Aperiodicity Control in ARX-Based Speech Analysis-Synthesis Method

Takahiro Ohtsuka† and Hideki Kasuya‡

†Graduate School of Engineering Utsunomiya University
‡Faculty of Engineering, Utsunomiya University
{takahiro,kasuya}@klab.ee.utsunomiya-u.ac.jp

Abstract

We present an improved algorithm for a robust speech analysis-synthesis method based on an auto-regressive with exogenous input (ARX) speech production model proposed previously. The speech analysis-synthesis method is capable of making an automatic estimation of vocal tract (formant) and voice source parameters from a speech utterance, generating accurate formant values even for very high-pitched voices. The improved algorithm presented in this paper incorporates aperiodic components included in the voice source signal, taking the dynamic nature of the speech production process into account. Perceptual experiments show that implementation of the aperiodic components in the analysis-synthesis is very effective in improving the perceived quality of synthetic speech, particularly for soft voices, typical of female voice quality.

1. Introduction

One of the authors and his colleagues developed a novel pitch-synchronous speech analysis-synthesis method based on an auto-regressive with exogenous input (ARX) speech production model [1]. The method makes an automatic estimation of the formant and voice source parameters of a speech utterance. An additional function for editing the source-formant parameters thus estimated allows the speech sounds to be re-synthesized for various voice qualities and speaking styles [2]. The method, however, encountered deficiencies in the analysis of high-pitch voices and weak-voiced sounds and inadequate spectral distortion resulting from the excluded formants having broad bandwidths and/or on the real axis. In the synthesis, it also introduced clicking sounds in the transition between the vocalic and consonantal segments due to wrong alignment of the formants, and often produced the buzzy voice quality of re-synthesized speech. In order to overcome these performance flaws, we proposed a novel algorithm based on the same ARX speech production model, and we showed experimentally that the improved method was indeed superior [3].

In this paper, we present a further improved algorithm, specifically, implementing a model of aperiodic components included in the voicing source signal, taking the dynamic nature of the speech production process into account. Perceptual experiments show that the inclusion of an aperiodic component model in analysis-synthesis is necessary to reproduce natural sounding synthetic speech, particularly for soft voices, typical of female voice quality.

2. ARX speech production model

The ARX speech production model is shown in Fig.1 and is represented by the linear difference equation

\[ s(n) + \sum_{k=1}^{p} a_k s(n-k) = \sum_{k=0}^{q} b_k u(n-k) + c(n) \quad (1) \]

where the input \( u(n) \) denotes a periodic voicing source signal and the output \( s(n) \) denotes a speech signal. A part of the glottal noise component is simulated by the white noise \( e(n) \). In the equation, \( a_i \) and \( b_i \) are vocal tract filter coefficients, and \( p \) and \( q \) are ARX model orders.

![Figure 1: ARX speech production model [1].](image)

We employ the Rosenberg-Klatt (RK) model to represent a differentiated glottal flow signal, including radiation characteristics. The RK waveform is given by

\[ g(n) = g_e(nT_s) \quad (2) \]

\[ g_e(t) = \begin{cases} 2at - 3bt^2, & 0 \leq t < OQT0 \\ 0, & elsewhere \end{cases} \]

\[ a = \frac{27AV}{4OQ^2T0}, \quad b = \frac{27AV}{4OQ^3T0^2} \quad (3) \]

where \( T_s \) is a sampling period, \( AV \) is an amplitude parameter, \( T0 \) is a pitch period, and \( OQ \) is an open quotient of the glottal open phase of the pitch period. The differentiated glottal flow waveform \( u(n) \) is generated by smoothing \( g(n) \) through the use of a low-pass filter, where the tilt of the spectral envelope is adjusted via a spectral tilt parameter, \( TL \).
3. Novel ARX-based analysis-synthesis algorithm

3.1. Analysis algorithm

The Kalman-filter based algorithm [1,2] revealed three critical flaws, when exposed to various types of speech utterances: unstable estimation of formant values for high-pitch voices, incorrect estimation of higher formants for weak-voiced sounds, and spectral distortion due to the exclusion of large-bandwidth formants. The improvements made in each of these three areas are detailed [3].

A least square (LS) method was employed instead of the Kalman filter, allowing the vocal tract filter coefficients to be estimated based on data from an analysis frame of variable length, even spanning several pitch periods. One of the most significant features of the ARX-based LS algorithm, as compared with the conventional linear prediction analysis method, is that multiple voicing source pulses are assumed in the estimation of the filter coefficients as shown in Fig.2.

