On Evaluation of the $F_0$ estimation based on time-varying complex speech analysis

Keiichi FUNAKI
Computing & Networking Center, University of the Ryukyus
funaki@cc.u-ryukyu.ac.jp

Abstract

We have already proposed a robust fundamental frequency ($F_0$) estimation based on robust ELS (Extended Least Square) time-varying complex-valued speech analysis for an analytic speech signal. It has been reported that the method performs better for IRS filtered speech corrupted by white Gauss noise or pink noise since speech spectrum can be accurately estimated in low frequencies. However, the evaluation was performed by using only time-invariant speech analysis, in which order of basis expansion was 1. In this paper, the performance of time-varying speech analysis is evaluated using Keele pitch database with respect to degree of voiced stationarity of frame. The evaluation demonstrates that the time-varying ELS-based robust complex analysis performs best for strong stationary voiced frame although it does not perform better for non-stationary voiced frame.

Index Terms: $F_0$ estimation, complex speech analysis, time-varying analysis, analytic signal

1. Introduction

An $F_0$ estimation plays an important role in speech processing such as speech coding, tonal speech recognition, speaker recognition, and speech enhancement. Needless to say, the $F_0$ estimation error results in a degradation of the performance. Speech processing is commonly applied in realistic noisy environment, hence, the performance is degraded seriously. Accordingly, the robust $F_0$ estimation is long lasting problem in speech processing and more robust algorithm is desired. We have already proposed robust $F_0$ estimation algorithm based on time-varying complex speech analysis for analytic speech signal [1][2]. Analytic signal is a complex-valued signal in which its real part is speech signal and its imaginary part is Hilbert transform of the real part. Since the analytic signal provides the spectrum only on positive frequencies, the signals can be decimated by a factor of two with no degradation. As a result, the complex analysis offers attractive features, for example, more accurate spectral estimation in low frequencies. In [1] and [2], complex LPC residual is used to calculate the criterion of weighted autocorrelation function (AUTOC) with a reciprocal of AMDF function[5]. The complex residual is calculated from analytic speech signal by means of time-varying complex AR (TV-CAR) speech analysis method [3][4]. In [1], MMSE-based TV-CAR speech analysis[3] is introduced and in [2], ELS-based TV-CAR speech analysis[4] is introduced to calculate complex LPC residual signal. It has been reported in [1] that the method can estimate more accurate $F_0$ for IRS (Intermediate Reference System) filtered speech corrupted by white Gauss noise. Moreover, it has been reported in [2] that the ELS-based complex speech analysis can perform better even for additive pink noise. However, these were the results for time-invariant speech analysis and we have never evaluated performance of the time-varying analysis. It could be expected that time-varying analysis can offer more accurate $F_0$ estimation. In this paper, the evaluation is carried out by comparing the time-varying speech analysis with the time-invariant analysis using Keele Pitch Database[6] with respect to degree of voiced stationarity.

2. TV-CAR Speech Analysis

2.1. 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. The term of $1/\sqrt{2}$ is multiplied in order to adjust the power of an analytic signal with that of the observed one.

2.2. Time-varying complex AR (TV-CAR) model

Conventional LPC model is defined by

$$Y_{LPC}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I} a_i z^{-i}}$$

where $a_i$ and $I$ are $i$-th order LPC coefficient and LPC order, respectively. Since the conventional LPC model cannot express the time-varying spectrum, LPC analysis cannot extract the time-varying spectral features from speech signal. In order to represent the time-varying features, the TV-CAR model employs a complex basis expansion shown as

$$a_t^i(t) = \sum_{l=0}^{L-1} g_t^i f_l^i(t)$$

where $a_t^i(t)$, $I$, $L$, $g_t^i$ and $f_l^i(t)$ are taken to be $i$-th complex AR coefficient at time $t$, AR order, finite order of complex basis ex-
pansion, complex parameter, and a complex-valued basis function, respectively. By substituting Eq.(3) into Eq.(2), one can obtain the following transfer function.

\[ Y_{TV-CAR}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I} \sum_{l=0}^{L-1} \hat{g}_{i,l} f_l(t) z^{-l}} \]  (4)

