Abstract—Fundamental frequency ($F_0$) estimation is important in speech processing such as speech coding, synthesis, recognition and so on. A present $F_0$ estimation method performs well under clean condition, however the performance deteriorates significantly in noisy environment. As a result, robust $F_0$ estimation against additive noise is demanded. We have previously proposed $F_0$ estimation methods based on Time-Varying Complex AR (TV-CAR) analysis whose criterion is the weighted correlation of the complex residual obtained by the TV-CAR analysis, sum of the harmonics for the complex residual spectrum, or so on. On the other hand, E.Azarov et al. have proposed an improved method of RAPT (Robust Algorithm for Pitch Tracking) using an instantaneous harmonics that is called IRAPT (Instantaneous RAPT). The IRAPT can perform better estimation than RAPT. Since IRAPT uses band-limited analytic signal to obtain harmonic frequencies, the complex residual signal obtained by the TV-CAR analysis can also be applied to the IRAPT. In this paper, novel $F_0$ estimation method using the instantaneous frequency based on the robust ELS (Extended Least Square) TV-CAR residual is proposed and evaluated.

I. INTRODUCTION

$F_0$ estimation still remains unsolved and challenging problem. It is one of the key technologies that determines the performance of speech processing. Autocorrelation method or Average Magnitude Difference Function (AMDF) are commonly used[1]. In addition, modified autocorrelation method[2] is also commonly used in which Linear Prediction (LP) residual is used to compute the autocorrelation instead of speech signal. The residual provides less formant components than speech signal does since residual is a output from an inverse filtering of LP. For this reason the method can estimate more accurate $F_0$. However, these methods suffer from error estimation in noisy environment. Thus, robust $F_0$ estimation against additive noise has been studying for a few decades.

We have been studying the robust $F_0$ estimation[3][4][5][6] based on time-varying complex AR (TV-CAR) speech analysis[7][8]. [3] and [4] adopt an weighted autocorrelation of the complex AR residual derived by the TV-CAR speech analysis for an analytic signal. An analytic signal is complex signal whose real part is an observed signal and whose imaginary part is the Hilbert transform of the real one. An analytic signal provides a spectrum only over positive frequency. For this reason, complex speech analysis can estimate more precise spectrum in low frequencies. In addition, formant structure except $F_0$ can be removed appropriately on complex residual than on real residual so that more accurate $F_0$ estimation can be realized by using complex one. Furthermore, we have reported that more accurate $F_0$ estimation can be realized by using complex residual using SRH(Summation of Residual Harmonic)[5].

On the other hand, RAPT (Robust Algorithm for Pitch Tracking) algorithm[9] is well-known and widely used $F_0$ estimation method since it does offer low delay, low computational amount and robust against noise, furthermore it is open source software. The improved RAPT named as IRAPT (Instantaneous RAPT) has been proposed that introduces instantaneous harmonics[10]. In [10], $F_0$ estimation is realized by using instantaneous frequency instead of NCCF (Normalized Cross Correlation Function) and input speech is time-warped by the estimated $F_0$ and $F_0$ is re-estimated by using the time-warped speech. Instantaneous harmonic in IRAPT method is calculated from speech, however the complex residual estimated by the TV-CAR analysis can be applied. Formant structure except $F_0$ can be removed on complex residual appropriately so that more accurate estimation can be expected by using the complex residual for IRAPT.

In this paper, $F_0$ estimation based on the IRAPT using complex residual estimated by the TV-CAR analysis is presented. Robust ELS-based TV-CAR analysis is introduced to obtain the complex residual. Performance are compared using Keele pitch database [11] among speech, the time-varying real-valued AR residual and time-varying complex AR residual based on MMSE(Minimizing Mean Squared Error)[7] or ELS[8].

The remainder of this paper is organized as follows. In Session II, the IRAPT method is explained. In Session III, the TV-CAR speech analysis based on MMSE or ELS is explained. In Session IV, the proposed algorithm is explained. In Session V, experimental results are shown and discussed.

