# FAST AND ROBUST JOINT ESTIMATION OF VOCAL TRACT AND VOICE SOURCE PARAMETERS

*Wen Ding*[1] , *Nick Campbell*[1] , *Norio Higuchi*[1] *and Hideki Kasuya*[2]

[1]ATR Interpreting Telecommunications Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
E_mail : ding@itl.atr.co.jp
[2]Faculty of Engineering, Utsunomiya University
2753 Ishii-machi, Utsunomiya 321, Japan
E_mail : kasuya@utsunomiya-u.ac.jp

## ABSTRACT

A new pitch-synchronous method of joint estimation is described to estimate vocal tract and voice source parameters from speech signals based on an ARX model. The method uses Kalman filtering to estimate the time-varying coefficients and simulated annealing to deal with the non-linear optimization of Rosenberg-Klatt parameters. A compact method is suggested in the algorithm in order to reduce the computation cost. Further, an automatic model order selection method is proposed to determine the proper analysis pole-order of the ARX model, based on the estimated formant bandwidths. The new method has been shown to be much faster than our previous method and the order selection technique has been shown to be effective. Finally, an ATR two-channel speech database including varying sentence-level prominence patterns is used to verify the proposed method.

## 1. INTRODUCTION

Joint estimation of vocal tract and voice source parameters has been shown to possess great potential applications in many research areas of speech processing, e.g., speech analysis / synthesis, perception of voice quality, speech coding, acoustic phonetics, etc. In a multi-language, multiple speaker translation system with a synthetic speech response, detection and reproduction of various prosodic features, such as intonation, speaking style and speaker characteristics are essential [1], [2], and joint estimation of vocal tract and voice source parameters may provide a powerful technique to detect, discriminate and generate these important features.

A large amount of effort has been put into measuring voicing source model parameters manually [3], [4], but these suffered from unstable subjective factors and the difficulty of processing a large speech database. Therefore, many methods of automatic estimation including glottal closure instant (epoch) detection have been suggested based on models such as adaptive inverse filtering [5], [6], two-channel analysis [7], glottal ARMA method [8], etc. However, most of these methods have been shown to have drawbacks and constraints in real applications, especially in continuous speech processing, [9], [10]. Generally speaking, because of the non-linear problem introduced by a glottal source model, most of these methods attempted to separate the processes of estimating vocal tract and voice source parameters.

The purpose of this paper is to devise a fast and robust pitch-synchronous joint estimation method. Joint estimation is realized by using Kalman filtering and simulated annealing. A fast and efficient technique is proposed here not only to speed up a previously proposed method [10] but also to provide an accurate estimation. Experimental results show improved estimation accuracy from the new model order detector. A corpus of 104 continuous speech conference-registration dialogues including different stress/prominence patterns is used to check the validity of the ARX method and to investigate the correlation between voice source parameters and perceived prominence.

## 2. ARX MODEL

Speech production process is represented as an auto-regressive model with an exogenous input (ARX):

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

where $s(n)$ and $u(n)$ denote an observed speech signal and an unknown input glottal waveform at time $n$, respectively. $a_i(n)$ and $b_j(n)$ are time-varying coefficients where $a_0(n) = 1$ and $b_0(n) = 1$. $p$ and $q$ are model orders, and $\varepsilon(n)$ is an equation error associated with the model. In the joint estimation, the input signal $u(n)$ is approximated by a Rosenberg-Klatt (RK) voicing source model [3]. There are four parameters of the RK model that need to be estimated: fundamental period $T_0$ $(1/F_0)$, amplitude of voicing $AV$, open quotient $OQ$ and spectral tilt $TL$.