The input-output relation is defined as

\[ y_f(t) = -\sum_{i=1}^{I} a_i(t)y_f(t-i) + u_f(t) \]

\[ y_f(t) = -\sum_{i=1}^{I} \sum_{l=0}^{L-1} \hat{g}_{i,l} f_l(t)y_f(t-i) + u_f(t) \]  (5)

where \( u_f(t) \) and \( y_f(t) \) 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.(3) 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.(5) can be represented by vector-matrix notation as

\[ \begin{align*}
\hat{y}_f &= -\Phi_f \bar{\theta} + \bar{u}_f \\
\bar{\theta}^T &= [\bar{g}_{f,1}^T, \bar{g}_{f,2}^T, \cdots, \bar{g}_{f,I}^T] \\
\bar{g}_{f,i}^T &= [\bar{g}_{i,1}^T, \bar{g}_{i,2}^T, \cdots, \bar{g}_{i,L}^T] \\
\bar{y}_f^T &= [y_f(I), y_f(I+1), y_f(I+2), \cdots, y_f(N-1)] \\
\bar{u}_f^T &= [u_f(I), u_f(I+1), u_f(I+2), \cdots, u_f(N-1)] \\
\bar{\Phi}_f &= [\bar{D}_f^1, \bar{D}_f^2, \cdots, \bar{D}_f^I] \\
\bar{D}_f^i &= [\bar{d}_{i,1}^T, \bar{d}_{i,2}^T, \cdots, \bar{d}_{i,L}^T] \\
\bar{d}_{i,l}^T &= [y_f(I-i)f_l^*(I), y_f(I-i+1)f_l^*(I+1), \cdots, y_f(N-1-i)f_l^*(N-1)]^T
\end{align*} \]  (6)

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

2.3. MMSE-based algorithm[3]

MSE criterion is defined by

\[ \bar{r}_f = [r_f^*(I), r_f^*(I+1), \cdots, r_f^*(N-1)]^T \]

\[ \bar{r}_f = \hat{y}_f + \bar{\Phi}_f \bar{\theta} \]  (7)

\[ r_f(t) = y_f^*(t) + \sum_{i=1}^{I} \sum_{l=0}^{L-1} \hat{g}_{i,l} f_l^*(t)y_f^*(t-i) \]  (8)

\[ E = \bar{r}_f^H \bar{r}_f = (\hat{y}_f + \bar{\Phi}_f \bar{\theta})^H (\hat{y}_f + \bar{\Phi}_f \bar{\theta}) \]  (9)

where \( \hat{g}_{i,l} \) is the estimated complex parameter, \( r_f^*(t) \) is an equation error, or 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.(9) with respect to the complex parameter leads to the following MMSE algorithm.

\[ (\hat{\Phi}_f^T \hat{\Phi}_f) \bar{\theta} = -\hat{\Phi}_f^T \hat{y}_f \]  (10)

Superscript H denotes Hermitian transposition. After solving the linear equation of Eq.(10), we can get the complex AR parameter \( a_i(t) \) at time \( t \) by calculating the Eq.(3) with the estimated complex parameter \( \hat{y}_f \).

2.4. ELS-based algorithm[4]

In the ELS algorithm, MMSE equation error is whitened by introduced AR filter and the parameter is estimated so as to minimize the whitened MMSE equation error. The ELS method can estimate more robust speech spectrum than MMSE algorithm, as a result, more robust \( F_0 \) estimation can be realized by ELS method [2].

3. \( F_0 \) Estimation Method


Autocorrelation function (AUTOC) is defined by

\[ f(\tau) = \frac{1}{N} \sum_{t=0}^{N-1} x(t)x(t+\tau) \]  (11)

where \( x(t) \) is target signal such as speech signal, LPC residual or so on, \( N \) is frame length and \( \tau \) means delay. \( F_0 \) is selected as peak frequency for Eq.(11) within certain range of \( F_0 \).

AMDF is defined as follows.

\[ p(\tau) = \frac{1}{N} \sum_{t=0}^{N-1} |x(t)-x(t+\tau)| \]  (12)

\( F_0 \) is selected as notch frequency for Eq.(12) within certain range of \( F_0 \).

