A NEW SPEECH SYNTHESIS SYSTEM BASED ON THE ARX SPEECH PRODUCTION MODEL

Weizhong Zhu and Hideki Kasuya

Faculty of Engineering, Utsunomiya University,
2753 Ishii-machi, Utsunomiya, 321 Japan.
E-mail: zhu@klab.ishii.utsunomiya-u.ac.jp

ABSTRACT

In this paper, we present a new formant-type speech analysis-synthesis system based on the ARX (Auto-Regressive with Exogenous Input) speech production model. The model consists of cascade formant-antiformant synthesizers driven by a voicing source and an unvoiced turbulent noise source. One of the key features of the proposed method is that we have an algorithm to automatically measure the voicing source, unvoiced source and formant-antiformant parameters of the synthesizer directly from natural speech waveforms. After having automatically obtained estimates of the parameters from natural speech, one can manipulate the estimates using a flexible editing tool that has been developed as a part of the system. By changing values of the fundamental frequency, glottal open quotient, spectral tilt parameter, turbulent noise level, formant-antiformant frequencies and bandwidths, we can synthesize natural sounding speech with various voice qualities including modal, breathy, tense, and whisper voice.

Acoustic correlates of these voice qualities could be systematically investigated using the proposed system. Since our analysis-editing-synthesis system has been developed on the MS-Windows platform, it is expected that it will be a useful tool in various basic areas of speech science and technology.

1. INTRODUCTION

Klatt's speech synthesizer [1] has made a large contribution to various areas of speech research. With his synthesizer, however, there are a large number of parameters to be controlled to generate natural sounding speech sounds. It is also very difficult to be trained to use it effectively. In this paper, we present a new speech analysis-editing-synthesis software system based on the ARX model which can be used with MS-windows, one of the most popular graphical user interface platforms in personal computers.

According to the source-filter theory of speech production, formant frequency is one of the most important parameters characterizing phonetic features of speech. In addition, voice source parameters also appear to be important in characterizing voice quality. We have proposed an adaptive pitch-synchronous analysis method to estimate the vocal tract (formant-antiformant) and voice source parameters from a natural speech waveform [2]. Using this method, a voicing source waveform is approximated by the Rosenberg-Klatt (RK) model [3] and the unvoiced source is represented by a white noise. The Kalman filter algorithm is used to estimate the formant-antiformant parameters from the coefficients of the ARX model.

After having automatically obtained estimates of the source and vocal tract parameters from natural speech, a flexible editing tool allows the user to manipulate all the estimated acoustic parameters by simply clicking a mouse.

We can synthesize natural sounding speech with various voice qualities by changing values of the fundamental frequency, glottal open quotient, spectral tilt parameter, turbulent noise level, and formant-antiformant frequencies and bandwidths.

The core of the system is inherited from the previous speech analysis, synthesis and evaluation system [4][5]. The system is developed in C++ Object Window with user-friendly interface, providing the user with the full advantages of the Windows applications.

2. ARX SPEECH PRODUCTION MODEL

2.1. ARX Model

Speech production process can be modeled as a time-variant IIR system with an equation error as follows,

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

where $s(n)$ and $u(n)$ denote a speech signal and a glottal waveform at time n, respectively. In the above equation, $a(n)$ and $b(n)$ are time-varying coefficients. p and q are model orders, and $e(n)$ is an equation error. When $e(n)$ is assumed to be white, the equation represents an ARX model.

By performing the Z-transform onto the equation (assuming time invariance), one gets the following equation,

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

where $S(z)$,$U(z)$ and $E(z)$ are the Z-transform of the speech signal $s(n)$, voice source $u(n)$, and equation error $e(n)$, respectively.

The ARX model consists 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 production process of unvoiced sounds is approximated by an AR model with a transfer function $1/A(z)$ driven by a white noise.