The vector notation of the coefficients and data are represented as

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

and

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

The error criterion used in the nonlinear optimization procedure is the mean-square estimation error ($MSEE$) of

the ARX model,

$$E = \frac{1}{N}\sum_{n=1}^{N} \varepsilon(n)^2. \quad (4)$$

## 3. JOINT ESTIMATION

Figure 1 shows a flowchart of the proposed joint estimation method. The optimization algorithm is based on simulated annealing (SA) and is employed to search for the best set of voice source parameters to minimize $MSEE$. It is not practical to estimate the fundamental period parameter $T_0$ by SA, since $T_0$ is defined explicitly as the elapsed time between two successive glottal closure instants [11] which can be obtained from the estimated RK glottal waveforms. As the initial value, the average $T_0$ of the first frame in every voiced segment is computed, and then subsequently determined from the negative peaks of the estimated RK waveforms. The integration of Kalman filtering and simulated annealing with model order determination is explained in detail in the following subsections.

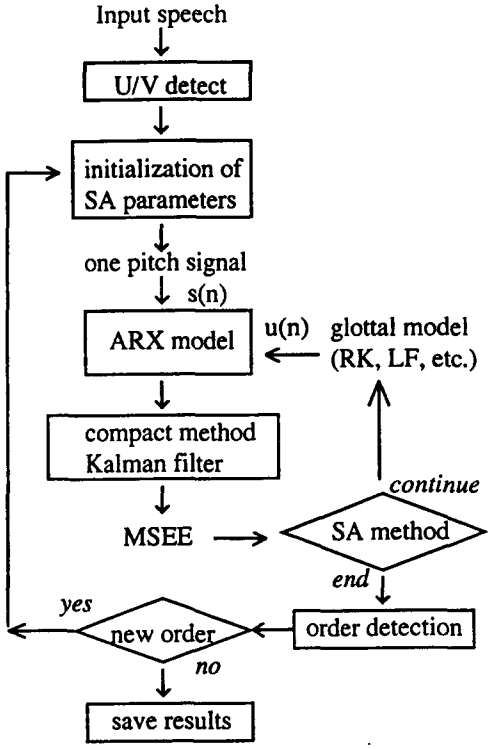


**Figure 1. Flowchart of the proposed joint estimation method.**

### 3.1. Algorithm

The algorithm for joint estimation is given in Table 1, based on the general structure described above. The method is implemented pitch-synchronously and the algorithm in the table describes the computation equations for one-pitch-period speech signals. The matrix $K(0)$ is only initialized at the beginning of the analysis as
$K(0) = E\{\theta(0)\theta(0)^T\} = diag(100, 100, \cdots, 100)$.

**Table 1. The novel joint estimation algorithm.**

1. **compute** the average pitch of the first 40ms speech in each voiced segments.
2. **get** speech signals of one pitch period $s(n)$ (n=1,$\cdots$, $T_0$).
3. **initialize** SA control variables and RK parameters
4. **generate** new RK parameters with SA.
5. **compute** $u(n)$ (n=1,$\cdots$, $T_0$) with the new parameters.
6. **for** n=1 to $T_0$,
   **if** the first SA loop,
   $(a)$ $g(n) = K(n-1)\varphi(n)[\varphi(n)^T K(n-1)\varphi(n) + 1]^{-1}$
   $(b)$ $\varepsilon(n) = s(n) - \varphi(n)^T\theta(n) - u(n)$
   $(c)$ $\theta(n+1) = \theta(n) + g(n)\varepsilon(n)$
   $(d)$ $K(n) = K(n-1) - g(n)\varphi(n)^T K(n-1)$
   $(e)$ $c(n) = \varphi(n)^T\theta(n)$
   **compute** $MSEE'$ based on eq. (4)
   **else**
   $(f)$ $\varepsilon(n) = c(n) - u(n)$
   **compute** $MSEE'$ based on eq. (4)
7. **check** the new RK parameters with SA based on the $MSEE'$.
8. **if accept** the new RK parameters
   **for** n=1 to $T_0$, **implement** eq. $(a)$, $(b)$, $(c)$, $(d)$, $(e)$ with current $u(n)$.
9. **if** SA loop continues, **goto** (4), **else for** $s(n)$ $(n = 1, 2, \cdots, T_0)$,
   **compute** $u(n)$ with RK parameters of the smallest $MSEE$, and
   **implement** eq. $(a), (b), (c), (d)$, and
   $(g)$ $K(n+1) = K(n) + \alpha * diag(K(n))$.
10. **compute** formant parameters from $\theta(n)$, and then **check** model order.
11. **save** parameters, and $T_0$ is set to the interval between current GCI and previous GCI, then **goto** (2).