In Shimamura method [5], the AUTOC is weighted by a reciprocal of the AMDF shown as Eq.(13). Since the weighting makes it possible to suppress other peaks, the method can estimate more accurate \( F_0 \) than AUTOC or AMDF. The value of \( m \) is set to be 1 in order to avoid the value of 0 at the denominator.

\[ G(\tau) = \frac{f(\tau)}{p(\tau) + m} \]  (13)

where \( f(\tau) \) and \( p(\tau) \) are AUTOC shown as in Eq.(11) and AMDF shown as in Eq.(12), respectively.

3.2. Proposed Method

In this paper, Shimamura criterion shown as Eq.(13) is applied to complex AR residual extracted by the robust TV-CAR speech analysis. The time-varying complex parameter is estimated and the complex AR residual is calculated with the estimated complex parameter with Eq.(8). Note that pre-emphasis is operated for speech analysis such as real-valued AR or TV-CAR speech analysis, and inverse filtering is applied for the non-pre-emphasized speech signal so as not to eliminate \( F_0 \) spectrum on the residual signal. Real part of AUTOC is used to calculate the AUTOC for complex-valued signal.
4. Experiments

Speech signals used in the experiment are 5 long sentences uttered by male speaker and 5 long sentences uttered by female speaker of Keele pitch database[6]. The speech signals are filtered by an IRS filter [7]. The IRS filter is band pass FIR filter whose frequency response corresponds to that for analog part of the transmitter of telephone equipment. In order to evaluate the proposed method for the speech data processed by speech coding, the IRS filter has to be introduced as in [1]. The experimental conditions are summarized in Table 1. Frame length is 25.6[msec] and frame shift length is 10[msec]. Analysis orders are 14 and 7 for real-valued analysis and complex-valued analysis, respectively. The basis expansion order L is set to be 1(time-invariant) or 2(time-varying) in the experiments. First order polynomial function is adopted as a basis function. White Gaussian noise or pink noise [8] is adopted for additive noise and the levels are 30, 20, 10, 5, 0, and -5 [dB]. In order to extract more accurate $F_0$, 3-point Lagrange’s interpolation is adopted.

Commonly used criterion for $F_0$ estimation, Gross Pitch Error(GPE), is adopted for objective evaluation. $F_0$ estimation error is defined as

$$e_p(n) = F_e(n) - F_l(n)$$

where $F_e(n)$ is true $F_0$ value and $F_l(n)$ is the estimated one. The true $F_0$ values are derived by pitch file in Keele database. In Eq.(14), if $|e_p(n)| > T \times \frac{|F_e(n)|}{100}$ then the estimation error is regarded as ERROR and GPE is the probability of the error frames. Otherwise, the estimation is regarded as SUCCESS and FPE is standard deviation of the error. Figures 1,2,3, and 4 show the experimental results setting the $T \times \frac{|F_e(n)|}{100}$. In Figures, (1) show the results of GPEs for additive white Gaussian noise. (2) show the results of GPEs for additive pink noise.

Figures 1 show the results for all voiced frames. In order to investigate the effectiveness of each method for degree of voiced stationarity of frame, voiced frame is categorized into 3 modes as follows. Mode 3 is the strongest voiced frame whose pitch prediction gain is larger than 9[dB]. Mode 2 is the ordinary voiced frame whose pitch prediction gain is larger than 6[dB] and less than 9[dB]. Mode 1 is weak voiced frame whose pitch prediction gain is less than 6[dB]. Pitch prediction gain PG is calculated by

$$PG = 10 \cdot \log_{10} \left( \frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} x(n)^2 - \sum_{n=0}^{N-1} x(n-T_0) x(n-T_0)} \right)$$

where $T_0$ is accurate fundamental period prepared in Keele Pitch Database. Figures 2 show the results for mode 3 frames. Figures 3 show the results for mode 2 frames. Figures 4 show the results for mode 1 frames.

