Chapter 2

Low Computational Robust F_0 Estimation of Speech Based on TV-CAR Analysis

Keiichi Funaki and Takehito Higa

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51694

1. Introduction

The F_0 estimation determines a performance of speech processing such as speech coding, tonal speech recognition, speaker recognition, and speech enhancement. F_0 estimation named "YIN" has been proposed [1] and it is being prevalently used around the world due to its high performance and open-source policy. Speech processing is commonly applied in realistic noisy environments; hence, the performance is degraded seriously. It is well known that YIN does not perform well for noisy speech although it does perform best for clean speech. Accordingly, more robust F_0 estimation algorithm is desired and the robust F_0 estimation is long lasting problem in speech processing. We have already proposed robust F_0 estimation algorithm based on time-varying complex speech analysis for analytic speech signal [2][3]. 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 [2] and [3], complex LPC residual is used to calculate the criterion of weighted autocorrelation function (AUTOC) with a reciprocal of Average Magnitude Difference Function (AMDF) [6]. The complex residual is calculated from analytic speech signal by means of time-varying complex AR (TV-CAR) speech analysis method [4][5]. In [2], MMSE-based TV-CAR speech analysis [4] is introduced and in [3], ELS-based TV-CAR speech analysis [5] is introduced to calculate complex LPC residual signal. It has been reported in [2] 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 [3] that the ELS-based complex speech analysis can perform better even for additive pink noise. Furthermore, in order to investigate the effective-

ness of the time-varying analysis, the performance was compared for the frame with respect to degree of voiced nature [7]. The experiments using IRS filtered speech corrupted by white Gauss 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. However the computational cost turns to be larger by introducing time-varying analysis. In this paper, in order to reduce the computational cost, pre-selection is introduced. The pre-selection is performed by peak picking of speech spectrum based on the TV-CAR analysis [8]. The evaluation is carried out using Keele Pitch Database [9]. The reminder of the chapter is organized as follows. In Section 2, TV-CAR speech analysis is explained. Analytic signal and Time-Varying Complex AR (TV-CAR) model are explained. Two kinds of the TV-CAR parameter estimation algorithms from an analytic signal, viz., MMSE and ELS methods are explained. In Section 3, F_0 estimation algorithm is explained in detail. Sample-based pre-selection is explained and frame-based final–selection is explained. In Section 4, experimental results are explained and these confirm the effectiveness of the proposed method.

2. TV-CAR speech analysis

In this section, ELS-based robust TV-CAR speech analysis method is explained. Before the explanation, analytic signal and TV-CAR model is explained, in which analytic signal is output of the TV-CAR model. In 2.6, the benefit of the robust TV-CAR analysis is explained by showing the estimated sprctra from natural speech.

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 an all-pole model as follows.

$$y^c(t) = \frac{y(2t) + j \cdot y_H(2t)}{\sqrt{2}} \tag{1}$$

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

http://dx.doi.org/10.5772/51694

$$Y_{LPC}\left(z^{-1}\right)=\frac{1}{1+\sum_{i=1}^{1}a_i z^{-1}} \tag{2}$$

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

$$a_i^c\left(t\right)=\sum_{l=0}^{L-1}g_{i,l}^c f_l^c\left(t\right) \tag{3}$$

where $a_i^c(t)$,I,L, $g_{i,l}^c$,l and $f_l^c(t)$ are taken to be i-th complex AR coefficient at time t, AR order, finite order of complex basis expansion, complex parameter, and a complex-valued basis function, respectively. By substituting Eq.(3) into Eq.(2), one can obtain the following transfer function. Eq.(4) means the TV-CAR model.

$$Y_{TVCAR}\left(z^{-1}\right)=\frac{1}{1+\sum_{i=1}^{I}\sum_{l=1}^{L}g_{i,l}^c f_l^c\left(t\right)z^{-i}} \tag{4}$$

The input-output relation is defined as

$$y^c\left(t\right)=-\sum_{i=1}^{I}\sum_{l=0}^{L-1}g_{i,l}^c f_l^c\left(t\right)y^c\left(t-i\right)+u^c\left(t\right) \tag{5}$$

where $u^c(t)$ and $y^c(t)$ are taken to be complex-valued input and analytic speech signal shown in Eq.(1), respectively. In the TV-CAR model, the complex AR coefficient is modeled by a finite number of arbitrary complex basis functions such as Fourier basis, wavelet basis or so on. 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

$$
\begin{aligned}
&\bar{d} = i,\ l^{f} \\
&\bar{\theta}^{T} = \left[\bar{g}_{0}^{T},\ \bar{g}_{1}^{T},\ \cdots,\ \bar{g}_{l}^{T},\ \cdots,\ \bar{g}_{L-1}^{T}\right] \\
&\bar{g}_{l}^{T} = \left[g_{1,l}^{c},\ g_{2,l}^{c},\ \cdots,\ g_{i,l}^{c},\ \cdots,\ g_{I,l}^{c}\right] \\
&\bar{y}_{f}^{T} = \left[y^{c}(I),\ y^{c}(I+1),\ y^{c}(I+2),\ \cdots y^{c}(N-1)\right] \\
&\bar{u}_{f}^{T} = \left[u^{c}(I),\ u^{c}(I+1),\ u^{c}(I+2),\ \cdots u^{c}(N-1)\right] \\
&\bar{\Phi}_{f} = \left[\bar{D}_{0}^{f},\ \bar{D}_{1}^{f},\ \cdots \bar{D}_{l}^{f},\ \cdots,\ \bar{D}_{L-1}^{f}\right] \\
&\bar{D}_{l}^{f} = \left[\bar{d}_{1,l}^{f},\ \cdots \bar{d}_{i,l}^{f},\ \cdots,\ \bar{d}_{I,l}^{f}\right] \\
&\bar{d}_{i,l}^{f} = \begin{bmatrix} y^{c}(I-i)f_{l}^{c}(I),\ y^{c}(I+1-i)f_{l}^{c}(I+1) \\ \cdots,\ y^{c}(N-1-i)f_{l}^{c}(N-1)^{T} \end{bmatrix}
\end{aligned}
\tag{6}
$$

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

2.3. MMSE-based algorithm [4]

There are several algorithms that estimate the TV-CAR model parameter from complex-valued signal such as MMSE, WLS(Weighted Least Square), M-estimation, GLS(Generalized Least Square), and ELS(Extended Least Square). The MMSE-algorithm is basic algorithm and used for initial estimation of the ELS. Before explaining the ELS, the MMSE algorithm is explained.

