A NOVEL APPROACH TO THE ESTIMATION OF VOICE SOURCE AND VOCAL TRACT PARAMETERS FROM SPEECH SIGNALS

Wen Ding and Hideki Kasuya
Faculty of Engineering, Utsunomiya University
2753 Ishii-machi, Utsunomiya 321, Japan
E-mail: ding@itl.atr.co.jp

ABSTRACT

This paper presents a novel adaptive pitch-synchronous analysis method for simultaneous estimation of voice source and vocal tract (formant / antiformant) parameters from the speech signal. The method uses a parametric Rosenberg-Klatt model to generate a glottal waveform and an autoregressive with exogenous input (ARX) model for representing speech production process. The time-varying coefficients of the model are estimated with an adaptive algorithm based on Kalman filter, while the parameters of the Rosenberg-Klatt model are optimized using the simulated annealing method. In addition, a new hybrid error criterion is used to optimize the glottal opening instant. Furthermore, in order to estimate the fundamental period parameter T_0, it is defined as two successive glottal closure instants, and is estimated automatically based on the obtained differentiated glottal waveforms. Experiments using two-channel speech signals (speech and electroglottograph (EGG) signal) and continuous speech show a good estimation performance.

1 INTRODUCTION

Among a large variety of speech processing methods, joint estimation of voice source and vocal tract parameters has been found the one of the most important. Its importance lies on the fact that it possesses potential applications in areas such as speech analysis, speech synthesis, perception of voice quality, speech coding, acoustic phonetics and modeling of the speech production process.

Many parametric voicing source models have been proposed to approximate a glottal volume velocity waveform, but the source parameters have been mainly measured by manual means, most commonly by an experienced researcher [1], [2]. This drawback is especially troublesome when a large amount of speech data is to be processed. Therefore, automatic esitmation of voice source parameters becomes an attractive research task. Alku proposed an automatic method based on pitch-synchronous iterative adaptive inverse filtering [3], but it employs rather an ad hoc formant estimation algorithm and thus is difficult to be applied for a large variety of speech data. Using the synchrovoice EGG signal, Milenkovic proposed a method to jointly determine a voice source model parameters together with the coefficients of the inverse filter [4]. However, it needs an extra channel signal (EGG) and estimation of the pulse positions is essentially a heuristic procedure, and there is no way to guarantee its convergence. In another attempt to realize the joint estimation, Fujisaki and Ljungqvist proposed a glottal autoregressive moving-average (GARMA) analysis by synthesis method, but it is sensitive to the position and length of the analysis frame [5].

The motivation behind this paper is to develop an accurate method of automatic joint estimation. In section 2, the speech production process is represented by an ARX model by incorporating an Rosenberg-Klatt (RK) voicing source model into an IIR filter. In section 3, the source parameters are identified together with the time-varying vocal tract transfer function so as to minimize the mean square equation error of the ARX model. The estimation method consists of an optimization procedure for voice source parameters and an adaptive algorithm for formant / antiformant estimation. Finally we give some experimental results for synthetic and natural speech to show the validity of the proposed method.

2 SPEECH SIGNAL MODELING

2.1 ARX Speech Modeling

Based on Fant's source-filter concept [6], speech production process can be modeled as a time-variant IIR system with an equation error described as the following equation:

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

Wen Ding is now with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

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, $a_i(n)$ and $b_j(n)$ are time-varying coefficients. p and q are model orders, and $\varepsilon(n)$ is an equation error associated with the model. $u(n)$ is the differentiated voicing source signal generated by the RK model in which the radiation characteristics of the lips are included.

Equation (1) represents an autoregressive with exogenous input (ARX) model, because the input signal $u(n)$ of the equation is not white, whereas $\varepsilon(n)$ is assumed to be white [7]. This is in contrast with ARMA model in which the input signal is assumed to be white, but both models can be viewed as a pole-zero model.

The vector notation of the coefficients and data is

$$\theta(n) = \{a_1(n), \cdots, a_p(n), b_1(n), \cdots, b_q(n)\}^T, \quad (2)$$

$$\varphi(n) = \{-s(n-1), \cdots, -s(n-p), u(n-1), \cdots, u(n-q)\}^T, \quad (3)$$

where $a_0(n) = 1$ and $b_0(n) = 1$. Now the ARX model can be expressed as,

$$s(n) = \theta(n)^T \varphi(n) + u(n) + \varepsilon(n). \quad (4)$$

By performing the Z-transform onto Eq.(1) (assuming time invariant system), one gets the following equation,

$$S(z) = \frac{B(z)}{A(z)} U(z) + \frac{1}{A(z)} E(z), \quad (5)$$

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 $\varepsilon(n)$, respectively. Based on the above description, Fig. 1 illustrates 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) with a voicing source signal. The speech production process of unvoiced sounds could be approximated by an AR model with a transfer function 1/A(z) and a white input .