The initial values of SA control variables and RK parameters are described in detail in [10].

In our previous method (old method) [10], when a new set of voice source parameters was generated in the SA loop, the equations of the Kalman filter from eq. $(a)$ to eq. $(e)$ were implemented to get a new $MSEE$. This computation is quite time consuming. Since most of the search iterations could be discarded without affecting $MSEE$, many iterations of the Kalman filter algorithm were performed fruitlessly.

The solution to this problem requires an improvement to the iterative computation of $MSEE$ instead of using the full-computing Kalman filter. This is because in the simple case, e.g., an all-pole version ($q$=0), the $\varepsilon(n)$ of eq. $(b)$ is never changed during the SA iteration of one pitch period. This indicates that the initial values of the current iteration of SA are set from the previously updated coefficients vector $\theta(n)$. The new $MSEE$ of the current loop can be computed by simply using eq. $(f)$. If the current $MSEE$ is smaller than the previous one, then the Kalman filter is used to update $\theta(n)$, otherwise we start the next SA loop with newly generated source parameters while keeping the

current $\theta(n)$. In such a way, the computation time can be cut exponentially. It is not difficult to expand this idea to the more general case ($p > 0$). Figure 2 shows an evaluation result of the time cost for one pitch period with an average $F_0$=100 Hz (10ms) on a SUN workstation.

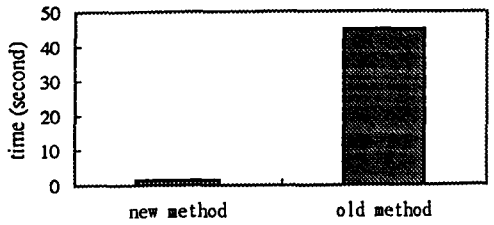


**Figure 2. Evaluation of estimation time.**

### 3.2. Model Order Selection

Using an incorrect pole-order usually causes an excess formant with a large bandwidth or a formant approximating the envelope of two spectral peaks in the low frequency domain. We assume that formant whose frequencies are below 3kHz should have a minimum average bandwidth value, and thereby estimate an optimal pole-order. The order selector monitors the estimated formant values only when the SA loop ends. If the difference of the average bandwidths is larger than a threshold $\xi$, an optimal order $p_{(opt)}$ will be selected from several defined candidates $P$ based on the following criterion,

$$p_{(opt)} = \min_{p \in P} \{ \frac{1}{N} \sum_{i=1}^{N} B_i^p, (F_i^p < 3000) \}. \qquad (5)$$

The benefits of this formant-driven rule are two-fold : (1) it has almost no effect on the computation time, and (2) it can remove any ghost formants. The algorithm recovers from ghost formant estimation errors by using an appropriate order or by reseting the coefficients if the order is not changed.

## 4. EXPERIMENTS AND DISCUSSIONS

Using model order selection, formant estimation errors can be checked immediately and eliminated by switching to a proper analysis order. Figure 3 shows a trajectory of the estimated first-five formant frequencies with and without order detection. The initial model order is $p$=12, $q$=0. The line marked with "x"shows the result without using order detection, where the first formant ($F1$) is mistakenly estimated (to near zero frequency) from pitch 13. On the other hand, using the order detection method (line marked "o"), when the difference of the average bandwidths compared with the previous pitch period became larger than the threshold $\xi$ at pitch 10, the model order was changed ($p$=14, $q$=0) based on eq. (5). This new order removed the $F1$ errors from pitch 13.