MSE criterion is defined by

$$
\bar{r}_{f} = \left[r^{c}(I), r^{c}(I+1), \cdots, r^{c}(N-1)\right]^{T} = \bar{y}_{f} + \bar{\Phi}_{f}\hat{\theta} \tag{7}
$$

$$
r^{c}(t) = y^{c}(t) + \sum_{I=1}^{I}\sum_{l=0}^{L-1}\hat{g}_{i,l}^{c} f_{l}^{c}(t)\, y^{c}(t-i) \tag{8}
$$

$$
E = \bar{r}_{f}^{H}\bar{r}_{f} = \left(\bar{y}_{f} + \bar{\Phi}\hat{\theta}\right)^{H}\left(\bar{y}_{f} + \bar{\Phi}\hat{\theta}\right) \tag{9}
$$

Where $\hat{g}_{i,l}^{c}$ is the estimated complex parameter, $r^{c}(t)$ is an equation error, or complex AR residual and E is Mean Squared Error (MSE) for the equation error. To obtain optimal complex

http://dx.doi.org/10.5772/51694

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.

$$\left(\overline{\Phi}_f^H \overline{\Phi}_f\right)\hat{\theta} = -\overline{\Phi}_f^H \overline{y}_f \tag{10}$$

Superscript H denotes Hermitian transposition. After solving the linear equation of Eq.(10), we can get the complex AR parameter ($a_i^c(t)$) at time *t* by calculating the Eq.(3) with the estimated complex parameter $\hat{g}_{i,l}^c$.

2.4. ELS-based algorithm [5]

Figure 1 shows block diagram of ELS estimation. If the equation error shown as in Eq.(8) 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 (Figure 1(2)).

$$r^c(t) = -\sum_{k=1}^{K} b_k^c r^c(t-k) + e^c(t) \tag{11}$$

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.(11) is called a whiten filter. The TV-CAR model can be represented using Eq.(5) and Eq.(11) as follows.

$$y^c(t) = -\sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t)\, y^c(t-i) - \sum_{k=1}^{K} b_k^c r^c(t-k) + e^c(t) \tag{12}$$

Eq.(12) is the ELS model shown as in Figure 1(3). The parameter is estimated so as minimize the MSE for the whitened equation error in the ELS algorithm whereas the parameter is estimated so as minimize the MSE for the equation error in the MMSE algorithm shown as in Figure 1(1).

Eq.(12) can be expressed by the following vector-matrix notation.

$$\overline{y}_f = -\overline{\Phi}_f \overline{\theta} - \overline{R}_f \overline{b} + \overline{e}_f = -\left(\overline{\Phi}_f \overline{R}_f\right)\begin{pmatrix}\overline{\theta}\\ \overline{b}\end{pmatrix} + \overline{e}_f \tag{13}$$

Where

$$\overline{R}_f = \begin{pmatrix} r^c(I-1) & r^c(I-2) & \cdots & r^c(I-K) \\ r^c(I) & r^c(I-1) & \cdots & r^c(I+1-K) \\ \vdots & \vdots & \ddots & \vdots \\ r^c(t) & r^c(t-1) & \cdots & r^c(t-K) \\ \vdots & \vdots & \ddots & \vdots \\ r^c(N-2) & r^c(N-3) & \cdots & r^c(N-1-K) \end{pmatrix}$$
$$\overline{b} = \left[b_1^c, b_2^c, \cdots, b_K^c \right]^T \tag{14}$$
$$\overline{e}_f = \left[e^c(I), e^c(I+1), e^c(I+2), \cdots, e^c(N-1) \right]^T$$

By minimizing the MSE for Eq.(13), one can get the following equation.

$$\begin{pmatrix} \overline{\Phi}_f^H \overline{\Phi}_f & \overline{\Phi}_f^H \overline{R}_f \\ \overline{R}_f^H \overline{\Phi}_f & \overline{R}_f^H \overline{R}_f \end{pmatrix} \begin{pmatrix} \hat{\theta} \\ \hat{b} \end{pmatrix} = -\begin{pmatrix} \overline{\Phi}_f^H \overline{y}_f \\ \overline{R}_f^H \overline{y}_f \end{pmatrix} \tag{15}$$

By applying the well-known inversion Matrix lemma to Eq.(15), one can obtain the following equation.

$$\left(\overline{\Phi}_f^H \overline{\Phi}_f \right) \hat{\theta}_{bias} = \overline{\Phi}_f^H \overline{R}_f \hat{b} \tag{16}$$

$$\hat{\theta} = \hat{\theta}_0 - \hat{\theta}_{bias} \tag{17}$$

The MMSE estimated parameter $\hat{\theta}_0$ contains the biased element $\hat{\theta}_{\text{bias}}$. The unbiased estimation of $\hat{\theta}$ is calculated by $\hat{\theta}_0 - \hat{\theta}_{|}bias$. The ELS algorithm is equivalent to the GLS (Generalized Least Square) algorithm and more sophisticated algorithm. Since the equation error $r^c(t)$ cannot be observed, the iteration algorithm is required by estimating the A(z) and B(z). The iteration procedure is shown as follows.

1. Initial $\hat{\theta}_0$ is estimated by MMSE (Eq.(10)).

2. The equation error is calculated by Eq.(8).
3. $\hat{b}$ is estimated so as to minimize Eq.(18) using $r^c(t)$.

4. The bias parameter $\hat{b}$ is calculated by Eq.(16).

5. The unbiased parameter $\hat{\theta}$ is calculated by Eq.(17).

6. Go to 2.

http://dx.doi.org/10.5772/51694

$$\frac{1}{2\pi j}\oint_{|z|=1}\left|R(z)B(z)\right|^2\frac{dz}{z}=0 \tag{18}$$

In Eq.(18), R(*z*) is *z*-transform of $r^c(t)$ and B(*z*) is the transfer function of the whiten filter. The procedures from 2 to 5 are iterated with the pre-determined number. The ELS algorithm estimates two kinds of AR filters, A(*z*) and B(*z*), iteratively. Since the ELS algorithm can estimate unbiased and less effected speech spectrum against additive noise, more accurate F_0 and formants frequencies can be estimated. Thus, more accurate F_0 trajectories can be estimated than the MMSE estimation.