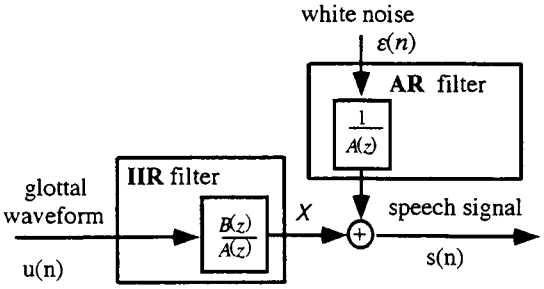

Figure 1: ARX model decomposed into an IIR filter and AR filter.

2.2 Voicing Source Model

The Rosenberg-Klatt model is used to represent a differentiated glottal waveform because of its capability of adjusting independently both the waveform and spectral slope as well as relatively easy implementation [1]. The differentiation simulates a lip radiation characteristic. This model uses a generator of a rudimentary waveform defined as

$$g(n) = \begin{cases} 2an - 3bn^2 & (0 <= n < T_0 \cdot \mathrm{OQ}) \\ 0 & (T_0 \cdot \mathrm{OQ} <= n < T_0) \end{cases} \quad (6)$$

where

$$a = \frac{27}{4} \cdot \frac{\mathrm{AV}}{\mathrm{OQ}^2 \cdot T_0},$$
$$b = \frac{27}{4} \cdot \frac{\mathrm{AV}}{\mathrm{OQ}^3 \cdot T_0^2},$$

in which T_0 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. Then $g(n)$ is filtered by a low-pass filter (LPF) to adjust the tilt of its spectrum using a spectral-tilting parameter TL. This filtered waveform will be referred to hereafter as a glottal waveform $u(n)$. From the definition, there are four parameters T_0, AV, OQ and TL which need to be estimated.

3 JOINT ESTIMATION

In the joint estimation algorithm, an adaptive procedure based on Kalman filter is applied for formant/antiformant estimation and an optimization procedure based on simulated annealing is used for source parameters estimation. Both of the above procedures are based on the minimization of the predicted mean-square equation error (MSEE) of the ARX model,

$$E = \frac{1}{N} \sum_{n=1}^{N} \{s(n) - \theta(n)^T \varphi(n) - u(n)\}^2. \quad (7)$$

Based on the Parseval theorem the time-domain error criterion in Eq. (7) can be transformed into a frequency-domain error criterion [7],[8], The algorithm, i.e. LMS, RLS, Kalman filter, requires that the prediction error be white. Therefore, it is necessary to pre-emphasize both $s(n)$ and $u(n)$ before analysis using such an operation as $s'(n) = (1 - z^{-1})s(n)$, $u'(n) = (1 - z^{-1})u(n)$. This operation flattens out the gross features of the weighted spectra throughout the frequency domain.

It seems reasonable to use glottal closure instants (GCI's) as the more reliable positions to estimate the fundamental period than using the structural features of the speech waveform [9]. In our case, the GCI can be obtained from the negative peak of the estimated RK glottal waveform. T_0 is therefore obtained as the elapsed time between two successive GCI's. In the following section, an automatic approach of T_0 estimation will be explained. Figure 2 shows the flowchart of the joint estimation. The optimization algorithm based on the simulated annealing method and the adaptive Kalman filter algorithm are explained in [10].

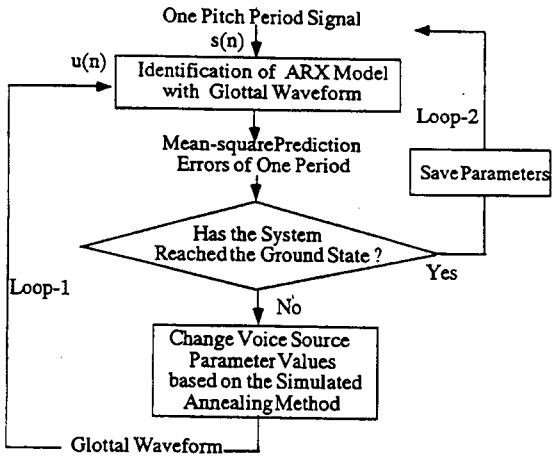

Figure 2: Flowchart of the proposed joint estimation method.

3.1 T_0 Estimation

In the ARX analysis method, T_0 is defined as the interval between two successive GCI's, since GCI's can be easily obtained from the estimated RK glottal waveforms. Thus it is possible to estimate automatically T_0 together with GCI's. Figure 3 gives a structual description of T_0 estimation.