Figure 4 illustrates the trajectories of the estimated parameters of a Japanese sentence ("nimotsuwa koredake

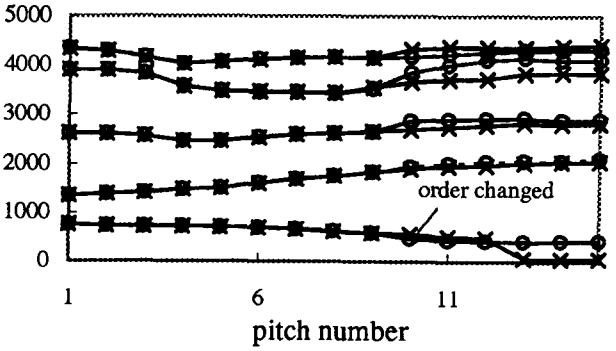


**Figure 3. The first five formant frequencies of /hai/, showing the difference between order selection ("o") and fixed-order estimation ("x").**

desuga"). The estimated values of unvoiced segments were excluded from the figure. The RK parameters and formant values are shown to be estimated stably. The $OQ$ and $TL$ values become larger near the end of the utterance, which may be due to a relaxation of the vocal cords.

A database of conference-registration dialogues including 104 Japanese sentences (speech and EGG signal) with shifting word prominence (showing different focal interpretations of the same utterance) was analyzed to verify the proposed method. Figure 5 shows the two-channel signal for /shu-dai/ in a Japanese utterance from the database. The glottal closure instant of the differentiated EGG (DEGG) signal is at the negative peak of each pitch period, and the glottal opening instant of the DEGG signal is located at the positive peak. It can be seen that the glottal opening instants and negative peaks of the RK glottal waveforms match well with those of the DEGG signals. A relationship between voice source parameters and the perceived prominence was found using both extraction methods as shown in Fig. 6.

## 5. CONCLUSIONS

A new pitch-synchronous joint estimation method has been proposed, in which we estimate the RK model parameters and formant values from speech signals. This method represents an improvement to our previously proposed method. Improved results have been achieved using a formant-driven order detection rule based on average minimum bandwidths. We use a compact method derived from simulated annealing and Kalman filtering, and the algorithm has been shown to cut the time cost. The proposed method has been applied to a two-channel speech database, without human interaction, and shown to achieve reliable estimates of voice source and vocal tract parameters.

## REFERENCES

[1] W.N. Campbell, "Prosodic influence on segmental quality", *Proc. ESCA*, Madrid, pp. 1011-1014, 1995.