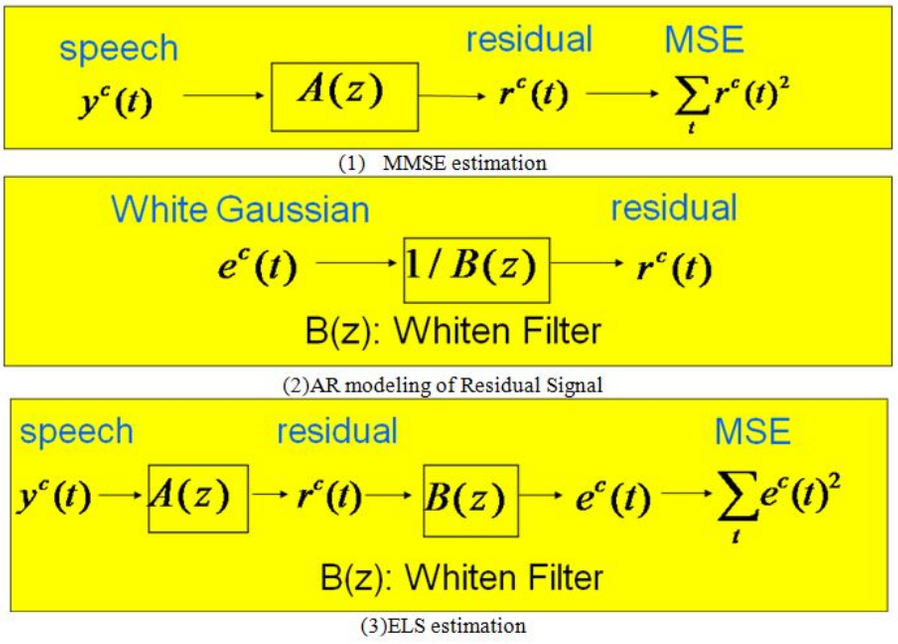


Figure 1. Block diagrams of MMSE and ELS estimation.

2.5. Benefit of robust TV-CAR speech analysis

In this paragraph, we explain the benefit of robust TV-CAR speech analysis by showing the estimated speech spectrum and explain its effectiveness on F_0 estimation of speech. Figure 2 shows example of the estimated speech spectra of natural Japanese vowel /o/ for analytic signal and conventional LPC analysis for speech signal.

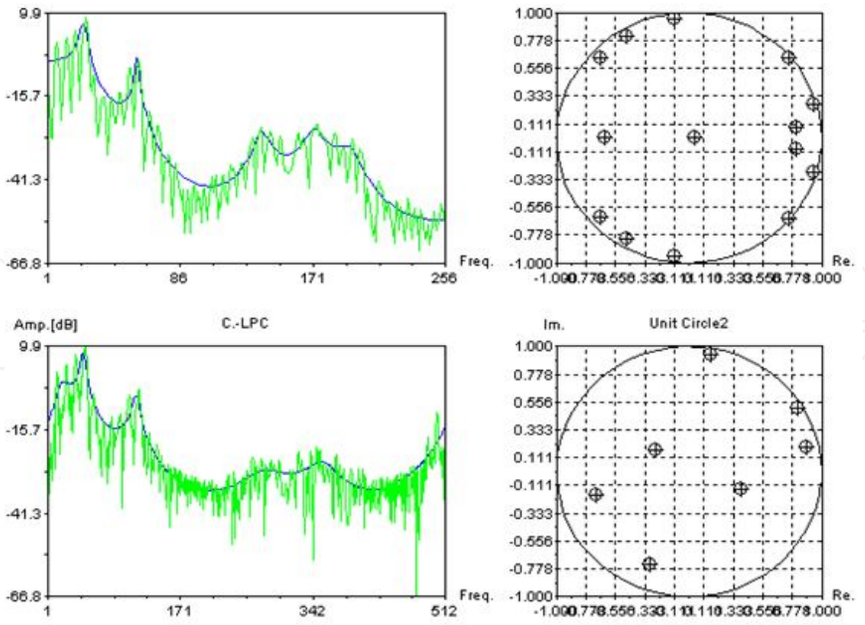
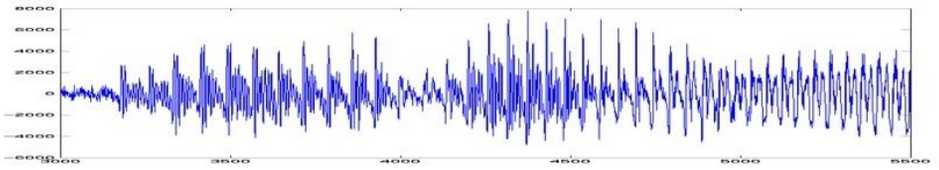


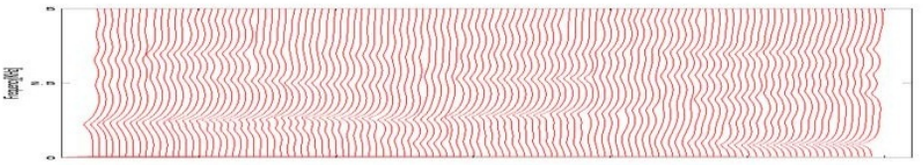
Figure 2. Estimated Spectra of vowel /o/ with complex and conventional LPC analysis.

