values of the voice source parameters are stored for each of the CV segments. On the basis of segmental information obtained by text analysis, the vocal tract and voice source templates are retrieved in the dictionary. Voice quality control algorithms, e.g., converting female to male voice, operate on the vocal tract templates, which is followed by concatenation process based on given duration of each mora. The vocal tract parameters are given to the formant/antiformant network. Prosodic information, on the other hand, is used to generate a basic time pattern of the voice source parameters and subjected to the voice quality control, e.g., generating breathy voice. The voice source parameters are then used to generate voice source signals which excite the formant/antiformant network. Moraic duration is determined on the basis of both prosodic and segmental information.

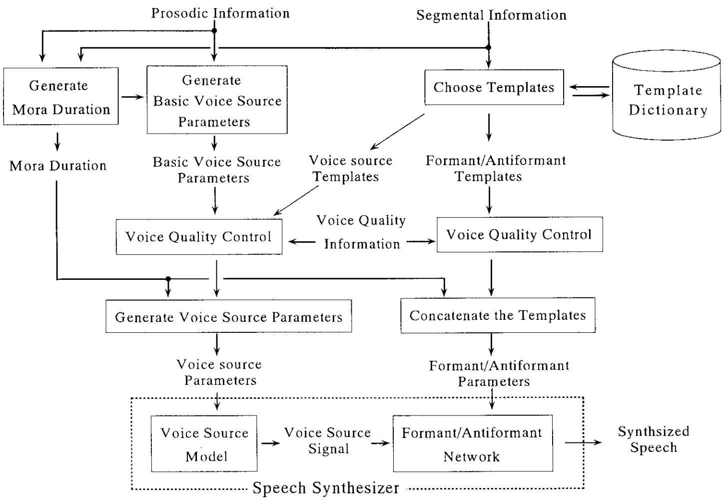

Fig.3 Block diagram of the speech synthesis system for man-machine dialogue.

5. Speaking Rate in Man-Machine Dialogue

5.1 Characteristics of Mora Duration in Dialogue

Temporal structure of Japanese mora duration was investigated in conversational speech and read speech and compared in terms of variability of the duration and degree of context dependency. A text to be read was made from the conversation, by syntactically smoothing the original utterance in the dialogue. This method allowed us to compare the characteristics of the duration in the two speaking modes, based on nearly the same phonemic context. Major findings are as follows: (1) moraic durational variation increases by 30 % as compared to the read utterance, (2) the average duration decreases by 21 % in the dialogue, (3) the duration within a breath group and within a prosodic phrase is affected much more in the dialogue than in the read utterance, and (4) putting emphasis does not always cause elongation of morae.

5.2 Preference Measurement of Speaking Rate of Synthetic Speech in Man-Machine Dialogue

As a man-machine dialogue experiment with synthetic speech produced by the proposed method, tourist information task was organized, in which a tourist asked the machine several questions about Nikko National Park by voice. Since speech understanding system was not available, a human experimenter recognized a question, chose an appropriate answer to the question from the entry of answers and generated synthetic speech. Tourists were required to make judgments on the preference of speaking rate of synthetic speech. Also made was the same preference judgment for read-style synthetic speech, results of which were compared with that for the dialogue. It was concluded from the experiment that (1) addition of variability of the moraic duration was not significant to make synthetic speech in the dialogue more natural, (2) faster speaking rates were preferred more in the dialogue than in the read speech, (3) people who spoke fast tended to prefer faster speaking rate, and (4) elderly people preferred slow speaking rate both in the dialogue and read style than younger people, indicating strong speaker dependency of the speaking rate.

6. Summary

A novel speech synthesis system has been developed, which allows a flexible control of the prosodic, segmental, and voice quality features, using multiple templates of a Japanese CV syllable consisting of formant/antiformant and voice source parameters. A new ARX analysis method was developed to automatically estimate the template parameters from the utterance. The method incorporated time-varying Kalman filter and simulated annealing algorithm to jointly estimate the voice source and vocal tract parameters. Man-machine dialogue experiment with synthetic speech showed that a voice response system should manage speaking rate depending on individual questioners.

Keywords: speech synthesis, formant template, concatenation method, speaking rate