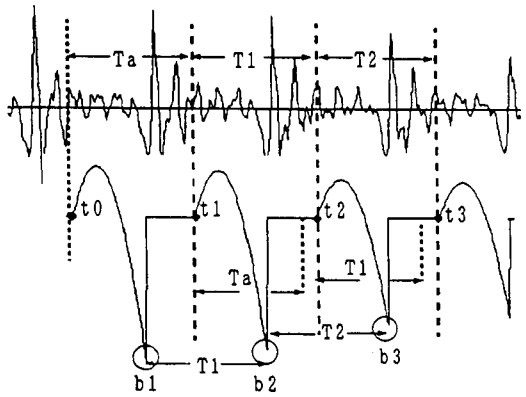

Figure 3: Estimation of RK model parameter T_0.

Firstly, the average fundamental period T_a is obtained. Next the first glottal opening instant GOI (t0) of RK waveform is assumed in a distance $0.3*T_a$ from the negative peak of the first segment of length T_a. Then the speech signals of length T_a are analyzed and the negative peak b1 of the estimated differentiated glottal waveform is obtained. For the second pitch period, the initial value of T_0 is set to T_a. The negative peak of RK glottal waveform b2 is then obtained by analyzing $s(n)$ with length T_a from t1. Now the time interval between b1 and b2 is the required fundamental period T1. In the next stage the speech segment between t1 and t2 are analyzed for signal duration of T1. In such a way, one can automatically extract the parameter T_0 of RK model.

3.2 A Hybrid Method for Estimation of Open Quotient Parameter

The open quotient OQ represents one of the voice source parameters related to the glottal opening phase. The glottal opening phase plays an important role in realizing various voice qualities of synthetic speech. Although the MSEE criterion is very effective in finding the optimal GCI's, the estimation of GOI's is not always accurate. This is mainly caused by the small amplitude of speech waveform at the GOI's.

In the proposed method, attentions have been concentrated on the harmonic components of the low frequency band as a possible error criterion to optimize the GOI's. The difference of the first ($H1$) and second ($H2$) harmonic components between the original speech and the predicted signal is selected as the criterion.

Firstly, it is necessary to investigate the relationship between $H1$ - $H2$ and the voice source parameters: AV, OQ, TL, and F_0 ($1/T_0$, fundamental frequency). Experiments were carried out with synthesized speech of the five Japanese vowels /a/, /i/, /u/, /e/ and /o/. The results showed that $H1 - H2$ mainly depended on OQ. Therefore, the difference of $H1 - H2$ between the original signal ($(H1 - H2)_{org}$) and the predicted signal ($(H1 - H2)_{est}$) can be used to achieve an approximate displacement between the optimal OQ (OQ_{opt}) and the estimated value OQ_{msee},

$$(H1-H2)_{org}-(H1-H2)_{est} = \alpha(F_0)*(OQ_{opt}-OQ_{msee}). \quad (8)$$

The optimal value of OQ becomes

$$OQ_{opt} = \gamma(F_0)*((H1-H2)_{org}-(H1-H2)_{est})+OQ_{msee}, \quad (9)$$

where $\gamma(F_0) = 1/\alpha(F_0)$. In the above experiments, the average value of $\gamma(F_0)$ between F_0=50 Hz and F_0=450 Hz is 0.018.

4 EXPERIMENTS AND DISCUSSIONS

In order to evaluate the performance of the new method, a two-channel signal, /aoiue/ ("blue top" in Japanese) uttered by a male, was analyzed. The results are shown in Fig. 4. The negative peaks of differentiated EGG (DEGG) signal in Fig. 4 (b) indicate that the GCI's, the negative peaks of RK glottal waveform in Fig. 4 (c), match well with the GCI's of DEGG signal. By comparing the DEGG signal with the RK waveform and the original speech with the re-syntheized speech, it is not unfair to say that even the other voice source parameters are also estimated correctly. On the other hand, the difference between the negative peaks of DEGG signals and those of RK glottal waveforms was around one sample point over the whole speech signal (74 pitch periods), which is shown in Fig. 5. For a male voice, one point

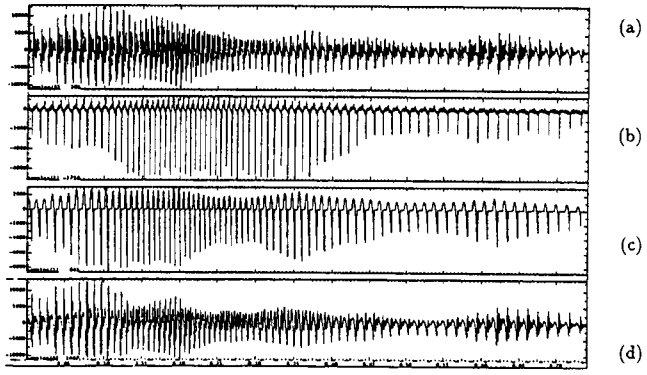

Figure 4: Results of natural speech /a-o-i-u-e/ using the proposed method. (a) speech waveform, (b) DEGG signal, (c) RK glottal waveform, and (d) re-synthesized speech waveform.

error of pitch period (about 1 Hz for 147 Hz average pitch) can be disregarded. However, for a female voice of about 300 Hz average pitch, one point error of pitch period may yield a perceived jitter in the synthesized speech. Therefore, for female voice and a low sampling rate, it is necessary to devise more detailed estimation of the GCI within two neighboring sampling points.