In Figure 2, left side denote the estimated spectra. Upper is for real-valued LPC analysis. Lower is for complex-valued LPC analysis. Blue line means estimated spectrum by LPC analysis and green line means estimated DFT spectrum. Right side means estimated poles from the estimated AR filter. Figure 3 shows the estimated running spectrum for clean natural speech /arayu/ and for the speech corrupted by white Gaussian (10[dB]). In Figure 3, (1) means speech waveform, (2),(3),(4),(5) and (6) mean the estimated spectrum by MMSE-based time-invariant real-valued AR speech analysis, by MMSE-based time-invariant complex-valued AR speech analysis (L=1), by MMSE-based time-varying complex AR (TV-CAR) speech analysis (L=2), by ELS-based time-invariant complex-valued AR speech analysis (L=1), and by ELS-based time-varying complex AR (TV-CAR) speech analysis (L=2), respectively. Analysis order I is 14 for real analysis and 7 for complex analysis. Basis function is 1st order polynomial function (1,t). One can observe that the complex analysis can estimate more accurate spectrum in low frequencies whereas the estimation accuracy is down in high frequencies. Since speech spectrum provides much energy in low frequencies, it is expected that the high spectral estimation accuracy in low frequencies makes it possible to improve the performance on F_0 estimation. Furthermore, the ELS analysis can estimate more accurate spectrum than MMSE, so that the ELS analysis makes it possible to estimate more accurate F_0. Time-varying analysis can estimate tive-varying spectrum from speech. It is expected that the time-varying analysis enables to estimate more accurate F_0 since F_0 is varying in the analysis interval.

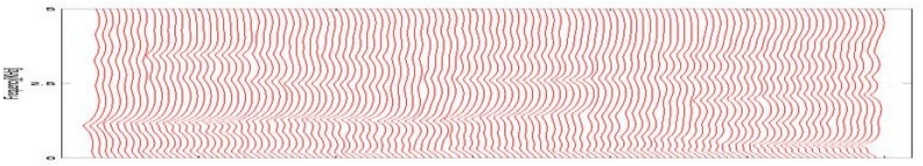
http://dx.doi.org/10.5772/51694



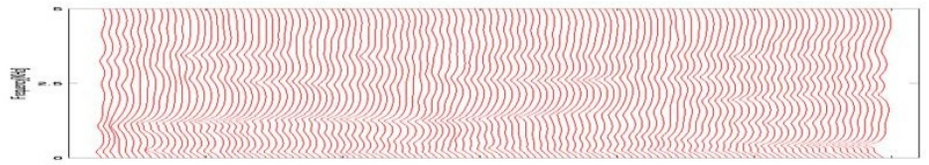
(1) clean speech waveform /arayu/



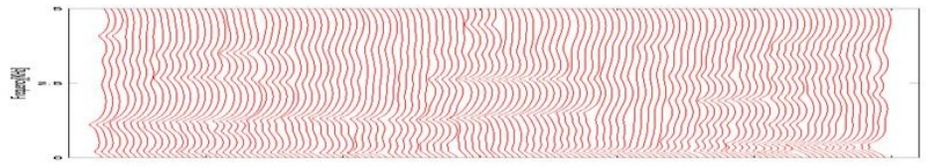
(2) MMSE-based time-invariant real AR analysis



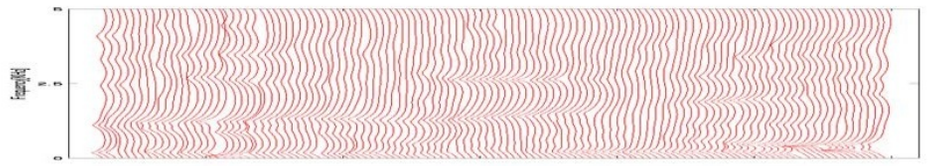
(3) MMSE-based time-invariant complex AR analysis (L=1)



(4) MMSE-based TV-CAR analysis (L=2)



(5) ELS-based time-invariant complex AR analysis (L=1)



(6) ELS-based TV-CAR analysis (L=2)

Figure 3. Estimated spectrum for noise corrupted speech /arayu/ (10[dB]).

3. F_0 Estimation method

Proposed method employs two-stage search of F_0. In first stage, pre-selection, F_0 and F_1 are estimated by using sample-based F_0 contour estimation [8]. In second stage, final-selection, F_0 is estimated by using frame-based F_0 estimation [3] within limited range based on the pre-estimated F_0 and F_1. The two-stage estimation makes it possible to reduce the computation with less degradation In 3.1, pre-selection algorithm is explained. In 3.2, final-selection algorithm is explained.

3.1. Sample-based pre-selection

F_0 and F_1 are estimated as the lowest two peak frequency, viz., glottal and first formant frequencies by peak-picking for the estimated time-varying speech spectrum. The procedure of F_0 and F_1 contour estimation is shown as in Figure 4

1. The set of complex-valued parameter $\hat{g}^{c}_{i,l}$ is estimated by the ELS algorithm for each analysis frame.
2. By using Eq.(3) and Eq.(4) with the estimated parameter $\hat{g}^{c}_{i,l}$, the speech power spectrum for each sample t is calculated, and the two peaks of the estimated spectrum are searched by the peak-picking.

The peak-pinking is carried out from low frequency to high frequency shown as in Figure 5. The estimated two peaks correspond to glottal formant (F_0) and first formant (F_1). The formant frequencies are estimated by solving the equation of the reciprocal of Eq.(4).