2.2. Voicing Source Model

The RK model is used to represent a differentiated glottal wave form because of its capability of adjusting independently both the waveform and spectral slope as well as of relatively easy implementation. This model uses a generator of a rudimentary waveform defined as

$$g(n) = 2an - 3bn^2, \quad 0 \le n \le T * OQ,$$
$$= 0, \quad T * OQ < n < T$$
$$a = \frac{27 * AV}{4 * (OQ^2 * T)}, b = \frac{27 * AV}{4 * (OQ^3 * T^2)},$$

where T is a fundamental period, AV an amplitude parameter and OQ an open quotient of the glottal open phase to the duration of a complete glottal cycle. A value of $g(n)$ is 0 in the open period. Then $g(n)$ is filtered by a low-pass filter to adjust the tilt of its spectral envelope using a spectral tilt parameter TL.

3. ANALYSIS ALGORITHM

A pitch-synchronous method has been used to estimate the vocal tract and voice source parameters from a natural speech waveform, where an extended Kalman filter is applied [2]. An efficient optimization method is proposed to extract the RK voice source parameters by minimizing the mean-squares equation error (MSEE).

Speech signal includes voiced, unvoiced, mixed-voice and silent segments. In the ARX analysis, the analysis is synchronized with an estimated pitch period in a voiced and mixed-voice segment, whereas the analysis frame is shifted by 5 ms in an unvoiced segment. The voiced and mixed-voice segments are defined by the user with a labeling tool which is implemented as a part of the system.

3.1. Pitch Estimation

In order to estimate the source parameters fast and accurately, an efficient pitch estimation method has been devised. We define the pitch period as the interval between two successive estimated glottal close instants (GCI's) which are calculated from the estimated RK parameters (see Fig. 1). In the RK model, the GCI is determined by OQ, if T is known. Since the value of MSEE is variable for different sets of the RK parameters, the MSEE can be viewed as a function of the RK parameters. We extract the RK parameters by minimizing the MSEE. OQ is the most sensitive to the MSEE among the RK parameters. In consideration of pitch extraction at the current cycle, we take speech waveform with a window length T which is equal to the previous pitch period and calculate the MSEE with an estimated AV, a default TL and an OQ being varied within a certain range. The position that has the minimum value of MSEE is denoted as the GCI. Assuming that the power of the RK source waveform in a pitch period is equal to the power of the residual signal calculated from the previous estimated ARX coefficients and speech waveform by inverse filtering, an estimated AV is calculated from RMS of residual signal with a default OQ set by the user. TL is an option value. All the option values including the model orders can be set by the user using an option setting dialogue. The extraction is made on cycle-to-cycle basis.

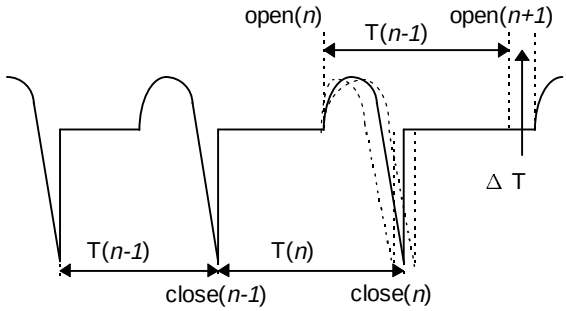

Figure 1: Procedure for searching the GCI's.

Figure 1 shows the way that we extract close(n) (current GCI) based on the previous GCI's. It also shows how to locate open($n+1$) (start position for the next period). The following procedure is used to extract these values.

(1) Take an analysis window length identical to the previous pitch period and calculate the MSEE with an estimated AV, a default TL and an OQ being varied within a certain range. Denote the position that has the minimum value of MSEE as close(n).

$$close(n) = open(n) + T(n-1) * OQ$$

(2) Take the pitch period T(n) from two successive GCI's.

$$T(n) = close(n) - close(n-1)$$

(3) Shift to the next period with T(n-1) and a adjusted value.

$$open(n+1) = open(n) + T(n-1) + \Delta T$$
$$\Delta T = (OQ - \alpha) * T(n-1)$$

where α is a specified value.

Besides the glottal waveform, a glottal noise component defined by the noise amplitude parameter (*NA*) is estimated from the prediction error in the glottal open phase,

$$NA = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} e^2(n)}$$

where N is the data length and $e(n)$ is the prediction error.

A segment of speech signal and an estimated RK waveform are shown in Figs. 2 and 3.