$\text{SP}$ (dotted line) means the results for speech, $\text{AN}$ (dotted line) means that for analytic signal, $\text{LPC}$ (red dotted line) means that for real AR residual with time-invariant MMSE-based speech analysis, $\text{TVR}$ (blue dotted line) means that for real AR residual with time-invariant ELS-based speech analysis, $\text{LPC}_E$ (magenta dotted line) means that for real AR residual with time-invariant ELS-based speech analysis, $\text{TVR}_E$ (green dotted line) means that for real AR residual with time-invariant ELS-based speech analysis, $\text{CLPC}$ (red solid line) means that for complex AR residual with time-invariant MMSE-based speech analysis, $\text{TVC}$ (blue solid line) means that for complex AR residual with time-invariant MMSE-based speech analysis, $\text{CLPC}_E$ (magenta solid line) means that for complex AR residual with time-invariant ELS-based speech analysis, $\text{TVC}_E$ (green solid line) means that for complex AR residual with time-invariant ELS-based speech analysis. Note that SP means the Shimamura method [5], viz., Shimamura criterion for speech signal In all figures, X-axis means noise level of 30, 20, 10, 5, 0, -5[dB]. Y-axis means GPE[%].

Figure 1 demonstrates that the time-varying speech analysis does not perform well while ELS-based robust time-invariant complex speech analysis ($\text{CLPC}_E$) performs best in terms of GPE. Figures 2, 3, and 4 show the very interesting results. $\text{TVC}_E$ performs better in mode 3. $\text{CLPC}_E$ performs slightly better in mode 2. $\text{TVC}$ performs better in mode 1. According to the results, we can conclude as follows. (1)Time-varying analysis does not perform well for ordinary voiced segment. (2)ELS based robust time-varying complex speech analysis can perform better for strong stationary voiced segment. However, the performance is not so high. Evaluation for each phoneme is future study.

<table>
<thead>
<tr>
<th>Table 1: Experimental Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech data</td>
</tr>
<tr>
<td>Male 5 long sentences</td>
</tr>
<tr>
<td>Female 5 long sentences</td>
</tr>
<tr>
<td>IRS filter</td>
</tr>
<tr>
<td>Target signal</td>
</tr>
<tr>
<td>(2)real AR residual</td>
</tr>
<tr>
<td>(3)analytic speech signal</td>
</tr>
<tr>
<td>(4)complex AR residual</td>
</tr>
<tr>
<td>Sampling</td>
</tr>
<tr>
<td>Analysis window</td>
</tr>
<tr>
<td>Shift Length: 10.0[ms]</td>
</tr>
<tr>
<td>$F_0$ search range</td>
</tr>
<tr>
<td>Real-valued AR</td>
</tr>
<tr>
<td>I=14, L=2 (time-varying)</td>
</tr>
<tr>
<td>Pre-emphasis</td>
</tr>
<tr>
<td>Complex-valued AR</td>
</tr>
<tr>
<td>I=7, L=2 (time-varying)</td>
</tr>
<tr>
<td>Pre-emphasis</td>
</tr>
<tr>
<td>Criterion</td>
</tr>
<tr>
<td>Noise</td>
</tr>
<tr>
<td>(2)pink noise[8]</td>
</tr>
<tr>
<td>Noise Level</td>
</tr>
<tr>
<td>Interpolation</td>
</tr>
</tbody>
</table>

5. Conclusions

This paper has evaluated the performance of robust fundamental frequency estimation algorithm based on the robust TV-CAR speech analysis. The estimation accuracy is evaluated by GPE with respect to degree of voiced nature, viz., mode 3.2 and 1. The experiments using IRS filtered speech corrupted by white Gaussian noise or pink noise demonstrate that ELS-based robust time-varying complex speech analysis can perform better for stationary voiced speech and ELS-based time-invariant speech analysis can perform better for ordinary voiced frame.

The proposed frame-based $F_0$ estimation can be introduced as a open-loop adaptive codebook in CELP speech coding such
as G729 and G722.2, or iLBC. Since closed-loop final adaptive codebook search is carried out within the neighboring range for the lag pre-selected by open-loop search, it can realize more accurate adaptive codebook search for noisy speech, as a result, the speech quality can be improved in realistic environment. Moreover, to investigate the performance for each phoneme is future study.

6. Acknowledgements
This work was supported by Grand-in-Aid for Scientific Research (C), Research Project Number:20500158.

7. References

Figure 1: Experimental Results (All Modes)

Figure 2: Experimental Results (Mode 3)

Figure 3: Experimental Results (Mode 2)

Figure 4: Experimental Results (Mode 1)