II. IRAPT ALGORITHM

Figure 1 shows block diagram of IRAPT[10]. The procedure is shown as follows. In block (1), down sampling to 6KHz is operated. In block (2), instantaneous frequencies are estimated from band limited analytic signal. In block (3), the duplicated values on the harmonic frequencies are discarded. In block (4), $F_0$ Candidate Generating function corresponding to the autocorrelation function is computed. In block (5), the best $F_0$ candidate is selected by using Dynamic Programming (DP).
like RAPT. In block (6), the down sampled signal is time-warp by using all-pass filter with the estimated $F_0$. In block (7), the instantaneous frequencies are estimated by using the time-warped signal like the block (2). In block (8), $F_0$ is estimated by using the instantaneous frequencies in block (7). In Figure 1, the method for $F_0$ estimation 1 (block (1) to (5)) is called IRAPT1 and the method for $F_0$ estimation 2 (block (1) to (8)) is called IRAPT2.

![Source signal](image)

**Figure 1:** Block diagram of IRA

### III. TV-CAR Speech Analysis

#### A. Analytic speech signal

Target signal of the time-varying complex AR (TV-CAR) method is an analytic signal that is complex-valued signal defined by

$$
y^c(t) = \frac{y(2t) + j \cdot y_H(2t)}{\sqrt{2}}
$$

where $y^c(t)$, $y(t)$, and $y_H(t)$ denote an analytic signal at time $t$, an observed signal at time $t$, and a Hilbert transformed signal for the observed signal, respectively. Notice that superscript $c$ denotes complex value in this paper. Since analytic signals provide the spectra only over the range of $(0, \pi)$, analytic signals can be decimated by a factor of two. $2t$ means the decimation and the term of $1/\sqrt{2}$ is multiplied in order to adjust the power of an analytic signal with that of the observed one.

#### B. Time-varying complex AR (TV-CAR) model

The TV-CAR model is defined as

$$
a^c_i(t) = \sum_{l=0}^{L-1} g^c_{i,l}f^c_l(t)
$$

$$
Y_{TV\text{CAR}}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I} a^c_i(t)z^{-i}}
$$

$$
Y_{TV\text{CAR}}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I} \sum_{l=0}^{L-1} g^c_{i,l}f^c_l(t)z^{-i}}
$$

where $I$ is AR order. The input-output relation is defined as

$$
y^c(t) = -\sum_{i=1}^{I} a^c_i(t)y^c(t-i) + u^c(t)
$$

$$
y^c(t) = -\sum_{i=1}^{I} \sum_{l=0}^{L-1} g^c_{i,l}f^c_l(t)y^c(t-i) + u^c(t)
$$

where $u^c(t)$ and $y^c(t)$ are taken to be complex-valued input and analytic speech signal, respectively. In the TV-CAR model, the complex AR coefficient is modeled by a finite number of arbitrary complex basis. Note that Eq.(2) parameterizes the AR coefficient trajectories that continuously change as a function of time so that the time-varying analysis is feasible to estimate continuous time-varying speech spectrum. In addition, as mentioned above, the complex-valued analysis facilitates accurate spectral estimation in the low frequencies, as a result, this feature allows for more accurate $F_0$ estimation if formant structure is removed by the inverse filtering. Eq.(4) can be represented by vector-matrix notation as

$$
\tilde{y}_f = -\Phi_f \hat{\theta} + \tilde{u}_f
$$

$$
\tilde{y}_f = [\tilde{y}_f^T, \tilde{y}_f^T, \cdots, \tilde{y}_f^T, \cdots, \tilde{y}_f^T]_1
$$

$$
\tilde{g}_f^c = [\tilde{g}_f^c, \tilde{g}_f^c, \cdots, \tilde{g}_f^c, \cdots, \tilde{g}_f^c]_1
$$

$$
\tilde{y}_f^c = [y^c(I), y^c(I+1), y^c(I+2), \cdots, y^c(N-1)]
$$

$$
\tilde{u}_f^c = [u^c(I), u^c(I+1), u^c(I+2), \cdots, u^c(N-1)]
$$

$$
\Phi_f = [\Phi_f^T, \Phi_f^T, \cdots, \Phi_f^T, \cdots, \Phi_f^T]_1
$$

$$
\Phi_f^c = [\Phi_f^c, \Phi_f^c, \cdots, \Phi_f^c, \cdots, \Phi_f^c]
$$

$$
\hat{\theta} = \sum_{i=1}^{I} \sum_{l=0}^{L-1} \tilde{g}^c_{i,l}f^c_l(I), y^c(I+1-i)f^c_l(I+1), \cdots, y^c(N-1-i)f^c_l(N-1)]^T
$$

where $N$ is analysis interval, $\tilde{y}_f$ is $(N-I, 1)$ column vector whose elements are analytic speech signal, $\hat{\theta}$ is $(L \cdot I, 1)$ column vector whose elements are complex parameters, $\Phi_f$ is $(N-I, L \cdot I)$ matrix whose elements are weighted analytic speech signal by the complex basis. Superscript $T$ denotes transposition.