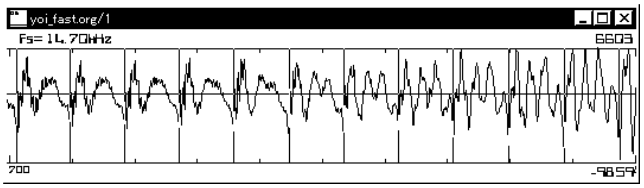

Figure 2: Original speech waveform.

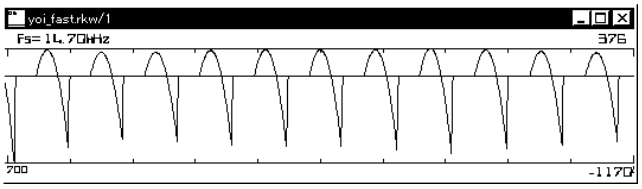

Figure 3: Estimated RK waveform.

3.2. Formant-Antiformant Estimation

In order to estimate the coefficients of the system transfer function $B(z)/A(z)$, one needs to know $u(n)$ and $s(n)$. We take as $u(n)$ a voice source waveform generated by the RK model using a set of assumed values of the parameters and as $s(n)$ an observed speech waveform. An extended Kalman filter algorithm is used to estimate the coefficients of the ARX model [2]. In the algorithm, the estimates of the coefficient vector and Kalman gain for the last point of the previous pitch period are used as the initial values of the current pitch period.

Formant and antiformant parameters are obtained by solving for the roots of the $A(z)$ and $B(z)$ polynomials, respectively. Although they can be calculated point by point within one pitch period from time-varying coefficients of $A(z)$ and $B(z)$, we save one set of formant-antiformant values within a pitch period for the synthesis for the following three reasons: (1) It takes immense computation time for the calculation and a large capacity for the storage of the parameters is needed, (2) Estimated formant values are not always stable, and (3) It is reasonable to assume the time-invariant nature of the vocal tract movement within one pitch period. To do so, we first calculate the formant-antiformant values at three time points within the pitch period , i.e. the last of the period, the middle of the open phase and the middle of the closed phase. We then select and save a set of the formant-antifformant values which has the maximum number of formants.

4. EDITING TOOL

Each of the estimated voice source parameters and formant-antiformant parameters is re-sampled every 5 ms and displayed on a corresponding graphical window. An example of the trajectories of estimated formant parameters is shown in Fig. 4. An editing tool can be applied to manipulate the parameters on the graphical window.

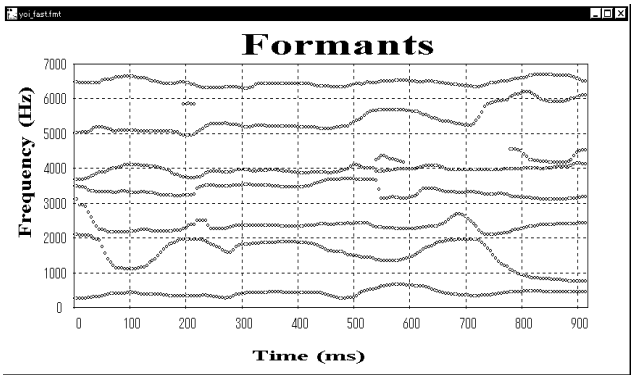

Figure 4: Trajectories of estimated formant parameters.

4.1. Editing the Parameters

A number of functions have been developed to edit the parameter being displayed as points on a graphical window. The following is the list of the functions.

1. Modify a value: Move the mouse to the point which is to be modified, press the left button and drag the mouse to a new value.
2. Add a new value: Press the Shift-key and click the left button at the point of the new value.
3. Delete a value: Move the mouse to the point to be deleted, press the Ctrl-key and click the left button.
4. Modify multiple values: Double-click at the start point and at the end point. The points between the start and the end are linearly changed.
5. Scale values: Double-click at the first point and at the last point on the window and input a scaling factor.

The right button is used to modify the bandwidth for the formant window.

4.2. Saving Graphical Window

A graphical window can be saved as a text file which contains data of main tittle and x-y tittle as well as the parameter values. The user can redraw the graphical window from a saved text file.

4.3. Saving All the Parameters

All the extracted acoustic parameters such as formant parameters, are also listed on the text-typed report window. By using the save command, the user can save the parameters as a text-typed data file. The data file can be accessed using other applications, such as Excel, to do further data analysis.