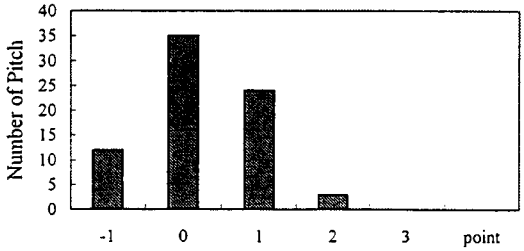

Figure 5: Difference between the negative peaks of the DEGG signal and GCI's of the RK glottal waveform ("–" means negative peak of RK comes after that of DEGG).

In a general form, an utterance including voiced, unvoiced, mixed-voice and silent segments, /muzkasii reedaio oboeru/ ("remember the difficult example" in Japanese) spoken by a Japanese male speaker, was also analyzed and the sound spectrograms of both original and re-syntheiszed sounds are shown in Fig. 6. By comparing the two sound spectrograms, it can be confirmed that voice source parameters as well as glottal noise amount parameter and formant trajectroies have been estimated accurately.

5 CONCLUSIONS

This paper proposed a novel pitch-synchronous method to deal with the joint estimation. It used an ARX model and a time varying IIR filter for representing the overall structure of speech production process. Without human interaction, the proposed method can achieve accurate estimation of RK parameters and formant/antiformant frequencies and bandwidths. The performance has been

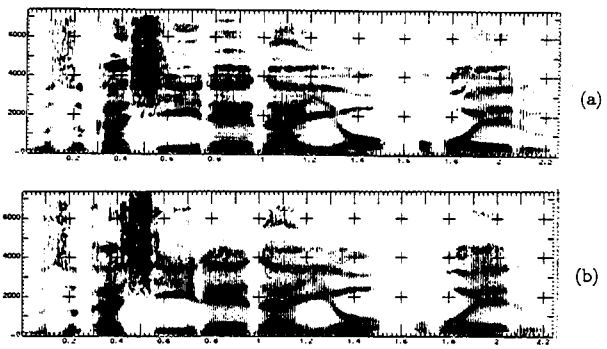

Figure 6: Results of /muzkasii reedaio oboeru/ (a) sound spectrogram of original speech, (b) sound spectrogram of re-synthesized speech.

verifed with EGG signal and inspiring results have been achieved for male voice. Moreover, the great potential ability of the proposed method in analyzing an utterance consisting of voiced, unvoiced and mix-voiced speech has also been shown.

References

[1] D. Klatt and L. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers," J. Acoust. Soc. Am., Vol. 87, pp. 820-857, 1990.

[2] I. Karlson, "Controlling Voice Quality of Synthetic Speech," Proc. ICSLP, Yokohama, pp. 1439-1442, 1994.

[3] P. Alku, "Glottal Wave Analysis with Pitch-synchronous Iterative Adaptive Inverse Filtering," Speech Communication, Vol. 11, pp. 109-118, 1992.

[4] P. Milenkovic, "Glottal Inverse Filtering by Joint Estimation of an AR System with a Linear Input Model," *IEEE Trans. ASSP*, Vol. 34, pp. 28-42, 1986.

[5] H. Fujisaki and M. Ljungqvist, " Estimation of Voice Source and Vocal Tract Parameters Based on ARMA Analysis and a Model for the Glottal Source Waveform," Proc. ICASSP, pp. 637-640,1987.

[6] G. Fant, *Acoustic Theory of Speech Production*, (Mouton, The Hague), 1960.

[7] S. Adachi, *System Identification Theory for Users* , (in Japanese) The Society of Instrument and Control Engineers, 1993.

[8] L. Ljung, *System Identification – Theory for the User* , Prentice Hall, 1987.

[9] W. J. Hess , Pitch Determination of Speech Signals - Algorithms and Devices, Springer-Verlag, Berlin, 1983.

[10] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous Estimation of Vocal Tract and Voice Source Parameters Based on an ARX Model", *IEICE Trans. Inf. & Syst.*, Vol. E78-D, No. 6, pp. 738-743, 1995.

Sound File References:

[a106s01.wav]

[a106s02.wav]