#### C. MMSE-based algorithm [7]

MSE criterion is defined by

$$
r_f = [r^c(I), r^c(I+1), \cdots, r^c(N-1)]^T
$$

$$
r_f = \tilde{y}_f + \Phi_f \hat{\theta}
$$

$$
y^c(t) = y^c(t) + \sum_{i=1}^{I} \sum_{l=0}^{L-1} \tilde{g}^c_{i,l}f^c_l(t)y^c(t-i)
$$

$$
E = \tilde{r}_f^H r_f = (\tilde{y}_f + \Phi_f \hat{\theta})^H (\tilde{y}_f + \Phi_f \hat{\theta})
$$

where $\tilde{g}^c_{i,l}$ is the estimated complex parameter, $r^c(t)$ is an equation error, viz., complex AR residual and $E$ is Mean Squared Error (MSE) for the equation error. To obtain optimal complex AR coefficients, we minimize the MSE criterion.
Minimizing the MSE criterion of Eq.(8) with respect to the complex parameter leads to the following MMSE algorithm.

\[
\hat{\Phi}_{ij}^H \hat{\Phi}_{ij} \hat{\theta} = \bar{\Phi}_{ij}^H \bar{y}_j
\]  

Superscript \(H\) denotes Hermitian transposition. After solving the linear equation of Eq.(9), we can get the complex AR parameter \(\hat{\alpha}_i(t)\) at time \(t\) by calculating the Eq.(2) with the estimated complex parameter \(\hat{g}_{ci,ti}\).

### D. ELS-based algorithm [8]

If the equation error shown as in Eq.(7) is white Gaussian, the MMSE estimation is optimal, however, it is rare case. As a result, MMSE estimation suffers from biased estimation. In the ELS method, an AR filter is adopted to whiten the equation error as follows.

\[
r^c(t) = - \sum_{k=1}^{L} b_k^c r^c(t-k) + e^c(t)
\]  

where \(b_k^c\) is \(k\)-th parameter of the AR filter whose order is \(K\) and \(e^c(t)\) is 0-mean white Gaussian of equation error at time \(t\). The inverse filter of Eq.(10) is called a white filter. The TV-CAR model can be represented using Eqs.(4) and (10) as follows.

\[
y^c(t) = - \sum_{i=1}^{J} \sum_{l=0}^{L-1} g_{i,l}^c f_i^c(t) y^c(t-i)
\]  

\[- \sum_{k=1}^{K} b_k^c r^c(t-k) + e^c(t)
\]

Eq.(11) is the ELS model. The parameter is estimated so as to minimize the MSE for the whitened equation error in the ELS algorithm whereas the parameter is estimated so as to minimize the MSE for the equation error in the MMSE algorithm.

### IV. PROPOSED METHOD

As mentioned above, the IRAPT method uses speech signal to estimate \(F_0\). Since formant structure can be removed by inverse filter, the residual signal is more similar than speech signal for \(F_0\) estimation. It is expected that \(F_1\) can be removed more appropriately by the complex AR analysis for analytic signal than by real-valued analysis since the complex analysis can resolve \(F_0\) and \(F_1\) owing to its higher frequency resolution in low frequencies. Furthermore, formant structure can be removed more appropriately by using a time-varying analysis with the estimated AR coefficient every sample. It is calculated by Eq.(7).

In the proposed method, complex AR residual estimated by the TV-CAR analysis from analytic signal is used as a source signal denoted in Figure 1 instead of speech. The MMSE-based TV-CAR analysis and robust ELS-based TV-CAR analysis are used to compute the residual and are compared by means of \(F_0\) estimation accuracy. The complex residual is evaluated as half sampling rate of the real-valued input signal.