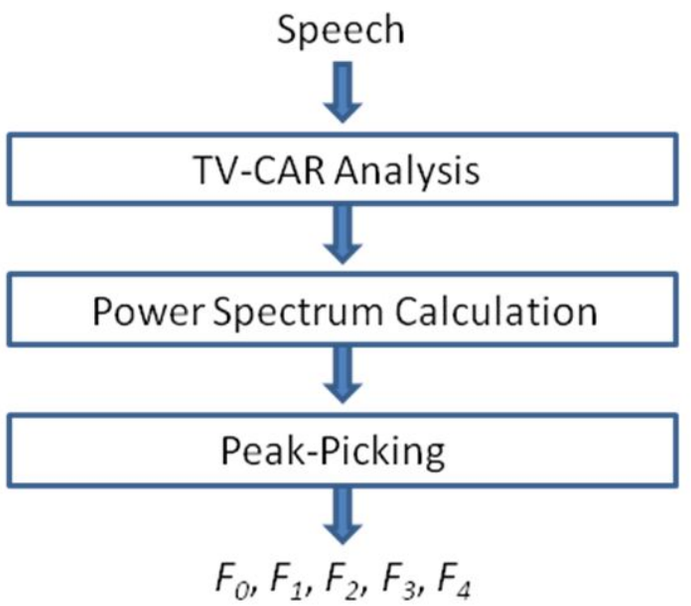


Figure 4. Flow of F_0 and F_1 contour estimation

http://dx.doi.org/10.5772/51694

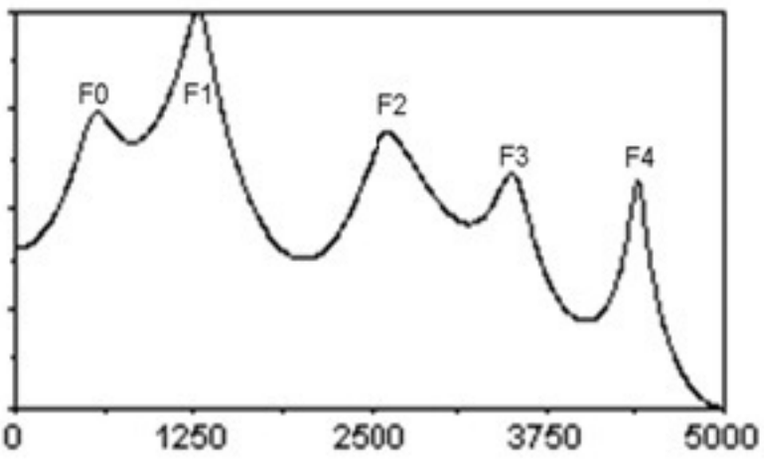


Figure 5. Peak Picking

3.2. Frame-based final-selection

In frame-based F_0 estimation, autocorrelation or AMDF is commonly used. In this paragraph, autocorrelation and AMDF are explained and then adopted weighted autocorrelation is explained.

Autocorrelation function (AUTOC) is defined by

$$f(\tau) = \frac{1}{N}\sum_{t=0}^{N-1} x(t)x(t+\tau) \tag{19}$$

where x(t) is target signal such as speech signal, LPC residual or so on, N is frame length and τ means delay. F_0 is selected as peak frequency for Eq.(19) within certain range of F_0.

AMDF is defined as follows.

$$p(\tau) = \frac{1}{N}\sum_{t=0}^{N-1} \left| x(t) - x(t+\tau) \right| \tag{20}$$

F_0 is selected as notch frequency for Eq.(20) within certain range of F_0. In Shimamura method [6], the AUTOC is weighted by a reciprocal of the AMDF shown as Eq.(21). 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.

$$G(\tau) = \frac{f(\tau)}{p(\tau) + m} \tag{21}$$

where f(τ) and p(τ) are AUTOC shown as in Eq.(19) and AMDF shown as in Eq.(20), respectively. In the frame-based method, Shimamura criterion shown as Eq.(21) is applied to complex AR residual extracted by the ELS-based TV-CAR speech analysis. The time-varying complex parameter is estimated and complex AR residual is calculated with the estimated complex parameter with Eq.(17). 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. F_0 is estimated within the range corresponding to 50-400[Hz]. In order to reduce the computational amount, the range is shortened by setting the upper value as follows.

$$\min\left(F_0^S + \left(F_1^S - F_0^S\right)\delta / 100, 400\right) \tag{22}$$

where F_0^s and F_1^s are estimated F_0 and F_1 by the sample-based pre-selection. Setting upper bound below F_1 can not only reduce the computational cost but also can reduce the estimation error.

4. Experiments

Speech signals used in the experiment are 5 long sentences uttered by 5 male speaker and 5 long sentences uttered by 5 female speaker of Keele pitch database [9]. Speech signals are filtered by an IRS filter [10]. The IRS filter is band pass FIR filter whose frequency response corresponds to that for analog part of the transmitter of telephone equipment. The frequency response is shown in Figure 6. In order to evaluate the proposed method for the speech data processed by speech coding, the IRS filter has to be introduced shown as in [2]. 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 Gauss noise or pink noise [11] 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

http://dx.doi.org/10.5772/51694

$$e_p(n) = F_e(n) - F_t(n) \tag{23}$$

where $F_t(n)$ is true F_0 value and $F_e(n)$ is the estimated one. The true values are derived by pitch file in Keele database. In Eq.(14), if $|e_{p(n)}| \geq F_t(n) \times THR/100$ 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 7,8,9 and 10 show the experimental results setting the THR as 10[%]. Figure 7 and 9 means the results for male speech. Figure 8 and 10 means the results for female speech. In Figures, (1) shows the results of GPEs or FPEs for additive white Gauss noise. (2) shows the results of GPEs or FPEs for additive pink noise. PROPOSED means the GPEs or FPEs for the proposed method with δ being 25. SP means the Shimamura method [6], viz., Shimamura criterion for speech signal. Other lines mean the GPEs or FPEs for the analysis method shown in Table 2. In all figures, X-axis means noise level of 30, 20, 10, 5, 0,−5[dB]. Y-axis means GPE[%] or FPE[Hz].

Figures 7 and 8 demonstrate that the proposed method can perform slightly better than the full-search method(TVC_E) for male speech while it can perform equivalently to the full search method(TVC_E) for female speech. Figures 9 and 10 show that the proposed one does not perform well in terms of FPE although the Shimamura method performs better in terms of FPE.