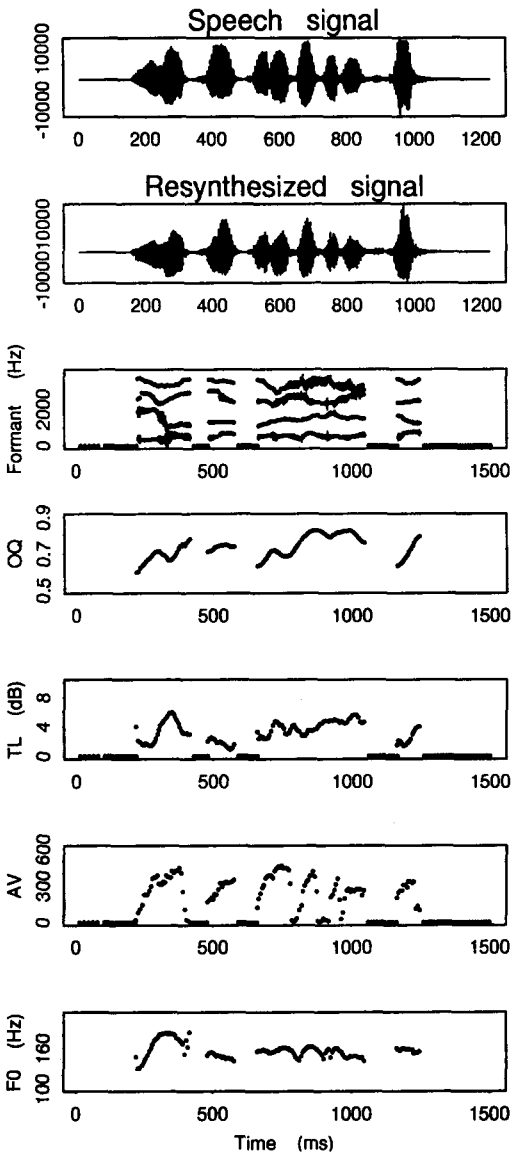


**Figure 4. Results of the sentence /nimotsuwa koredake desuga/.**

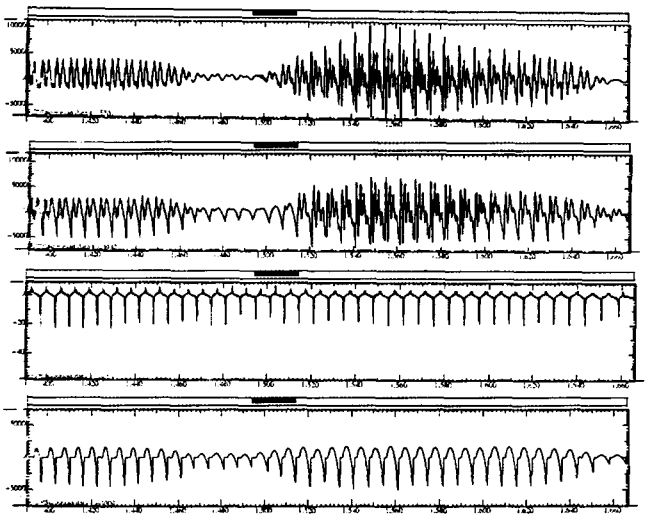

**Figure 5. Comparing EGG with estimated RK glottal waveform for /shu-dai/. Top to bottom: original speech, resynthesized speech, DEGG signal and estimated RK glottal waveform.**

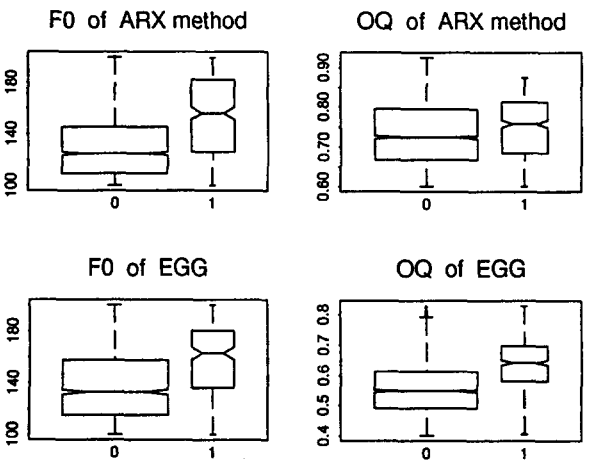


**Figure 6. Relationship between $F_0$, $OQ$ and prominence with the ARX method and EGG-based method. ("1" indicates prominence and "0" non-prominence.)**


[2] M. Hashimoto and N. Higuchi, "Spectral mapping for voice conversion using speaker selection and vector field smoothing", *Proc. ESCA*, Madrid, pp. 431-434, 1995.

[3] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male speakers", *J. Acoust. Soc. Am.*, Vol. 87, pp. 820-857, 1990.

[4] I. Karlson, "Controlling voice quality of synthetic speech", *Proc. ICSLP*, Yokohama, pp. 1439-1442, 1994.

[5] P. Alku, "Glottal wave analysis with pitch-synchronous iterative adaptive inverse filtering", *Speech Communication*, Vol.11, pp. 109-118, 1992.

[6] 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.

[7] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis", *IEEE Trans. ASSP*, Vol. 24, pp. 730-740, 1986.

[8] 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.

[9] T. V. Ananthapadmanabha, STL-QPSR, Vol. 2-3, pp. 1-24, 1984.

[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.

[11] W. J. Hess, *Pitch determination of speech signals - algorithms and devices*, Springer-Verlag, Berlin, 1983.