### V. EXPERIMENTS

As the source signal to IRAPT in Figure 1, the following four signals are evaluated. 1) Speech signal, 2) LPC residual, 3) the complex residual by MMSE-based TV-CAR, 4) the complex residual by ELS-based TV-CAR. Those are obtained by the TV-CAR analysis. The downloaded Matlab program from [12] site is used. Experimental conditions are shown in Table 1. The experiments were carried out with Keele Pitch Database[11] corrupted by white Gauss or Pink noise[13]. Note that the Keele Pitch database contains five long sentences uttered by different five male speakers and five long sentences uttered by five female speakers. Each sentence is more than 30 seconds long and total length is around 6 minutes. The noise corrupted speech is filtered by an IRS(Intermediate Reference System) filter[14] for speech coding application. \(F_0\) estimation performance is evaluated by using GPE(Gross Pitch Error) and FPE(Fine Pitch Error). Figure 2 is the result for additive white Gaussian noise, and Figure 3 is the result for additive Pink noise. In figures, (1) means a graph of GPE of 10\%[10] and (2) means that of FPE of 10\%. X-axis denotes noise level [dB], Y-axis denotes GPE\% in (1), FPE\% in (2). Four lines of the figures are as follows.

(a) \(\text{IRAPT2 (square solid line)}\) is GPE/FPE for the speech of IRAPT2[10].
(b) \(\text{TVR}_\text{IRAPT2} \) (diamond solid line) is GPE/FPE of IRAPT2 for the real-valued AR residual estimated by the time-varying real-valued analysis.
(c) \(\text{TVC}_\text{IRAPT2C}_\text{ELS} \) (red line) is GPE/FPE of IRAPT2 for the complex AR residual estimated by the ELS-based TV-CAR analysis.
(d) \(\text{TVC}_\text{IRAPT2C} \) (blue line) is GPE/FPE of IRAPT2 for the complex AR residual estimated by the MMSE-based TV-CAR analysis.

Figures 2 and 3 demonstrate as follows. The performance of the proposed method based on the MMSE or ELS is much better than that of SRH-based method[5] and YIN[15]. In [5], the exactly same data were used. GPE is considerably improved by introducing residual compared to original IRAPT2 and the complex residual can improve more than the real-valued AR residual. The MMSE-based TV-CAR analysis performs better than ELS-based analysis. Improvement by complex residual is remarkable for white Gaussian noise. However, in the case of Pink noise, a little improvement can be seen. FPE improves for both white Gauss noise and Pink noise than the original IRAPT2. However, when the Pink noise level is high, the real AR residual performs better. On the other hand, the ELS method can improve the FPE for the case. The initial estimation of ELS is MMSE and the difference between them is iteration number. Consequently, best performance can be obtained by controlling the iteration number. The reason why the ELS analysis cannot perform better is that pre-emphasis is not adequate and \(F_0\) components are removed.
Table 1: Experimental conditions

<table>
<thead>
<tr>
<th>Speech data</th>
<th>Keele Pitch Database[11]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 long Male sentences</td>
</tr>
<tr>
<td></td>
<td>5 long Female sentences</td>
</tr>
<tr>
<td>Sampling</td>
<td>10kHz/16bit</td>
</tr>
<tr>
<td>Analysis window</td>
<td>Window Length: 25.6[ms]</td>
</tr>
<tr>
<td></td>
<td>Shift Length: 10.0[ms]</td>
</tr>
<tr>
<td>Complex-valued AR</td>
<td>$I = 7, L = 2$ (time-varying)</td>
</tr>
<tr>
<td>Basis</td>
<td>$f_c(t) = t^l / l!$</td>
</tr>
<tr>
<td>Pre-emphasis</td>
<td>$1 - z^{-1}$</td>
</tr>
<tr>
<td>Real-valued AR</td>
<td>$I = 14, L = 2$ (time-varying)</td>
</tr>
<tr>
<td>Basis</td>
<td>$f_c(t) = t^l / l!$</td>
</tr>
<tr>
<td>Pre-emphasis</td>
<td>$1 - z^{-1}$</td>
</tr>
<tr>
<td>ELS analysis</td>
<td>$K = 5$</td>
</tr>
<tr>
<td>Noise</td>
<td>White Gauss or Pink noise[13]</td>
</tr>
<tr>
<td>Noise Level</td>
<td>30,20,10,5,0,-5[dB]</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

We have proposed the improved IRAPT using the TV-CAR residual. The complex residual is introduced as a input of IRAPT. Evaluation using IRS filtered Keele pitch database corrupted by white Gauss noise or Pink noise has been carried out. According to the experimental results, we can find out that the performance of IRAPT were improved remarkably by using complex residual signals. The MMSE-based TV-CAR analysis can perform best, however, for only pink noise whose noise level is high, the FPE is not better. The ELS-based analysis can perform better for this drawbacks. To introduce adaptive pre-emphasis is future study.

REFERENCES