Speech data	Keele Pitch database [9] Male 5 long sentences Female 5 long sentences
IRS filter	64-th FIR [10]
Target signal	complex AR residual Sampling 10kHz/16bit
Analysis window	Window Length: 25.6[ms] Shift Length: 10.0[ms]
***F_0* search range**	50 to Eq.Eq.(22)
Complex-valued AR	I=7, L=2 (time-varying) Pre-emphasis 1 – *z*–1
Criterion	AUTOC/AMDF [6]
Noise	(1)white Gauss noise (2)pink noise [11] Noise Level 30,20,10,5,0,-5[dB]
Interpolation	3 point Lagrange's

Table 1. Experimental Conditions

	Line	Real or Complex	Non or TV	MMSE or ELS
LPC	Red Dotted	Real	Non	MMSE
TVR	Blue Dotted	Real	TV	MMSE
LPC_E	Magenta Dotted	Real	Non	ELS
TVR_E	Green Dotted	Real	TV	ELS
CLPC	Red Solid	Complex	Non	MMSE
TVC	Blue Solid	Complex	TV	MMSE
CLPC_E	Magenta Solid	Complex	Non	ELS
TVC_E	Green Solid	Complex	TV	ELS

Table 2. Analysis methods

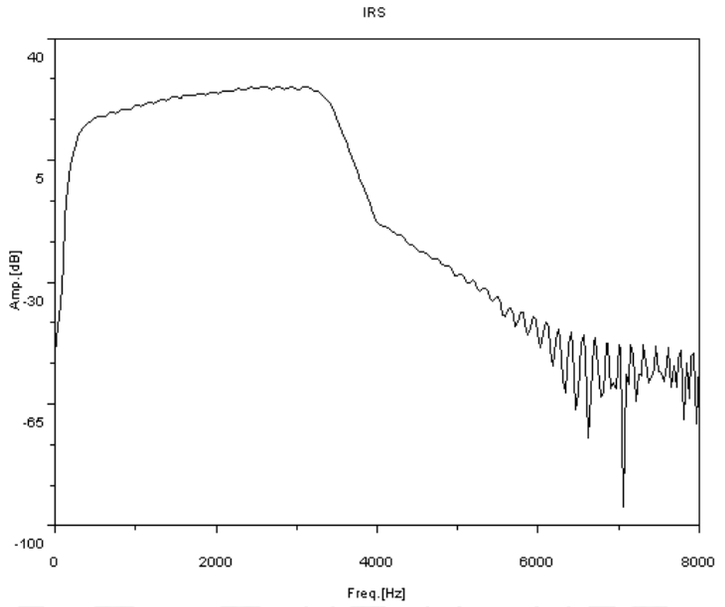
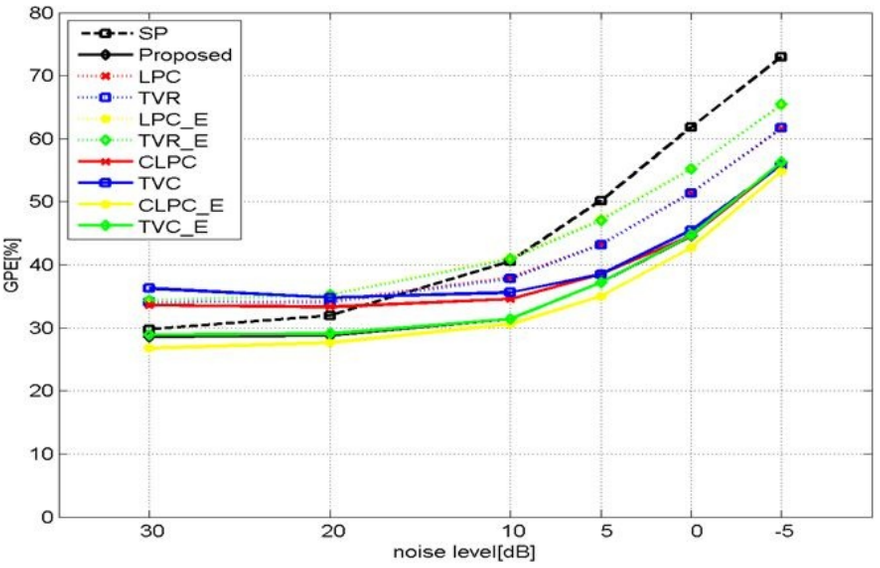
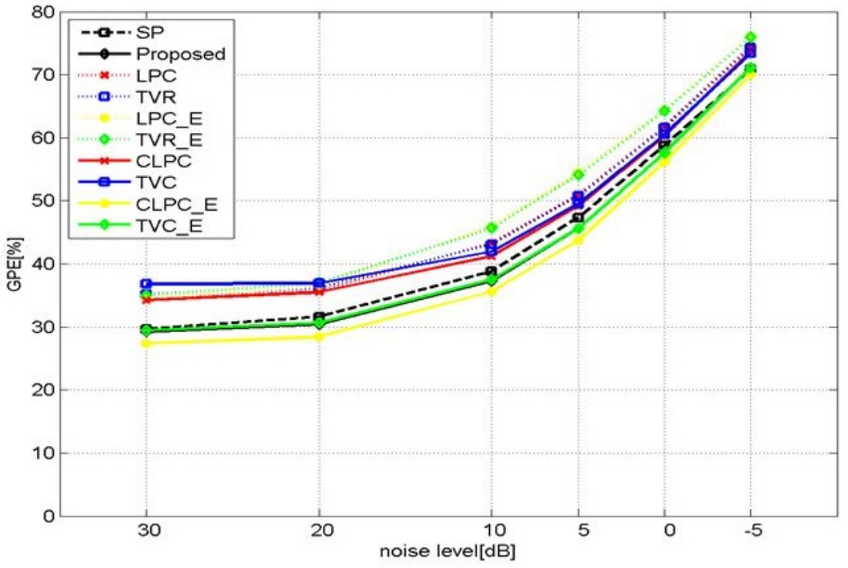