We introduced an adaptive prefilter for the input and output signal of the ARX model in order to achieve better estimation of the higher formants (Fig.2). The adaptive prefilter is comprised of inverse filter coefficients obtained from the autoregressive (AR) analysis of the RK voicing source waveform.

To compensate for the spectral distortion resulting from the excluded poles, we introduced a second order adaptive filter that is directly computed from the excluded poles.

These three revisions of the ARX-based analysis method reported previously proved to be effective for a large variety of utterances across a wide range of speakers including infants [4].

3.2. Synthesis algorithm

A cascade formant synthesizer is used in the ARX-based method instead of Klatt’s cascade/parallel configuration [5] to synthesize both voiced and unvoiced speech. The RK voicing source model generates voiced speech, whereas the M-sequence, pseudo random binary signal, is applied to synthesize unvoiced speech. The previous synthesis algorithm [1,2] introduced clicking sounds in the transition between the vocalic and consonantal segment due to discontinuity of the formants and often produced the buzzy voice quality of re-synthesized speech.

The clicking sounds turned out to be due to discontinuity of the formants, which occurs in two cases: 1) when the number of formants between two successive frames is not identical, and 2) when a formant frequency changes abruptly. Dynamic programming was employed to achieve the optimum match of the formants between the two frames using a distance measure defined by connection and disconnection costs [3]. This worked very well in eliminating the inadequate clicking sounds.

In the mean time, the buzzy quality of re-synthesized speech was reduced to a large extent by incorporating two major sophistications; fine control of the glottal closure instant (GCI) of the RK source waveform and randomization of the group delay in a higher frequency range [3]. This enabled the GCI to be manipulated with much higher resolution than the sampling period. The combination of these two key features made on the RK waveform has been successful in reducing the buzzy quality of re-synthesized speech.

4. Implementation of aperiodicity

The breathy voice quality is a common phenomenon in female voices [6] and in vowel segments in some phonetic contexts, e.g. preceded by certain unvoiced consonants. The breathiness is closely related to aperiodicity of the speech signal. Voiced fricatives and intervocalic /h/ sounds in Japanese also accompany both periodic and aperiodic components. In order to handle the aperiodic nature of the voice source signal, we extend the idea of randomizing the group delay of the RK voicing source waveform in the higher frequency region. A critical problem in this case is the automatic decision of the high-pass cut-off from frame to frame for the frequency range in which aperiodicity dominates.

4.1. Automatic estimation of aperiodic frequency range

A speech signal is filtered with a bank of the band pass filters. A modified autocorrelation function (MAF) is then computed from each of the band-pass filter outputs. The degree of MAF in the analysis frame is used to assess aperiodicity. As the dynamic nature of speech affects the degree of MAF, a number of factors need to be considered, most important of which appear to be vocal tract movement, pitch period change and voicing source amplitude change. These should be compensated for before calculating the MAF.

**Compensation for vocal tract movement**

The vocal tract filter parameters are estimated every 5 ms using the method described earlier and used to constitute a time-variant inverse filter for the speech signal. The inverse filter coefficients are interpolated every sample-point using the parameter values estimated every 5 ms. This time-variant inverse filtering of speech signals is expected to effectively remove the influence of vocal tract movement within the analysis frame.

**Compensation for pitch period change**

Pitch periods change smoothly in a continuous speech utterance due to intonation and accentuation involved in the utterance. This smooth change in the pitch from period to period obviously affects the MAF. In order to compensate for this effect, inverse-filtered speech signals are time-scaled so as to have identical periods over two successive pitch periods in the analysis frame, before calculating the MAF.
Compensation for voicing source amplitude change

A normalized MAF is employed to compensate for the effect of variations in amplitude, as follows [7]:

\[ R(\tau) = \frac{\sum s(n)s(n + \tau)}{\sqrt{\sum s^2(n)\sum s^2(n + \tau)}} \]

4.2. Determination of cut-off frequency

The cut-off frequency of the aperiodic frequency range is determined as follows:

STEP 1

1. Carry out the time-variant inverse filtering and obtain an inverse filtered signal \( v(n) \).
2. Take two segments of two pitch periods by windowing \( v(n) \): 

\[ v_{01}(n) = v(n)w(n) \]
\[ v_{02}(n) = v(n + T0/T_s)w(n) \]

3. Modify time scale of \( v_{02}(n) \) using the following equations:

\[ v_{02}'(n) = \frac{v_{02}(n)}{v_{02}(n/\alpha)} \]
\[ = \frac{\alpha}{N} \sum_{k=L}^{M} \left( \sum_{m=0}^{N-1} v_{02}(m)e^{-j\frac{\pi}{N}km} \right)e^{j\frac{\pi}{N}kn} \]
\[ L = -\min\left(\frac{N}{2} - 1, \frac{N}{2\alpha}\right) \]
\[ M = \min\left(\frac{N}{2}, \frac{N}{2\alpha}\right) \]
\[ N > \max\left(\frac{2T0}{T_s}, \frac{2\alpha T0}{T_s}\right) \]

The chirp-z transform and fast Fourier transform can be used to calculate the equations quickly.