5. FORMANT SYNTHESIZER

A cascade formant synthesizer is used to synthesize the voiced and unvoiced speech. The RK model is used to synthesize the voiced sound, whereas M-series white noise is used to synthesize the unvoiced sound.

5.1. Cascade Formant Synthesizer

The synthesizer is composed of the second-order resonators in cascade form. The spectrum for each resonator is expressed as

$$H(z) = \frac{a}{1 - bz^{-1} - cz^{-2}},$$
$$b = 2\exp(-\pi B / f_s) * \cos(2\pi F / f_s),$$
$$c = -\exp(-2\pi B / f_s),$$
$$a = 1 - b - c,$$

where *F*, *B* and *fs* are formant frequency, bandwidth, sampling frequency, respectively.

5.2. Adjustment of the Voice Amplitude Parameter

In the formant synthesizer, the gain of each second-order resonator at 0 Hz is always 0 dB. But the amplitude of the vocal tract transfer function $B(z)/A(z)$ estimated by the Kalman filter at 0 Hz may not be equal to 0 dB. As a result, the estimated value of *AV* based on the MSEE criterion can not be used directly in such a formant synthesizer. There are several ways to adjust *AV*: (1) minimize the difference of RMS between original speech and synthesized speech, (2) minimize mean-squares error between the original speech and synthesized speech, (3) minimize the difference of positive or negative peaks between original speech and synthesized speech. We have found that the RMS criterion yields a rather reasonable envelop of *AV* sequences as compared with the other methods.

5.3. Synthesis with Glottal Closure Instants

As we define the pitch period *T* as an interval between the two successive GCI's in the synthesis, the RK glottal waveform of one pitch is generated by changing the component of the RK glottal waveform in the close phase with that in the open phase (see Fig. 5). In this way, GCI's keep the same values, even when values of OQ are changed.

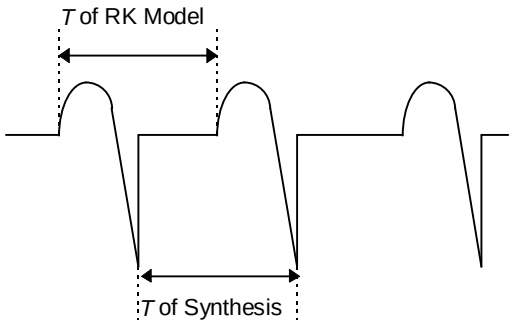

Figure 5: Pitch period procedure for synthesizing speech with GCI's.

6. CONCLUSION

A new speech synthesis system based on the ARX speech production model has been developed on the MS-Windows platform. One of the advantages of this analysis-editing-synthesis system is that it separates the voice source characteristics from those of the vocal tract. It allows reconstruction of speech after independently manipulating the acoustic parameters, to produce speech of various sound qualities. It is expected to be used as a useful tool in various basic areas of speech science and technology.

Acknowledgment : The authors would like to thank Ding.,W. and Matsushita, T. for providing Kalman filtering and formant synthesizer programs. This work was partly support by Grant-in-Aid for Developmental Scientific Research from the Ministry of Education, Science and Culture, Japan and International Communication Foundation of ASHIKAGA Bank, Tochigi, Japan.

7. REFERENCES

1. Klatt, D. H., "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, Vol. 67, pp.971-995, 1980.

2. Ding, W., Kasuya, H. and Adachi,. S., "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model," *IEICE Trans. Inf. & Syst.*, Vol.E78-D, pp.738-743,1995.

3. Klatt, D. H. and Klatt, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, Vol. 87, pp.820-857, 1990.

4. Zhu, W. Z., Endo, Y., Kasuya, H. and Kikuchi, Y., "An integrated speech analysis, synthesis and evaluation system on MS-Windows," *Technical Report of IEICE*, SP95-106, May 1995(In Japanese) .

5. Zhu, W. Z., Kikuchi, Y., Endo Y., Kasuya, H., Hirano, M. and Ohashi, M., "An integrated acoustic evaluation system of pathologic voice," *Proc. ICSLP 94*, Yokohama, S32-11, September 1994.