Figure 6. Frequency response of IRS filter

http://dx.doi.org/10.5772/51694

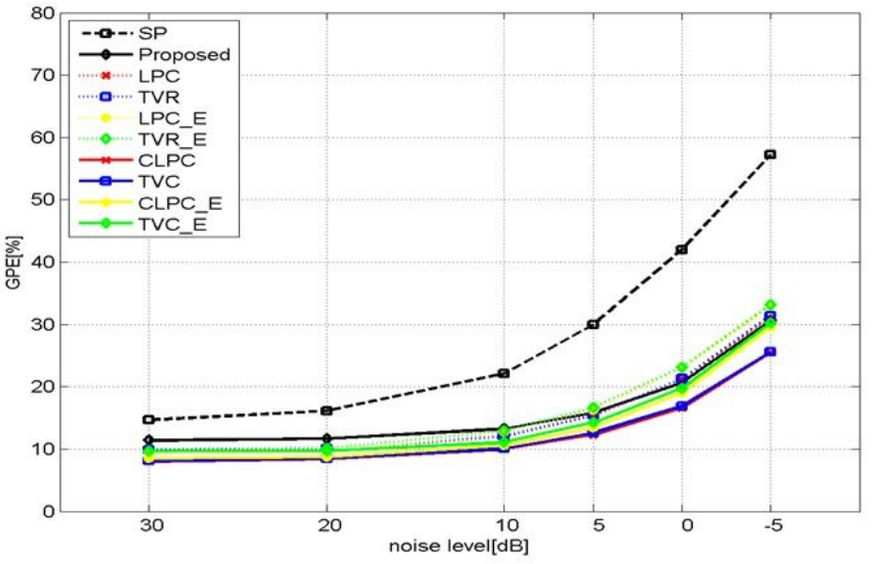


(1) GPEs for additive white Gauss noise

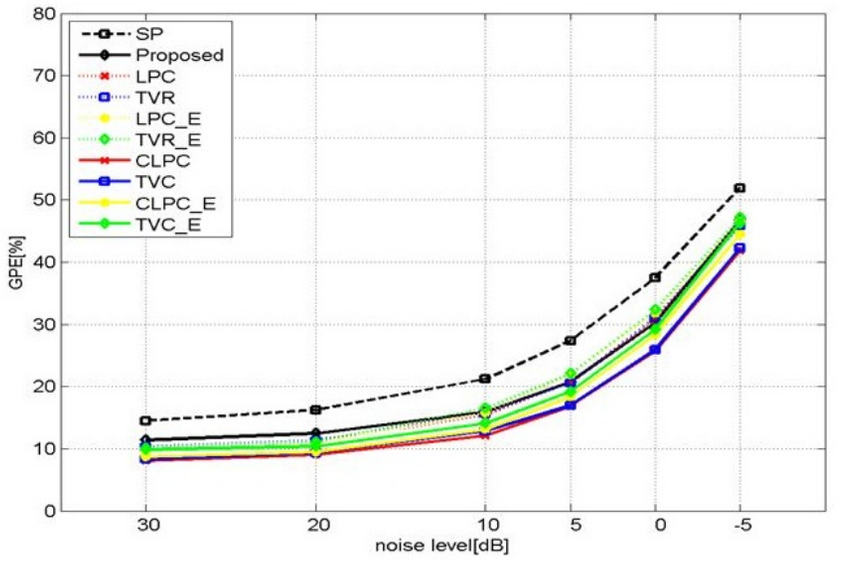


(2)GPEs for additive pink noise

Figure 7. Experimental Results for Male speech



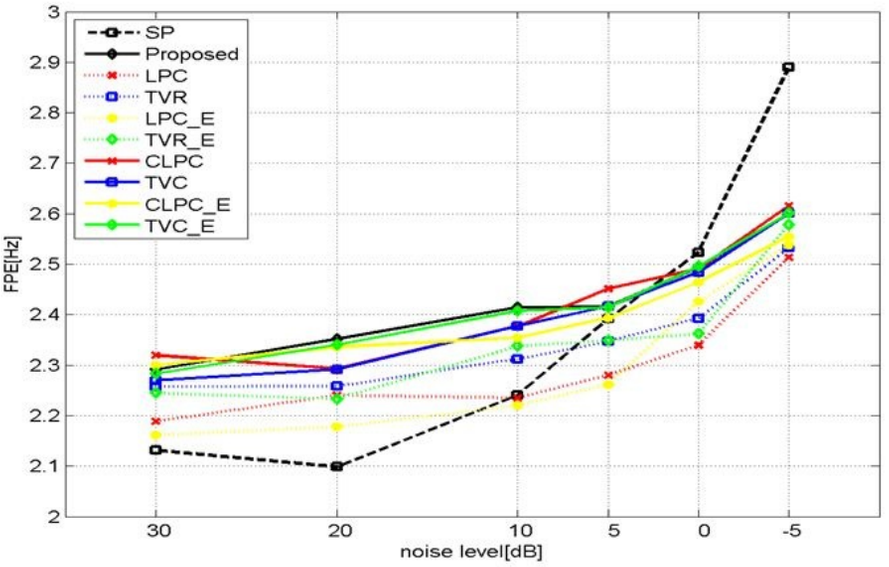
(1) GPEs for additive white Gauss noise



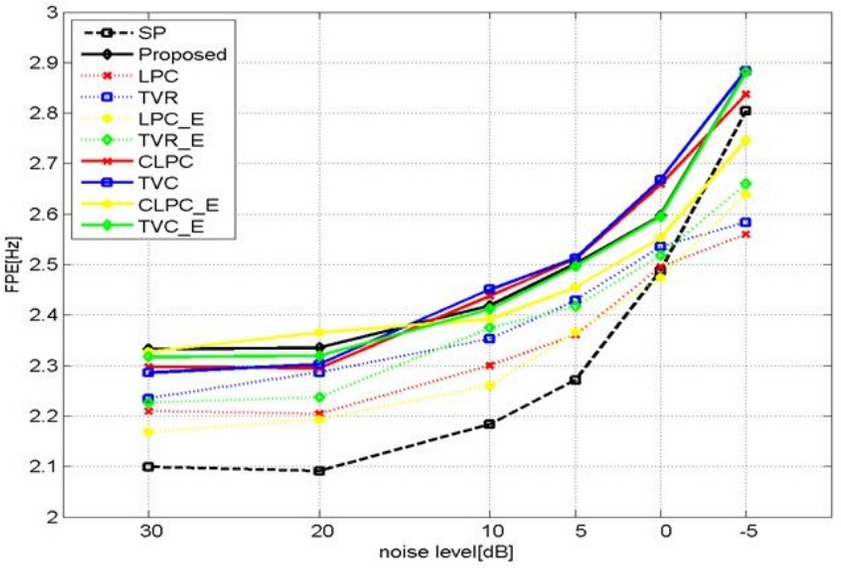
(2)GPEs for additive pink noise

Figure 8. Experimental Results for Female speech

http://dx.doi.org/10.5772/51694

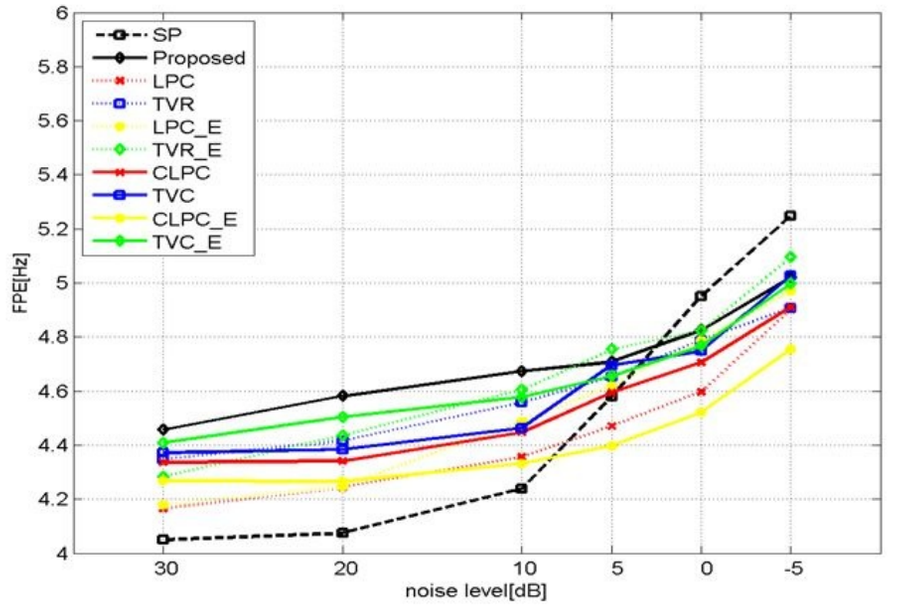


(1) FPEs for additive white Gauss noise

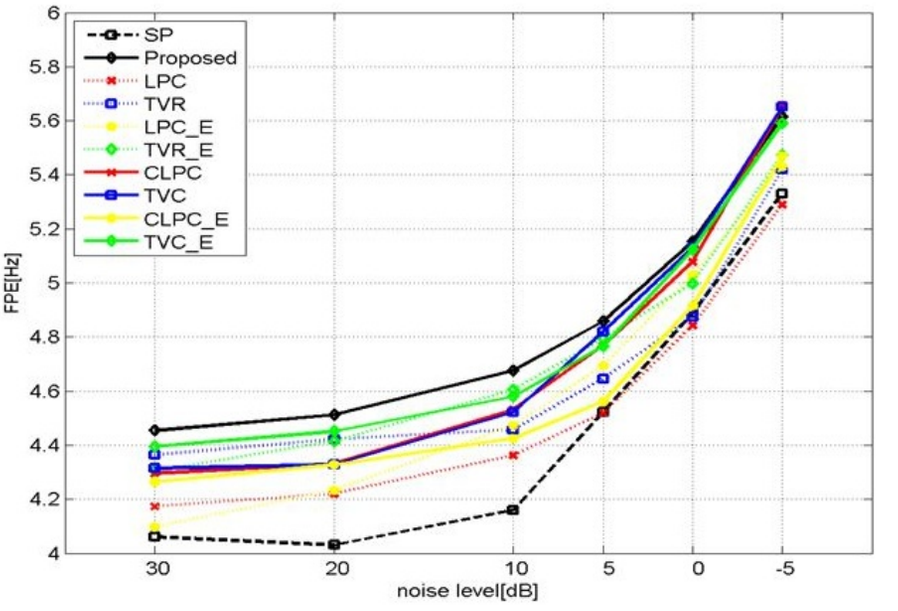


(2) FPEs for additive pink noise

Figure 9. Experimental Results for Male speech



(1) FPEs for additive white Gauss noise



(2)FPEs for additive pink noise

Figure 10. Experimental Results for Female speech

5. Conclusions

This paper proposed fast robust fundamental frequency estimation algorithm based on robust TV-CAR speech analysis. The method provides two stage of search procedure, pre-selection and final-selection. In the pre-selection, F_0 and F_1 are estimated by using time-

http://dx.doi.org/10.5772/51694

varying F_0 contour estimation. In the final-selection, F_0 is estimated for only the shorten range based on the pre-selected F_0 and F_1. The proposed method can perform better for male speech in terms of GPE with reduced computation.

Acknowledgements

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

Author details

Keiichi Funaki[1*] and Takehito Higa[2]

*Address all correspondence to: funaki@cc.u-ryukyu.ac.jp