4. Take a set of \((\tau', \alpha')\) that gives the maximum value of the normalized correlation \( R_0(\tau, \alpha) \) of \( v_{01}(n) \) and \( v_{02}(n) \).

STEP 2

1. Get a band-passed signal of the i-th channel \( v_i(n) \).
2. Get \( v_{11}(n) \) and \( v_{12}(n) \) by windowing \( v_i(n) \) according to STEP1-2.
3. Calculate \( R_s(\tau) \), where \( \tau \) varies around \( \tau' \).
4. The cut-off frequency \( \omega_c \) is the center frequency of the lowest frequency band which gives \( \max_{\tau} R_s(\tau) < \theta \).

STEP 3

1. Smooth the time-variant cut-off frequency \( \omega_c \) by five-point median filtering.

Time-variant \( \omega_c \) values are used instead of a fixed \( \omega_c \) proposed in the previous method [3] to re-synthesize the input speech.

5. Experiments

5.1. Examples of aperiodicity control

Figure 3 illustrates examples of the RK voicing source waveform with various degrees of aperiodicity, in which (a) is a voicing source pulse without aperiodicity, and (b) and (c) are the ones having aperiodicity in the frequency range above 3,500 Hz and 800 Hz, respectively. The figure shows fluctuations superimposed on the waveform, as the group delay is randomized. The extent of the fluctuation becomes larger as the cut-off frequency decreases.

A narrow-band sound spectrogram of an original female speech utterance /(sa)kihodo/ “a little while ago” and the cut-off frequency estimated are shown in Fig 4(a). We can see in the figure that aperiodicity dominates in the frequency range above 700 Hz in the voiced /h/ where aspiration noise exists. In the frequency range above 3,000 Hz for the vowel /i/, which follows the consonant /k/, the aperiodicity is strong. This acoustic feature obviously reflects the aspirated voice quality of the vowel /i/. Figures 4(b) and (c) show the sound spectrograms of re-synthesized speech with and without aperiodicity control, respectively. The randomization of group delay can clearly be seen to contribute significantly to mimicking the aspirated feature of the vowel /i/.

5.2. Perceptual evaluation of proposed method

Perceptual experiments were conducted to evaluate the quality of re-synthesized speech generated by the improved method, testing the effectiveness of adding aperiodicity to the voicing source signal. The speech samples used were continuous full sentence utterances of four females and one child. Four subjects participated in the experiments and were asked to choose the more natural stimulus of two synthetic stimuli, with and without aperiodicity control. The speech samples re-synthesized by the proposed method were judged to be more natural 85% of the time.

5.3. Comparison with LSP-based method

The line spectrum pair (LSP) method was taken up as an example of the parametric analysis-synthesis methods and was compared with the ARX-based method described above in terms of the perceived overall quality with and without the pitch change.
of ±10%. The speech samples and the subjects were the same as those used in the evaluation experiment. The subjects preferred the quality of speech samples re-synthesized by the ARX-based method 78% of the time.

6. Discussion

Figure 4 evidently indicates the significance of the proposed model of aperiodic components. In fact, we can clearly perceive the difference in the voice qualities between the two speech segments with and without aperiodicity control; the inclusion of the aperiodicity model successfully reproduces the breathy voice quality of the vowel /i/ preceded by the unvoiced plosive /k/. The perceptual evaluation, however, did not result in an overwhelming score for the speech samples produced by the aperiodicity model. This is considered to be due to the fact that the subjects evaluated the overall quality of re-synthesized speech, since they were not instructed to concentrate on specific sound segments. To validate the proposed method, therefore, additional evaluation experiments are necessary. It should be noted that the combination of the fine control of the glottal closure instant and the aperiodicity control model is responsible for the improvement.

7. Summary

A novel method, based on the ARX speech production model, for automatically estimating of vocal tract and voice source parameters was presented. A model of the aperiodicity of natural speech was included in the method, and when coupled with a sophisticated synthesis procedure, the method was shown through perceptual experiments to produce more natural re-synthesized speech than the well-known LSP method. Speech samples are available at http://www.klab.ee.utsunomiya-u.ac.jp/.

8. References