1 Computing & Networking Center, University of the Ryukyus, Okinawa, Japan

2 Graduate School of Engineering and Science, University of the Ryukyus, Okinawa, Japan

References

[1] Alan de Cheveigne and H.Kawahra, YIN (2002). A fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America , 111(4), 1917-1930.

[2] K., Funaki, et al. (2007). Robust F0 Estimation Based on Complex LPC Analysis for IRS Filtered Noisy Speech. IEICE Trans. on Fundamentals Aug , E90-A(8)

[3] K., Funaki. (2008). F_0 estimation based on robust ELS complex speech analysis. Proc. EUSIPCO-2008, Lausanne, Switzerland Aug

[4] K., Funaki, Y., Miyanaga, & K., Tochinai. (1998). On a time-varying complex speech analysis. Proc. EUSIPCO-98,Rodes,Greece Sep

[5] K., Funaki. (2001). A time-varying complex AR speech analysis based on GLS and ELS method. Proc.EUROSPEECH2001, Aalborg Denmark Sep

[6] T., Shimamura, & H., Kobayashi. (2001). Weighted Autocorrelation for Pitch Extraction of Noisy Speech. IEEE Trans. Speech and Audio Processing , 9(7), 727-730.

[7] K., Funaki. (2010). On Evaluation of the F_0 Estimation Based on Time-Varying Complex Speech Analysis. Makuhari, Japan Sep

[8] K., Funaki. (2011). F_0 Contour Estimation Using ELS-based Robust Time-Varying Complex Speech Analysis. IEEE DSP/SPE workshop, Sedona, AZ, USA Jan

[9] Keele Pitch Database University of Liverpool http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html

[10] ITU-T Recommendation G.191. (2000). Software tools for speech and audio coding standardization. Nov.

[11] NOISE-X92,. http://spib.rice.edu/spib/selectnoise.html.